

# **Molecular biology based strategies to aid assembly in *de novo* genome projects**

Darren Heavens BSc (Hons.)

PhD by Publication

University of East Anglia

School of Biological Sciences

March 2018

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived there from must be in accordance with current UK copyright law. In addition, any quotation or extract must include full attribution.

# Abstract

This thesis presents and critically assesses work undertaken and published between 2009 and 2018. It evaluates the benefits, limitations and impact of novel approaches to next generation sequencing library construction for *de novo* genome projects developed by the author.

Since the first fully sequenced genome was published in 1978, DNA sequencing technology has advanced rapidly and costs reduced significantly. Next generation sequencers capable of sequencing millions of DNA molecules in parallel revolutionised the genomics industry. Today, if the right strategies are adopted, prokaryotic genomes can be fully sequenced in a matter of hours for a few hundred pounds and a high degree of contiguity achieved in even the most challenging eukaryotic genomes within a few weeks for tens of thousands of pounds.

Chapter 2 describes the design and application of a bespoke, high throughput bacterial artificial chromosome sequencing pipeline designed to sequence complex eukaryotic genomes harbouring a wide variety of repeat structures. Chapter 3 focuses on novel approaches to optimise insert size in amplification-free, paired-end library construction and Chapter 4 discusses innovative solutions to construct large insert, highly complex long mate pair libraries which have much tighter insert size distributions than previously published methods. Chapter 5 demonstrates the application of the methods discussed in earlier chapters in wheat *de novo* genome projects, highlighting the benefits the author's approaches bring to sequencing a complex polyploid plant genome.

The presented methods establish new ways of thinking about next generation sequencing library construction, pushing the boundaries of complexity and maximising spatial information.

**Keywords:** Genome assembly, next generation sequencing, DNA, *de novo*, amplification-free paired-end libraries, long mate pair libraries, bacterial artificial chromosomes.

# Table of contents

Abstract	2
Table of contents	3
List of figures	7
List of tables	9
List of accompanying material	10
Acknowledgements	13
1 Introduction	15
1.1 Genome complexity	16
1.1.1 Determining genome size	16
1.1.2 Variations in genome size	17
1.1.3 Ploidy	18
1.1.4 Sequencing coverage	18
1.1.5 Repetitive DNA sequences	19
1.2 NGS based genome sequencing	20
1.2.1 Single and paired-end library construction	20
1.2.2 LMP library construction	22
1.2.3 Next generation sequencing	22
1.2.3.1 454 Pyrosequencing	22
1.2.3.2 Illumina sequencing	24
1.3 Genome assembly	26
1.3.1 Assembly algorithms	26
1.3.2 Assembling contigs using DBGs	26
1.3.3 Scaffolding contigs	28
1.4 Genome project quality control	29
1.4.1 CN50 and SN50	29
1.4.2 KAT plots	29

1.4.3 BUSCO analysis	30
1.5 Genome project strategies	31
1.6 Summary	31
2 Increasing sequence contiguity in barley by decomplexing the genome	32
2.1 Bacterial Artificial Chromosomes (BACs)	32
2.2 Barley	33
2.3 Alternative strategies to sequence barley BACs	34
2.3.1 Different pooling strategies for sequencing barley BACs	34
2.3.2 Targeted approaches to sequencing barley BACs	35
2.4 Development of a novel BAC sequencing pipeline	36
2.4.1 DNA extraction	36
2.4.2 BAC paired-end library construction	40
2.4.3 BAC LMP library construction	42
2.4.4 Additional developments to the BAC sequencing pipeline	43
2.4.5 Outputs from the BAC sequencing pipeline	44
2.5 Comparing the different barley BAC sequencing strategies	44
2.5.1 Gene bearing versus WGS barley BACs	44
2.5.2 Comparison of the 2017 approaches	45
2.6 Summary	45
3 Optimising NGS paired-end library insert sizes.	47
3.1 NGS shotgun library construction	48
3.1.1 The benefits of controlling NGS library insert sizes	48
3.1.2 The benefits of read length and paired-end sequencing	49
3.1.3 The benefit of amplification-free, paired-end libraries	50
3.2 Evolution of paired-end library construction	52
3.2.1 Development of TALL libraries	52
3.2.2 Evaluation of DISCOVAR libraries	55
3.2.3 Improving DISCOVAR libraries	57

3.2.4 Maximising spatial information in paired-end libraries	59
3.3 Summary	61
4 Enhancing LMP library characteristics	63
4.1 Established LMP library construction strategies	63
4.1.1 Different approaches to constructing LMP libraries	63
4.1.2 Investigating the benefit of multiple insert size LMPs	65
4.1.3 Reducing costs and improving Nextera LMP outputs	66
4.2 Development of a unique LMP library construction protocol	67
4.2.1 The benefit of controlling LMP insert size and distribution	67
4.2.2 Improving LMP library complexity	70
4.2.3 Further improvements to my LMP protocol	71
4.3 Summary	74
5 Improving wheat <i>de novo</i> genome assemblies	76
5.1 Wheat	76
5.2 Decomplexing the wheat genome	78
5.2.1 Sequencing wheat BACs	78
5.2.2 Sequencing flow sorted wheat chromosomes	79
5.2.3 Sequencing wheat progenitors	80
5.3 WGS wheat genome project strategies	80
5.4 TGAC wheat genome assemblies	81
5.4.1 CS42 assembly	81
5.4.2 Additional wheat line genome projects	83
5.4 Summary	86
6 Discussion	88
6.1 Short read sequencing	89
6.2 Long range spatial information	91
6.2.1 Optical mapping	91
6.2.2 Hi-C	92

6.2.3 Single molecule sequencers	92
6.2.4 Improving assemblies through analysis of the gene space	94
6.3 Linked Reads	95
6.3.1 Using standard paired-end and LMP libraries	95
6.3.2 Introducing the 10x Genomics Chromium	96
6.4 DNA integrity	97
6.5 Future strategies for <i>de novo</i> genome projects	97
6.6 Summary	99
Definitions	101
Glossary	104
References	106
Appendix 1: Letters of support	114
Appendix 2: Publications submitted	125

# List of figures

**Figure 1.1:** The variation in genome size.

**Figure 1.2:** Classes of DNA sequence repeats.

**Figure 1.3:** The basis of 454 sequencing.

**Figure 1.4:** The structure of an Illumina compatible paired-end library molecule.

**Figure 1.5:** The basis of Illumina sequencing.

**Figure 1.6:** Using  $k$ -mers to build a de Bruijn graph.

**Figure 1.7:** The principle of scaffolding contigs.

**Figure 1.8:** A KAT plot of a *S. coelicolor* assembly.

**Figure 2.1:** The percentage of contaminating host *E. coli* DNA present in 48 BAC DNA extractions before DNase treatment, as determined by qPCR and after DNase treatment, as determined by FastQC.

**Figure 2.2:** An overview of the high-throughput BAC DNA extraction pipeline.

**Figure 2.3:** Bioanalyzer electropherograms of Nextera libraries using the manufacturer provided buffers.

**Figure 2.4:** The effect of varying DNA input on library size profiles in the optimised Nextera based library construction protocol.

**Figure 2.5:** Overview of the BAC paired-end library construction and sequencing pipeline.

**Figure 3.1:** 454 Pyrosequencing read length distribution plots.

**Figure 3.2:** The consequences of E-gel size selection and amplification in paired-end library construction.

**Figure 3.3:** Paired-end library Bioanalyzer electropherograms.

**Figure 3.4:** TALL, DISCOVAR and Improved DISCOVAR library insert size distribution plots.

**Figure 3.5:** KAT plot of the *S. verrucosum* DISCOVAR assembly.

**Figure 3.6:** The effect of different size exclusion parameters on library insert size distribution.

**Figure 4.1:** The effect of increasing DNA concentration on transposase mediated fragmentation.

**Figure 4.2:** The presence of smaller than target insert size molecules in LMP libraries.

**Figure 4.3:** Agilent TapeStation genomic tape electropherogram of freshly extracted WLA DNA.

**Figure 4.4:** The absence of smaller than target insert size molecules in LMP libraries.

**Figure 5.1:** KAT plots for the TGAC (A) and the IWGSC chromosome survey(B) CS42 wheat assemblies.

**Figure 5.2:** KAT plot for the four new hexaploid wheat lines assembled at EI.

## List of tables

**Table 2.1:** Assembly metrics for targeted approaches to sequencing barley BACs.

**Table 2.2:** Primer sequences for qPCR assays for pIndigo-BAC5 and *E. coli* DH10B.

**Table 3.1:** 454 Pyrosequencing sequence outputs.

**Table 3.2:** The effect on Q30 of simultaneously sequencing libraries with different insert sizes.

**Table 3.3:** Assembly metrics for *S. verrucosum* and *S. tuberosum*.

**Table 4.1:** LMP library characteristics for different genome projects.

**Table 4.2:** DNA recovered post size selection and library characteristics for ELF based CS42 LMP libraries.

**Table 4.3:** The proportion of unique, true LMPs in subsampled CS42 and WLA LMP reads.

**Table 5.1:** Assembly metrics for different wheat based genome assemblies.

**Table 5.2:** BUSCO analysis for five different CS42 wheat genome projects.

**Table 5.3:** Assembled content and contig/ scaffold number for five wheat lines sequenced at EI.

**Table 6.1:** The current cost and DNA requirements for different NGS library construction methods at EI.

## List of accompanying material

**Publications submitted for this PhD by Publication:** This thesis is based on the manuscripts listed in chronological order below. My role in these publications was to create, optimise and apply the novel NGS library construction protocols I present in this thesis. I was responsible for all aspects of the library construction experimental design and I provided input into writing and editing each of the manuscripts.

In Appendix 1 there are letters of support from some of my co-authors and, in Appendix 2, the original publications are reproduced with the kind permission of the relevant journals.

Barke J, Seipke RF, Grüschorow S, **Heavens D**, Drou N, Bibb MJ, Goss RJ, Yu DW, Hutchings MI. A mixed community of actinomycetes produce multiple antibiotics for the fungus farming ant *Acromyrmex octospinosus*. *BMC Biol.* 2010 Aug 26;8:109<sup>1</sup>.

Complete genome sequence of the proteolytic *Clostridium botulinum* type A5 (B3') strain H04402 065. Carter AT, Pearson BM, Crossman LC, Drou N, **Heavens D**, Baker D, Febrer M, Caccamo M, Grant KA, Peck MW. *J Bacteriol.* 2011 May;193(9):2351-2. doi: 10.1128/JB.00072-11. Epub 2011 Mar 4. PMID:21378191<sup>2</sup>.

Genome sequence of the vertebrate gut symbiont *Lactobacillus reuteri* ATCC 53608. **Heavens D**, Tailford LE, Crossman L, Jeffers F, Mackenzie DA, Caccamo M, Juge N. *J Bacteriol.* 2011 Aug;193(15):4015-6. doi: 10.1128/JB.05282-11. Epub 2011 May 27. PMID:21622738<sup>3</sup>.

Draft genome sequence of *Streptomyces* strain S4, a symbiont of the leaf-cutting ant *Acromyrmex octospinosus*. Seipke RF, Crossman L, Drou N, **Heavens D**, Bibb MJ, Caccamo M, Hutchings MI. *J Bacteriol.* 2011 Aug;193(16):4270-1. doi: 10.1128/JB.05275-11. Epub 2011 Jun 17. PMID:21685285<sup>4</sup>.

A method to simultaneously construct up to 12 differently sized Illumina Nextera long mate pair libraries with reduced DNA input, time, and cost. **Heavens D**, Accinelli GG, Clavijo B, Clark MD. *Biotechniques*. 2015 Jul 1;59(1):42-5. doi: 10.2144/000114310. eCollection 2015 Jul. PMID:26156783<sup>5</sup>.

W2RAP: a pipeline for high quality, robust assemblies of large complex genomes from short read data. Clavijo B, Garcia Accinelli G, Wright J, **Heavens D**, Barr K, Yanes L, and Di Palma F. *bioRxiv* 110999; doi: <https://doi.org/10.1101/110999><sup>6</sup>.

An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, Borrill P, Kettleborough G, **Heavens D**, Chapman H, Lipscombe J, Barker T, Lu FH, McKenzie N, Raats D, Ramirez-Gonzalez RH, Counce A, Peel N, Percival-Alwyn L, Duncan O, Trösch J, Yu G, Bolser DM, Namaati G, Kerhornou A, Spannagl M, Gundlach H, Haberer G, Davey RP, Fosker C, Palma FD, Phillips AL, Millar AH, Kersey PJ, Uauy C, Krasileva KV, Swarbreck D, Bevan MW, Clark MD. *Genome Res.* 2017 May;27(5):885-896. doi: 10.1101/gr.217117.116. PMID: 28420692<sup>7</sup>.

A chromosome conformation capture ordered sequence of the barley genome. Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, Bayer M, Ramsay L, Liu H, Haberer G, Zhang XQ, Zhang Q, Barrero RA, Li L, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Staňková H, Vrána J, Chan S, Muñoz-Amatriaín M, Ounit R, Wanamaker S, Bolser D, Colmsee C, Schmutz T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Chaillyan A, Sampath D, **Heavens D**, Clissold L, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, McCooke JK, Tan C, Wang P, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Borisjuk L, Houben A, Doležel J, Ayling S, Lonardi S, Kersey P, Langridge P, Muehlbauer GJ, Clark MD, Caccamo M, Schulman AH, Mayer KFX, Platzer M, Close TJ, Scholz U, Hansson M, Zhang G, Braumann I, Spannagl M, Li C, Waugh R, Stein N. *Nature*. 2017 Apr 26;544(7651):427-433. doi: 10.1038/nature22043. PMID:28447635<sup>8</sup>.

Construction of a map-based reference genome sequence for barley, *Hordeum vulgare* L. Beier S, Himmelbach A, Colmsee C, Zhang XQ, Barrero RA, Zhang Q, Li L, Bayer M, Bolser D, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Staňková H, Vrána J, Chan S, Muñoz-Amatriaín M, Ounit R, Wanamaker S, Schmutz T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Sampath D, **Heavens D**, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, McCooke JK, Tan C, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Houben A, Doležel J, Ayling S, Lonardi S, Langridge P, Muehlbauer GJ, Kersey P, Clark MD, Caccamo M, Schulman AH, Platzer M, Close TJ, Hansson M, Zhang G, Braumann I, Li C, Waugh R, Scholz U, Stein N, Mascher M. *Sci Data*. 2017 Apr 27;4: 170044. doi: 10.1038/sdata.2017.44. PMID: 28448065<sup>9</sup>.

A critical comparison of technologies for a plant genome sequencing project. Paajanen P, Kettleborough G, Lopez-Girona E, Giolai M, **Heavens D**, Baker D, Lister A, Wilde G, Hein I, Macaulay I, Bryan G and Clark M. *bioRxiv* 201830; doi: <https://doi.org/10.1101/201830><sup>10</sup>.

Independent assessment and improvement of wheat genome assemblies using Fosill jumping libraries. Lu F-H, McKenzie N, Kettleborough G, **Heavens D**, Clark M.D, and Bevan MW. *bioRxiv* 219352; doi: <https://doi.org/10.1101/219352><sup>11</sup>.

## Acknowledgements

This thesis combines over 40 years of theoretical knowledge, nearly 30 years of practical experience and the last nine years of my research. I have been very fortunate to work with some very talented individuals, some of who get a mention below.

Sir Isaac Newton is reputed to have once said:

"If I have seen further, it is by standing on the shoulders of giants."

I sincerely hope that I have stood on my peers' shoulders, learnt something from them, and then built on that knowledge to further advance science.

To my supervisors Neil Hall and Iain Macaulay I thank you for your time, patience and guidance while assembling this thesis. I would have not begun this PhD without the support of Daniel Swan, Kirsten McLay and Matt Clark and I really miss the debates we used to have, so thank you for convincing me to write it. I must thank my darling wife Jane. It was her belief in me that got this started and faith in my ability that got me finished. For my two daughters, Alexa and Jemima, I hope that one day you will find time to read this and be proud of your Dad. To my parents, I thank you for your continued support and unwavering love.

The list of colleagues past and present who have influenced my career is a book in itself so, I apologise that I cannot name you all individually. Graham, who gave me my first job in genomics, and along with Ken, Alan and Richard taught me everything from first principles without a kit in sight and laid the foundations for my molecular biology understanding that I still use. Without their inspirational teaching I would not still be working in genomics. None of this would have been possible without Jon Clarke, who rescued me and brought me to Norwich, for which I am eternally grateful. The team at the John Innes Genome Laboratory were exactly what I needed. A great balance of work and play so thank you Davo, Beth, Helen and Stu, especially for the extra-curricular activities. Ian Bedford was also instrumental in me writing this thesis. As an early UEA PhD by publication student, his incite into the demands of

undertaking the task were refreshingly honest. I was so grateful when starting writing, of knowing what to expect.

A special mention should go to Andrew, aka 'Mr Cheltenham', who has been the one constant during my time in Norwich. He is the 'Top Man' and coffee breaks would not be the same without him. I also need to thank Bernardo, Gonza and Jon who develop the assembly algorithms and assemble complex genomes at Earlham Institute (EI). Interacting with them and understanding their bioinformatic capabilities, desires and how their minds work has led to the development of many of the techniques presented in this thesis. Without their continual feedback and demanding requirements, I would not have a thesis to defend. I would also like to thank all my publication co-authors, especially those providing me with letters of support and furnishing them with such kind words, and my colleagues past and present. The world of work has been a wild roller coaster and well worth the ride!

# 1 Introduction

Genome projects aim to provide an accurate sequence against which others can be compared. For agronomically important plants such as bread wheat (*T. aestivum*) and barley (*H. vulgare*), decoding their genomes has the potential to help identify the genetic basis of important traits such as yield, nutritional value, disease resistance and drought tolerance. In 2016/ 17, global production of wheat was 755 million metric tons and 147.9 million metric tons of barley were harvested. For wheat, the production in the United States in 2016 was worth \$9.1 billion<sup>13</sup>. With the UN predicting the global population to potentially rise to 16 billion by 2100<sup>14</sup>, world food production needs to increase significantly. Providing good quality reference genomes will hopefully allow breeders to improve their selection programmes and rapidly introduce new varieties that will contribute toward global food security.

The first complete DNA sequence of an organism, the 5.4 Kbp bacteriophage PhiX, was published in 1978<sup>15</sup>. This was followed by a succession of high profile genome projects. *H. influenzae* was the first fully sequenced prokaryote in 1995<sup>16</sup>, *S. cerevisiae* the first eukaryote in 1996<sup>17</sup>, *C. elegans* the first animal in 1998<sup>18</sup>, *A. thaliana* the first plant in 2000<sup>19</sup> and the first drafts of the human genome were published in 2001<sup>20,21</sup> and deemed complete (to 99.99 % accuracy) in 2003<sup>22</sup>. Each of these were sequenced using the same dideoxy sequencing chemistry, sequencing up to 1 Kbp per read<sup>23</sup>. For eukaryotic genome projects, Sanger sequencing was both expensive and time consuming. The budget for the 3 Gbp Human Genome Project (HGP) was >£10 million and it took 13 years to complete. This limited the number of genomes that would be sequenced using this technology.

With the introduction of the first commercial Next Generation Sequencing (NGS) instrument, the 454 pyrosequencer<sup>24-27</sup>, closely followed by the Solexa (now Illumina) Genetic Analyser<sup>28</sup>, increased sequence yields transformed genomic research. Due to much shorter read lengths of between 25 and 100 bp, and a good reference to compare against, many early adopters of NGS technology employed the instruments for resequencing of humans as the race toward a \$1,000 genome intensified<sup>29-31</sup>.

As read lengths and outputs increased, and costs reduced, numerous opportunities were created to optimise and develop NGS library construction protocols to aid *de novo* genome assembly. Novel, laboratory based methods developed by the author to help improve assembly accuracy and genome contiguity are presented and discussed in this thesis. Many of the protocols appear in the publications listed in Appendix 2 so this thesis does not have a dedicated material and methods chapter. Where unpublished methods are discussed, details to replicate the studies are written within the relevant results chapters.

## **1.1 Genome complexity**

When undertaking a genome project, consideration needs to be given to genome complexity. It is the combination of genome size, ploidy and the nature of repetitive DNA sequences that dictates the amount of sequence required and helps define the strategies needed for genome project success.

### **1.1.1 Determining genome size**

Genome size in base pairs can be determined empirically by measuring the picograms of DNA within a single haploid cell, this is known as the C-value<sup>32</sup>, or by using *k-mers*<sup>33</sup>.

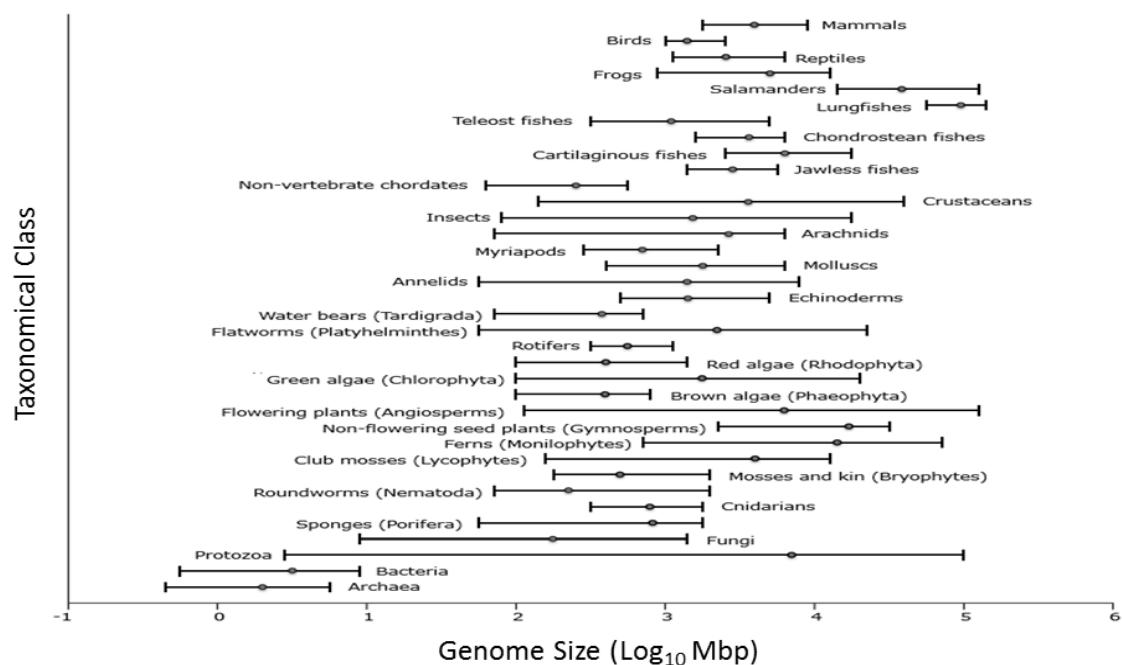
By comparing the molecular mass of the four component nucleotides of DNA- Adenine (A), Thymine (T), Guanine (G) and Cytosine (C), it is possible to calculate the average mass of a nucleotide base pair. This can then be used to calculate the number of nucleotides in 1 pg as 977.8 Mbp. For the hexaploid bread wheat (Chinese Spring 42, (CS42)) a C-value of 17.33<sup>34</sup> and for barley of 5.29<sup>35</sup> suggests genome sizes of 16.95 Gbp and 5.16 Gbp respectively.

*K-mers* represent genomic sequences of length *k* which can contain all possible combinations of nucleotides. For a *k* length of 17 bp there could potentially be >17 billion different sequences. *K-mers* can be used to estimate genome size although

technical biases such as those caused by amplification and sequencing errors or for biological reasons, such as repetitive sequences, can affect accuracy. Using a *k-mer* that is large enough to map uniquely within the genome, the *k-mer* frequency is determined to calculate the coverage. Genome size can then be calculated by dividing the total number of *k-mers* by the coverage.

### 1.1.2 Variations in genome size

Genome size varies widely. Viruses are the smallest life forms on earth and can be RNA based or DNA based. They range in size from 1.8 Kbp<sup>36</sup> to 2.5 Mbp<sup>37</sup>. Genome size ranges of different prokaryotes and eukaryotes are shown in Figure 1.1.



**Figure 1.1:** The variation in genome size. Genome sizes plotted as  $\log_{10}$  Mbp per haploid genome for many different taxonomical classes. Average genome size within a class is shown by the dot on the line.

Figure adapted from an image by Gregory<sup>38</sup>.

### **1.1.3 Ploidy**

For most of their life cycle, prokaryotes with a single copy of their chromosome per cell, are deemed haploid. Eukaryotes are predominantly diploid, having two copies of each chromosome per cell. Some species have more than two copies and are termed polyploids. This phenomenon occurs due to whole genome duplication events. When whole genome duplication is within the same species, which usually occurs in errors during meiosis or mitosis causing the fusion of gametes, then these are termed autopolyploids. When it is between closely related species, via hybridisation, these are known as allopolyploids. Polyploidy is much more common in plants than animals with an example of an autotetraploid being the cultivated potato (*S. tuberosum*)<sup>39</sup> and an allohexaploid being bread wheat<sup>7</sup>.

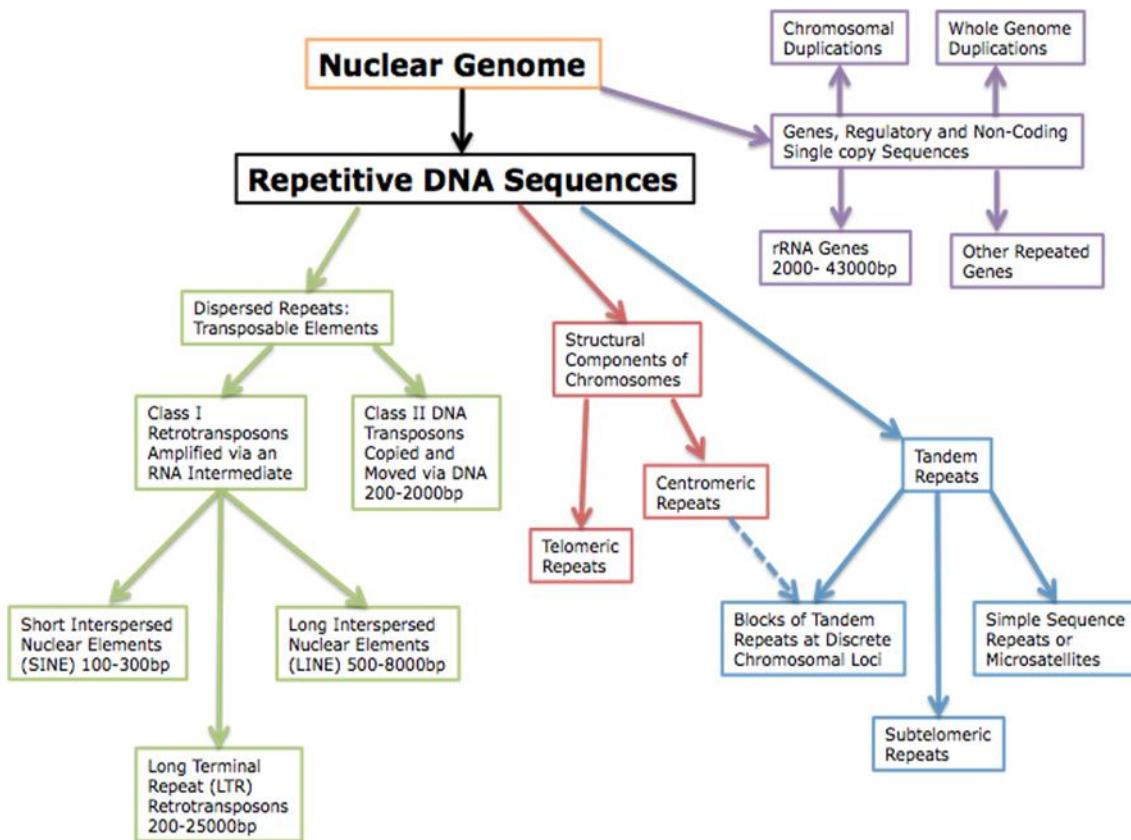
Assembling the genome of polyploids can be more difficult than diploids or haploids. The presence of significant amounts of homology between the different sets of chromosomes can make resolving and orienting these regions more challenging.

### **1.1.4 Sequencing coverage**

Genome size and ploidy dictate how much sequence is required to assemble a complete genome. It is generally accepted that for haploids and diploids >30x genome coverage of single/ paired-end NGS libraries is sufficient, rising to >60x for polyploids. If Long Mate Pair (LMP) libraries are constructed, a total of >30x genome coverage is targeted, irrespective of ploidy, and this is usually across libraries with at least two different insert sizes<sup>40,41</sup>. Sequencing to this depth helps identify variants such as Single Nucleotide Polymorphisms (SNPs), insertions and deletions and ensures that every nucleotide is covered multiple times and at different points within different reads. This allows for sequencing errors to be identified and corrected.

### 1.1.5 Repetitive DNA sequences

Generating highly contiguous genome assemblies is dependent upon being able to identify the unique sequence flanking any given repeat sequence. The major classes of repeat structures and their size ranges in eukaryotes are shown in Figure 1.2.



**Figure 1.2:** Classes of DNA sequence repeats. The major divisions of repetitive DNA sequence found in eukaryotic genomes. Where defined size ranges are known, these are shown.

Figure adapted from an image in Biscotti *et al.*<sup>42</sup> and data from Treangen and Salzberg<sup>43</sup>.

Of these, it is the dispersed repeats that are the most difficult to resolve in genome assembly projects due to their size and copy number. Short Interspersed Nuclear

Elements (SINEs) are present at around 15 % and Long Interspersed Nuclear Elements (LINEs) at around 21 % of the human genome<sup>42</sup>. The collection of Long Terminal Repeat (LTR) retrotransposons present in maize account for around 75 % of its genome<sup>44</sup>.

## 1.2 NGS based genome sequencing

DNA needs to be modified and platform-specific adapters introduced to each end of the molecule to enable them to be sequenced on NGS instruments. This process is called library construction. Protocols presented in this thesis were developed for the construction of NGS compatible paired-end and LMP libraries.

### 1.2.1 Single and paired-end library construction

Single and paired-end libraries are sometimes referred to as shotgun libraries and can be sequenced from one end to generate a single-end read, or from both ends to generate a paired-end read. By sequencing from both ends, spatial information relating to the distance between the reads can be used to improve contiguity in genome assemblies.

Early protocols to manipulate DNA to construct libraries suitable for next generation sequencing were based on *in vivo* cloning technologies. They typically required between 1 and 5 µg of DNA >10 Kbp and targeted insert sizes up to 400 bp. DNA is first fragmented by physical means using either a nebuliser or by ultrasonication. This fragmented DNA can either have a 5' or a 3' overhang or be blunt ended. During end repair, DNA polymerase I extends 5' to 3', like most polymerases, but it also has 3' to 5' single strand exonuclease activity ensuring that most molecules become blunt ended. A Phospho-Nucleotide Kinase (PNK) is also used to phosphorylate the 5' nucleotide to enable adapter ligation.

Blunt end molecules are then subjected to addition of a single adenine to the 3' end of each DNA strand in a process known as A tailing. This uses Klenow Fragment and

these A tailed molecules are then subjected to ligation of adapter molecules which have a 3' T overhang using a DNA ligase. Ligation is performed in the presence of polyethylene glycol (PEG) which acts as a crowding agent, effectively increasing the concentration of the DNA and ligase making the reaction more efficient. Using Y shaped adapters, which have the appropriate NGS platform-specific sequences, ensures that both strands of ligated molecule have the potential to be sequenced. This increases final library yields and by employing dual indices, library multiplexing potential is maximised.

If sufficient input material is used, adapter ligated molecules can be sequenced. However, for some early applications, it was recommended to amplify and enrich for viable library molecules using PCR. Amplification biases have been well reported<sup>45,46</sup>, especially for extremes of GC content, so PCR should be avoided where possible. If required, cycle numbers should be minimised and a suitable *Taq* polymerase such as Kapa HiFi<sup>47</sup> used to maintain library fidelity and complexity.

Recently, methods harnessing the ability of transposases to randomly insert sequence tags into the genome have become popular<sup>48</sup>. Transposases, both fragment the DNA and provide a common sequence to help introduce barcodes in a process called tagmentation. Recommended DNA requirements are 50 ng of input DNA >10 Kbp. By controlling the ratio of DNA to transposase it is possible to control library insert sizes. The more DNA to transposase the larger the insert size. Using the Nextera Tn5 transposase<sup>49</sup>, DNA is fragmented by having 43/ 44 bp adapter sequences inserted within the DNA molecules. These adapters are not ligated to the molecule on both strands so a nick translation step at 70 °C using a non-hot start *Taq* polymerase is introduced ahead of the conventional PCR cycles. Using the inserted sequence as a template to prime off and amplify the genome, PCR primers can be designed to introduce barcodes and sequences that make the final libraries compatible with NGS instruments. A conventional 10 to 16 cycle PCR step provides sufficient library molecules for sequencing.

## 1.2.2 LMP library construction

The LMP library construction methods presented in this thesis have been optimised using the Illumina Nextera LMP kit. Suggested DNA input requirements range from 1 to 4 µg depending whether a gel based size selection is used (4 µg input recommended) and it is suggested that DNA molecular weight is at least 3x the targeted insert size.

The Nextera Tn5 transposase inserts 19 bp biotinylated adapters into the DNA via fragmentation, this is followed by a strand displacement step before a suitable size selection method is employed to recover fragments of the desired size. A DNA ligase circularises molecules overnight before an exonuclease is employed to remove any uncircularised or nicked DNA. Circularised DNA is then fragmented using ultrasonication before molecules containing the biotin labelled adapter junction are enriched for through binding to streptavidin coated magnetic beads. Bound molecules are then processed as described for *in vivo* cloning based paired-end libraries with the necessary amplification minimised.

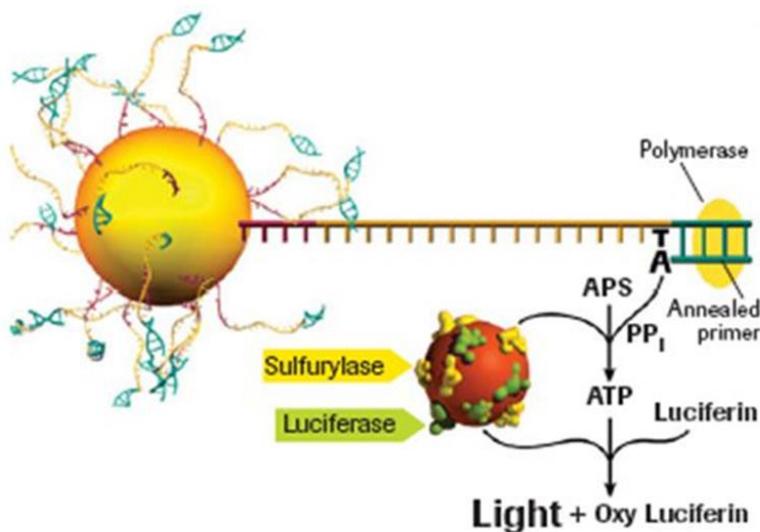
On sequencing, the identification of >25 bp of sequence either side of the 38 bp biotinylated adapter junction molecule is required to distinguish the true LMP reads apart from paired-end reads. Libraries prepared in this manner can suffer from low complexity due to excessive losses during processing, especially in size selection. Low yielding samples typically require more PCR cycles which generates more potential duplicate library molecules. Determining the number of unique reads, therefore, is an important QC step for LMP libraries.

## 1.2.3 Next generation sequencing

### 1.2.3.1 454 Pyrosequencing

The principle of 454 pyrosequencing detection is shown in Figure 1.3. If a nucleotide is incorporated, or a string of nucleotides, then light is emitted which is proportional

to the number of nucleotides added. By flowing nucleotides across in a set order, the sequence can be determined by the presence/ absence and intensity of light.



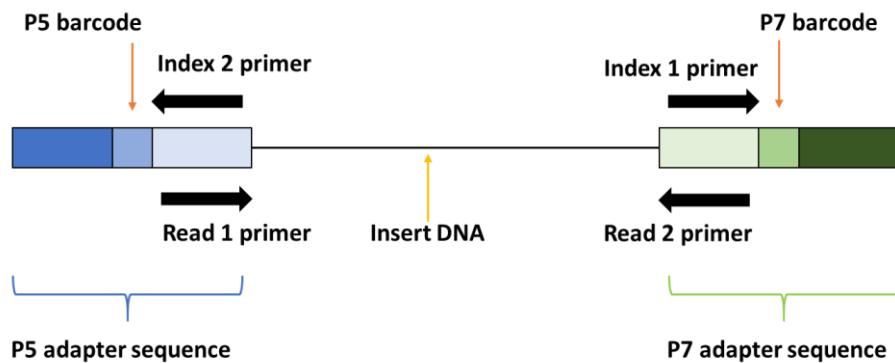
**Figure 1.3:** The basis of 454 pyrosequencing. When a nucleotide gets incorporated into a sequence, sulfurylase converts APS to ATP using the released PP<sub>i</sub> and then Luciferase converts the luciferin and ATP to light and oxy luciferin.

Reproduced from archived 454 promotional material.

Original instruments could sequence up to 100 bp per read and generate up to 25 Mbp per four-hour run. Later modifications included increasing the average read length above 500 bp and generating >500 Mbp per eight-hour run with their FLX instruments and they broke through the 1 Gbp per run barrier with the FLX+ instrument with reads up to 1 Kbp and run times up to 16 hours. Using single-end and LMP reads it was this technology that generated the sequence data for the prokaryotic genomes submitted as part of this thesis<sup>1-4</sup> and these provided a good baseline against which future library construction and genome assembly protocols could be judged. However, the cost of this technology was prohibitive and in 2013 Roche announced that the 454 division would cease trading in 2016.

### 1.2.3.2 Illumina sequencing

The structure of a viable, dual indexed Illumina compatible paired-end library molecule is shown in Figure 1.4 and the principle of their sequencing by synthesis (SBS) technology is shown in Figure 1.5.

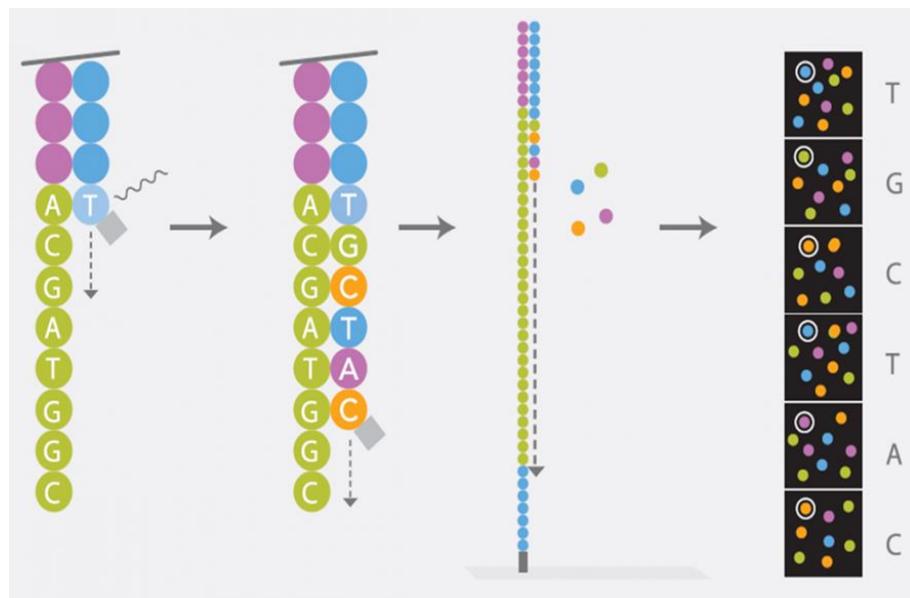


**Figure 1.4:** The structure of an Illumina compatible paired-end library molecule. Viable library molecules have a P5 adapter sequence at one end and a P7 adapter sequence at the other. The 5' end of each adapter enables them to be attached to the oligo lawn of an Illumina flow cell and bridge amplification is used to form clusters. Sequences at the 3' end of each adapter allow for sequencing of the inserted DNA and barcodes. During sequencing, read 1 is generated first followed by index 1, then index 2 and finally read 2.

The technology was initially capable of generating 25 bp of a sequence from a single-end of millions of reads in parallel within a lane of a flow cell and instruments could generate 1 Gbp of sequence data in a week. As read lengths were considerably shorter than the 1 Kbp generated by Sanger sequencing, Illumina instruments are often referred to as short read sequencers.

Each of the four nucleotides has a different colour fluorophore which acts as a reversible chain terminator. Once the nucleotide has been incorporated and the signal read, the fluorophore is then removed and washed away, reverting the nucleotide to

a conventional deoxynucleotide which is receptive to the addition of the next reversible terminator nucleotide.



**Figure 1.5:** The basis of Illumina sequencing. Fluorophore labelled reversible, dideoxynucleotides are added one at a time and the incorporation detected. The fluorophore is then cleaved off making the nucleotide available to be extended and then the next nucleotide can be added and the whole process repeated.

Reproduced from [www.illumina.com](http://www.illumina.com).

Since its launch, there have been rapid improvements in read lengths from ever growing numbers of clusters. Today, sequence reads up to 300 bp can be generated from each end of a library molecule and >1 Tbp of sequence data produced from a single instrument in 3 days. Although newer instruments recommend library insert sizes <400bp, earlier machines had the capability to cluster and sequence libraries with inserts up to 1 Kbp.

The Illumina reads generated are currently the most accurate of the NGS platforms, with accuracy >99.9 %, and it is the cheapest per base pair, with 1 Gbp of data

costing <£50 to generate. As a result, it has quickly become the most widely adopted NGS system in the scientific community.

## 1.3 Genome assembly

Genome assembly is the process of integrating sequence reads to faithfully reconstruct the genome of the sequenced organism. This is usually a two-step process. First paired-end reads are aligned to form contigs, a term first coined by Staden to represent a contiguous stretch of DNA sequence<sup>50</sup>, and then the paired or LMP reads can be used for scaffolding to determine the order of contigs relative to each other.

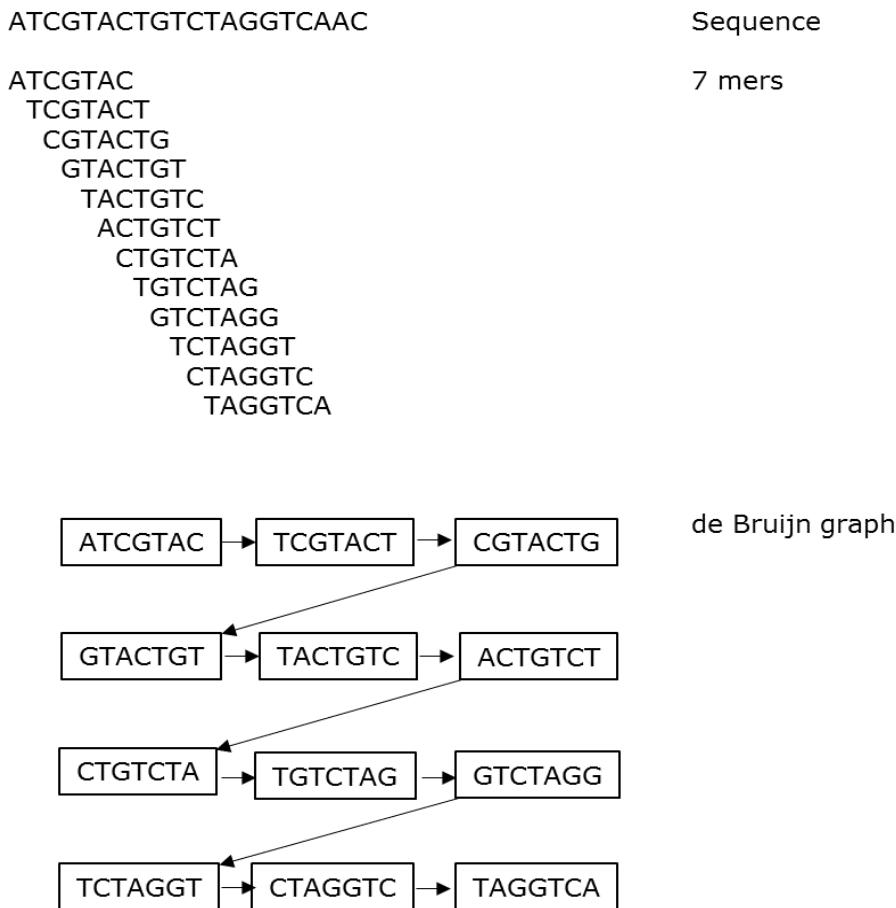
### 1.3.1 Assembly algorithms

With Sanger sequencing reads approaching 1 Kbp in length, assembly programs such as the TIGR<sup>51</sup> and Celera<sup>52</sup> assemblers used algorithms based on consensus overlap to identify reads with shared content. With NGS platforms producing vast quantities of much shorter reads, contigging programs such as ABySS<sup>53</sup> and Velvet<sup>54</sup>, and scaffolding programs such as SOAPdenovo<sup>55</sup>, introduced de Bruijn graphs (DBG)<sup>56,57</sup> due to the vast amounts of data produced and the reduced computing requirements. Today, so called third generation sequencing platforms, such as the Pacific Biosciences (PacBio) RSII and Oxford Nanopore Technology (ONT) MinION, are consistently generating reads >1 Kbp and genome assembly is increasingly returning to the consensus overlaps based on variations of the Celera Assembler<sup>58</sup> with packages such as CANU<sup>59</sup>.

### 1.3.2 Assembling contigs using DBGs

Assemblies based on data generated using protocols presented in this thesis employed DBG assemblers due to the relatively short sequence reads generated on the Illumina platform. DBG assemblers work by slicing reads into all the possible

$k$ -mers of length  $k$  and then uses these to build a DBG. The principle of producing a DBG using  $k$ -mers is shown in Figure 1.6. By overlapping the  $k$ -mers for the last  $k-1$  nucleotides, the path can be determined. Where  $k$  is greater than the size of a repeat then it should be resolved, but if  $k$  is smaller than the size of a repeat, there can be multiple paths in and out of the repeat.



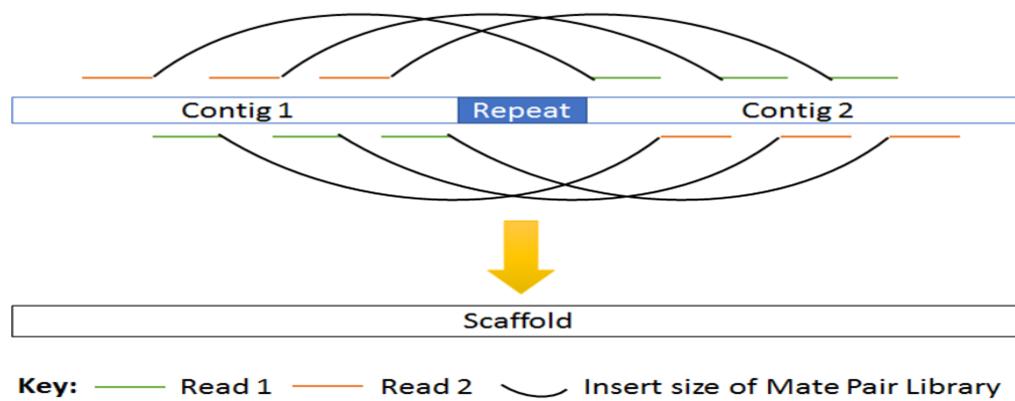
**Figure 1.6:** Using  $k$ -mers to build a de Bruijn graph. A sequence is sliced into all possible 7 mers which are then used to create a directed graph to represent the sequence.

Increasing the length  $k$  can improve specificity and lead to better assemblies. Therefore, Illumina based *de novo* genome assembly projects tend to use longer sequence reads (2x 250 bp) than resequencing projects (2x 150 bp). Paired-end library insert size can also be a factor for improved contiguity and this will be discussed in more detail in Chapter 3.

### 1.3.3 Scaffolding contigs

The principle of scaffolding with LMP reads is shown in Figure 1.7. Using LMP libraries with insert sizes greater than the size of repetitive DNA sequences, it is possible to identify unique sequences a known distance apart and order contigs accordingly. Scaffolding algorithms determine the estimated insert size of the LMP library by identifying paired reads that both map within a single contig. This information is then used to identify paired reads which map across two contigs, and determine the distance between them. It is the presence of multiple LMP reads that connect the same contigs that confirms the spatial relationship.

Scaffolds can consist of any number of contigs and it is the presence of repetitive DNA sequences within the genome which are longer than the insert size of the LMP that prevents further contigs being linked.



**Figure 1.7:** The principle of scaffolding contigs. When LMP libraries are sequenced, reads a defined distance apart can be determined. These are then used to orientate and position contigs relative to each other. Multiple reads mapping helps confirm the link between the contigs.

## 1.4 Genome project quality control

Tools available to assess genome assembly accuracy and contiguity include contig and scaffold N50 (CN50 and SN50), the *K-mer* Analysis Tool (KAT) plots<sup>60</sup> and Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis<sup>61,62</sup>. Together they help validate assemblies and determine genome accuracy and completeness.

### 1.4.1 CN50 and SN50

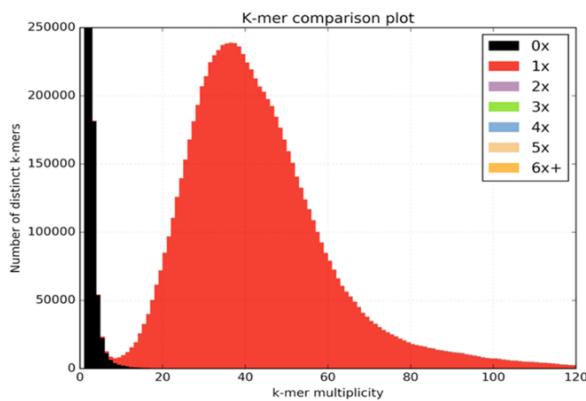
N50 is commonly used to identify the size of the contig/ scaffold above which more than half the genome is assembled. As a measure of contiguity, larger numbers are indicative of better assemblies. However, some scientists disagree as to whether N should be used<sup>63</sup>.

For those that use N50 to describe a length of sequence, they go on to use L50 to describe the number of sequences that it takes for the cumulative length of either contigs or scaffolds to be >50 % of the assembled content. In this scheme N is used to define a length and L a number. This has resulted in several scientists preferring to use L50 for the size, rather than N50, leading to considerable confusion. In this thesis, if L is used in a publication and is given a length in base pairs it will be treated as N and the number will be in bold italics to reflect this.

### 1.4.2 KAT plots

KAT plots are an efficient way to determine how accurate an assembly is and can help identify sequence biases and contaminants. Using the *k-mer* content of the paired-end reads they can be searched for within the assembly. Reads absent from the assembly are characterised by black sections below the main red peak and sequencing errors by a black peak along the y-axis of the graph. Red peaks along the y-axis represent *k-mers* in the assembly but not in the reads and the main red peak, the paired-end *k-mer* content that appears once within the assembly. Peaks that are neither red or black represent duplications which are either true, to the right of the

main peak at twice or greater the multiplicity of the main peak, or down to duplications in the assembly, above the main red peak. A KAT plot for a paired-end only assembly of the *S. coelicolor* M145 genome using data from an amplification-free paired-end library sequenced with 2 x250 bp read and using a  $k$  value of 200 with 37x coverage is shown in Figure 1.8. This assembly had a CN50 of 288kb in 70 contigs >500 bp.



**Figure 1.8:** A KAT plot of a *S. coelicolor* assembly. The graph shows a single red peak with no black underneath it. There is also no red to the left of the black peak on the y-axis. These are key indicators of a good assembly.

### 1.4.3 BUSCO analysis

BUSCO analysis measures the completeness of genome assembly based on the expected gene content. Single copy orthologs present in at least 90 % of the species are searched for within an assembly. Assembly accuracy is determined based on which genes are reported to be complete, duplicated, fragmented or missing and can be used to make informed decisions about potential sequencing or assembly errors. Characteristics of a good assembly include >90 % of the genes being complete and <1 % of the genes being fragmented. For polyploid genomes such as wheat, the presence of many homeologous genes will mean that a high proportion of the single copy orthologs will appear duplicated.

## 1.5 Genome project strategies

Several strategies can be adopted to improve contiguity for repeat rich genomes. Reducing genome complexity can be achieved through partitioning of the genome into smaller chunks by flow sorting chromosomes or utilising Bacterial Artificial Chromosome (BAC) libraries and these strategies are discussed in Chapters 2, 5 and 6. Maintaining library complexity and controlling the size distribution of molecules within a library can be achieved by minimising the need to amplify material and optimising fragmentation and size selection, these attributes are discussed in Chapters 2 through 6.

These approaches have the potential to make assembly a much simpler task. They reduce variability in what is being assembled and controlling this spatial information is a great asset to genome assemblers. It helps hone algorithms and makes mathematical modelling more straightforward. Knowing the potential insert size distribution between two reads can help reduce the number of undetermined bases in an assembly and helps connect sequences more accurately thus providing greater contiguity.

## 1.6 Summary

From humble beginnings, when it required several individual reactions and multiple sequencing runs to complete even the simplest genome, we are at a point when even the most complex genomes can be sequenced and assembled to high degree of contiguity within a matter of weeks. The current rate of advancement promises much and positions science to enter the pangenomic era for even the most challenging of genome projects.

## 2 Increasing sequence contiguity in barley by decomplexing the genome

During 2017 there were several high-profile publications of highly repetitive, grass genomes each showing ever increasing contiguity. These include publication of the most contiguous barley genome presented to date<sup>8</sup> and its sister publication detailing the methods employed<sup>9</sup>. Both are submitted as part of this thesis. My role was to develop a custom, low cost, high-throughput BAC sequencing pipeline: to culture, extract DNA and construct NGS compatible paired-end libraries from individual BACs and LMP libraries from pools of 384 BACs.

In this chapter I outline some of the challenges associated with sequencing BACs and highly repetitive genomes, describing how my approach overcame these. By comparing it against other laboratory based strategies, I highlight the benefits and limitations these strategies bring to genome projects.

The sequence assemblies discussed in this chapter were generated at EI by Dharanya Sampath (barley BACs) and Jon Wright (wheat BACs and barley Whole Genome Sequencing (WGS)) and the 434 unique 9 mer barcodes designed by Matt Clark.

### 2.1 Bacterial Artificial Chromosomes (BACs)

BAC clones were developed to amplify up to 300 Kbp of DNA allowing scientists to work on specific chromosomal regions of interest<sup>64</sup>. Starting with high molecular weight DNA >400 Kbp, partial digests using restriction enzymes increases the chances that inserts within a given BAC would overlap with inserts of other clones. These restriction digested fragments were separated on an agarose gel and bands cut out targeting molecules >100 Kbp and the DNA recovered.

Fragmented, size selected DNA molecules were cloned into a vector which consists of

- i) the sequence necessary for replication within a host bacterial cell, usually *E. coli*,
- ii) an antibiotic resistance gene, usually chloramphenicol, allowing for this to be used

as a selectable marker and iii) the *F* factor sequence from *E. coli* which ensured that they appeared as single copy within the host cell. BAC clones were then electro or chemically introduced into a competent host cell, usually *E. coli* DH10B. A suitable titre was used to generate sufficient single, discernible clones which were then picked into glycerol stocks. Based on average insert sizes of 130 Kbp, 12x coverage of the genome is targeted, and theoretically, this would result in >99 % of the genome being present within the library.

BAC DNA can be digested with a suitable restriction enzyme to produce a fingerprint of the clone insert consisting of different size DNA fragments which can be separated by agarose gel electrophoresis<sup>65</sup>. By comparing the fragment patterns BACs sharing common, multiple different sized bands are deemed to contain overlapping inserts. This information can then be used to produce a Minimal Tile Path (MTP) which would contain the fewest number of BACs to cover the whole genome. This approach was used for the publicly funded HGP, using primer walking to sequence chromosome anchored BACs in their entirety.

## 2.2 Barley

Barley has an estimated genome size of 5.1 Gbp and like many grass species has a high proportion of dispersed repeats with an estimated 75.33 % of its genome being Class I retrotransposons and 5.6 % Class II DNA transposons. In 2012 the International Barley Sequencing Consortium (IBSC) generated 55x coverage of a PCR amplified, 500 bp average insert paired-end library adding 2.5 Kbp insert LMP data. They reported a SN50 of 1.4 Kbp<sup>66</sup>. A CN50 of 1.5 Kbp was later achieved by Sanchez-Martin *et al.* when using flow sorting to isolate and then sequence chromosome 2H of the barley cultivar Forma. They generated 10x coverage using an amplification-free paired-end library with an average insert size of 500 bp<sup>67</sup>. Both studies highlighted the lack of contiguity in WGS approaches to sequencing barley.

With the level of repetitive DNA and lack of contiguity from WGS projects, BACs were considered the ideal vehicle to deconvolute and sequence the barley genome. Increasing outputs and improved barcoding capabilities created new opportunities for whole BAC sequencing on NGS instruments.

## 2.3 Alternative strategies to sequence barley BACs

### 2.3.1 Different pooling strategies for sequencing barley BACs

In 2006, an early NGS based study reporting the 454 pyrosequencing of barley BACs was published<sup>68</sup>. Using 100 bp single-end reads, and comparing the outputs against Sanger sequencing, they highlighted the benefit of NGS approaches in assembling the gene space but the presence of repetitive DNA sequences hindered contiguity. Of the four BACs sequenced, the best assembly required >50x coverage and contained 65 contigs.

As 454 read length increased, strategies to individually barcode BACs were developed which showed improved contiguity<sup>69</sup>. Pooling 48 non-overlapping BAC clones and generating 200 bp+ reads to an average of 26x coverage, Steuernagel *et al.* achieved an average CN50 of 48 Kbp with fewer than 10 contigs per BAC. Although the assembly metrics were impressive, library construction and sequencing costs >£250 per BAC made this approach unviable for screening the 85,000+ BACs in the Barley MTP.

BAC pooling strategies were developed further by Lonardi *et al.* to increase throughput and decrease costs<sup>70</sup>. They used a shifted transversal design<sup>71</sup> to pool BACs that formed contigs in the physical map for barley and sequenced them on the higher throughput Illumina instruments. Targeting an average 150x coverage to ensure each of the BACs within the pool had at least 50x coverage, they sequenced with 2x 100 bp reads. By deconvoluting the pooling, they could achieve single BAC resolution and obtained CN50s between 5.8 and 8.1 Kbp depending on coverage. With no deconvolution, they achieved a CN50 of 4.2 Kbp for 169 BACs and 3.8 Kbp for 2,197 BACs. They also attempted WGS of barley with paired-end and 2, 3 and 5 Kbp insert LMPs, improving the SN50 to 2.8 Kbp.

### 2.3.2 Targeted approaches to sequencing barley BACs

The IBSC used 454FLX reads and 2x 100 bp Illumina reads to sequence 5,341 gene rich BACs to supplement their physical map and 937 random clones using the 454FLX<sup>66</sup>. Starting with 1ml cultures and using conventional *in vivo* cloning techniques to construct paired-end libraries, up to 67 BACs were pooled and size selected by cutting bands out of agarose gels and viable library molecules enriched for by performing 10 cycles of PCR ahead of sequencing.

Munoz-Amatriain *et al.* sequenced 15,711 gene bearing BACs from the barley library and adopted the shifted transverse design, generating assemblies for 15,622 BACs<sup>72</sup>. Taking 2,197 BACs at a time they generated 169 BAC pools with 13 pools per layer and 7 layers. This reduced the number of libraries constructed down to 637. Traditional alkali lysis based DNA extractions were employed with DNA pools created by hand. Illumina library construction methods were used and on sequencing, >40 % of the data was shown to be contaminating host *E. coli* DNA which was filtered out ahead of assembly.

The Leibniz Institute on Aging—Fritz Lipmann Institute (FLI) and Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) sequenced MTPs of barley chromosomes 1H, 3H and 4H and Beijing Genomics Institute (BGI) sequenced chromosomes 5H, 6H and 7H. Some BACs had been previously sequenced using a combination of different 454 and Illumina library construction protocols, and many of these were not repeated. For the remaining BACs they adopted the BAC culturing and Illumina library construction strategy from the 2012 IBSC paper adding LMP data with 10 and 20 Kbp inserts.

Assembly metrics for these targeted approaches to sequencing barley BACs are shown in Table 2.1.

Publication	Number of BACs	Sequencing Strategy	Average Contigs per BAC	CN50 (Kbp)	SN50 (Kbp)
<b>IBSC 2012<sup>66</sup></b>	3,158	454 single-end	21	30.9	-
	937	454 single-end	20	43.0	-
	2,183	Illumina paired-end	31	6.8	-
<b>Munoz-Amatriain<sup>72</sup></b>	15,622	Illumina paired-end	20	23.9	-
<b>IBSC 2017<sup>8</sup></b>	69,761	Illumina paired-end + LMP	n/a	n/a	82.3
<b>EI 2017<sup>8</sup></b>	17,317	Illumina paired-end + LMP	24	16.9	95.3

**Table 2.1:** Assembly metrics for targeted approaches to sequencing barley BACs.

## 2.4 Development of a novel BAC sequencing pipeline

Working on the principle that 1 ng of BAC DNA with a 135 Kbp inserts equates >6.5 million copies, I developed a novel BAC sequencing pipeline focussing on low input library construction to reduce costs and increase throughputs.

### 2.4.1 DNA extraction

Traditional methods to optimise BAC DNA extraction involve measuring turbidity at 600 nm over time to identify when cells enter the lag phase of growth. This information is then used to harvest cells when they are still in the exponential phase of growth to improve yield and DNA quality. When processing thousands of BACs simultaneously with different size inserts, this is not feasible.

Working in 384 well format, I grew clones on LB agar supplemented with chloramphenicol to confirm clone viability. I then optimised culture volumes,

incubation times and DNA extraction protocols based on the alkali-lysis method of Beckman Coulter's CosMC beads<sup>73</sup> and wrote bespoke programs on a 96-tip head Beckman Coulter FX<sup>P</sup> liquid handling instrument. My initial experiments involved evaluating incubation times between 15 and 24 hours, miniaturising reaction volumes, switching between 96 and 384 well cultures and reusing tips by employing hydrogen peroxide and water washes to denature and remove any DNA which could potentially cross-contaminate other samples.

Using quantitative PCR<sup>74-76</sup> (qPCR), I developed bespoke assays to determine copy number of the pIndigo-BAC5 vector and DH10B *E. coli* in extracted DNA. QPCR works by detecting the synthesis of DNA using double strand specific intercalating dyes such as SYBR Green. By measuring the background fluorescence over the first five PCR cycles, a threshold value is determined. The point at which fluorescence is detected above this threshold value is calculated and is known as the Ct value. With each cycle theoretically doubling the amount of double stranded DNA product, a difference in Ct value of 1 represents a copy number difference of two. For a tenfold difference in copy number the difference in Ct value would be 3.3.

Using the primers shown in Table 2.2, I generated amplicon specific standards ranging from 2e<sup>3</sup> and 2e<sup>8</sup> molecules/  $\mu$ l for each assay.

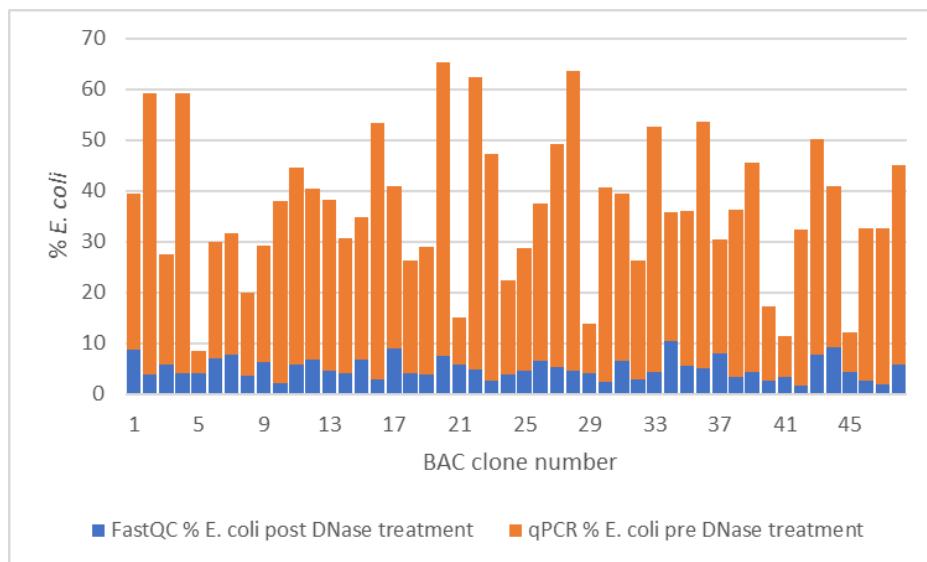
Primer	Sequence
<b><i>E. coli</i> Forward</b>	CTGAACTGTGGCTCAGCAAA
<b><i>E. coli</i> Reverse</b>	CGCTCAAGGGAAAGGTTAT
<b>pIndigo-BAC 5 Forward</b>	TAGAAACTGCCGGAAATCGT
<b>pIndigo-BAC 5 Reverse</b>	TCCGGCCTTATTACACATTC

**Table 2.2:** Primer sequences for qPCR assays for pIndigo-BAC5 and *E. coli* DH10B.

For the qPCR assay, I combined 10  $\mu$ l of the Kapa Biosystems 2x qPCR master mix with 1  $\mu$ l of 10  $\mu$ M forward primer, 1  $\mu$ l of 10  $\mu$ M reverse primer, 1  $\mu$ l of standard or

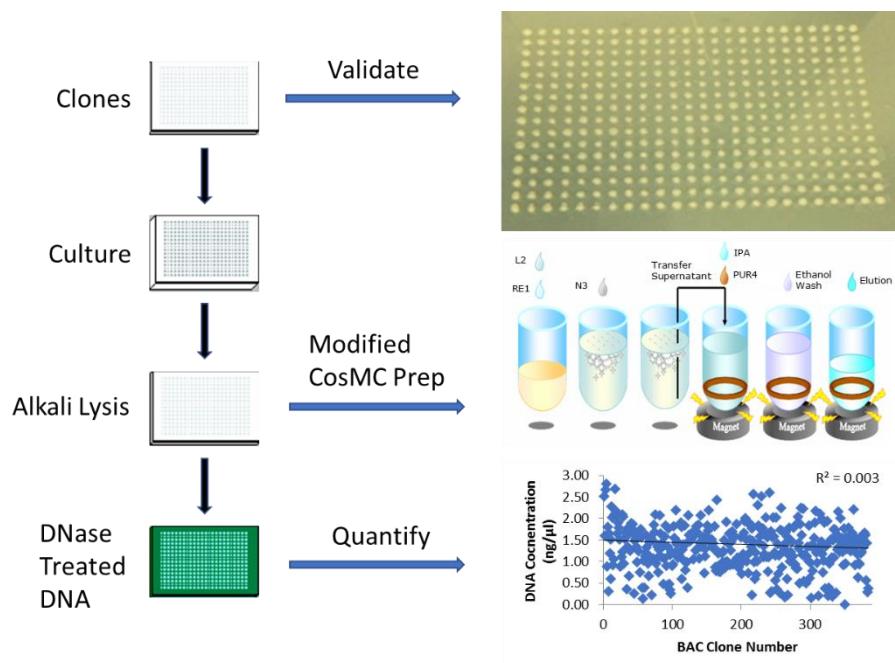
BAC extracted DNA and 7  $\mu$ l of water. The reactions were incubated for 5 minutes at 95  $^{\circ}$ C followed by 40 cycles of 45 seconds at 95  $^{\circ}$ C and 30 seconds at 60  $^{\circ}$ C on an Applied Biosystems Step One qPCR instrument. The Ct values for the known copy number samples were used to generate the standard curves and then the copy number for the unknown samples calculated by comparing their Ct values against these.

This revealed over 50 % of DNA was *E. coli* in some extractions and highlighted the need to employ an ATP dependent DNase to remove the contaminating host DNA. To check for host contamination, FastQC<sup>77</sup> was adapted to screen for *E. coli* DH10B during post-Illumina run output analysis. The percentage of *E. coli* in 48 BAC DNA extractions, as measured by qPCR before DNase treatment, and, as measured by FastQC after treatment and sequencing, is shown in Figure 2.1. For these BACs the average contamination before DNase treatment was 31 % and after treatment 5%. This confirmed the benefit of the DNase treatment in reducing the *E. coli* levels which would help maximise sequence coverage of the target BACs.



**Figure 2.1:** The percentage of contaminating host *E. coli* DNA present in 48 BAC DNA extractions before DNase treatment, as determined by qPCR (orange) and after DNase treatment, as determined by FastQC (blue).

An overview of the BAC DNA extraction pipeline is shown in Figure 2.2.



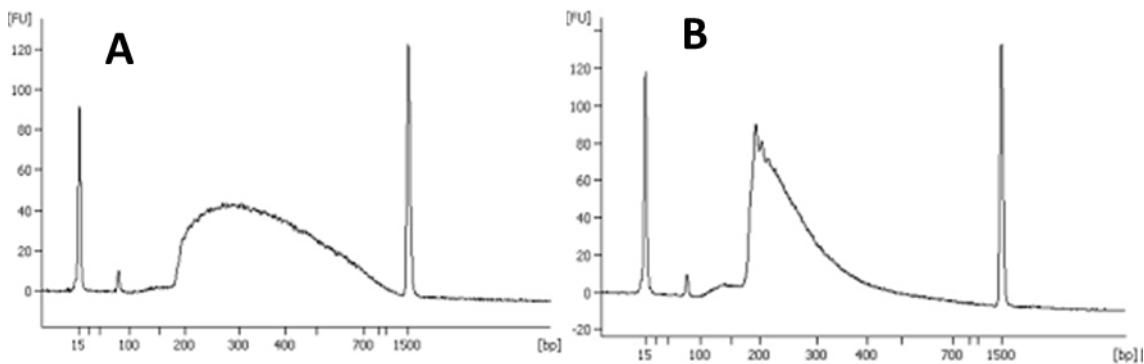
**Figure 2.2:** An overview of the high-throughput BAC DNA extraction pipeline.

CosMC prep figure reproduced from [www.beckmancoulter.com](http://www.beckmancoulter.com)

By the end of development, I could extract DNA from >4,000 BAC clones using a standard 384 CosMC prep reaction kit. I used individual tips for each clone, 384 well plates as solution reservoirs and I could complete BAC replication, culture and DNA extraction for a full economic cost (FEC) <50p per BAC when processing 2,304 clones per day. Over 85 % of the BACs tested had DNA yields between 0.5 and 2 ng/ µl in 20 µl and based on fingerprint data, the estimated insert sizes for the BACs in the barley MTP ranged from 80 to >200 Kbp. Obtaining consistent yields across hundreds of BACs confirmed that my DNA extraction pipeline was robust and reproducible and that my optimised conditions could be used independent of insert size.

## 2.4.2 BAC paired-end library construction

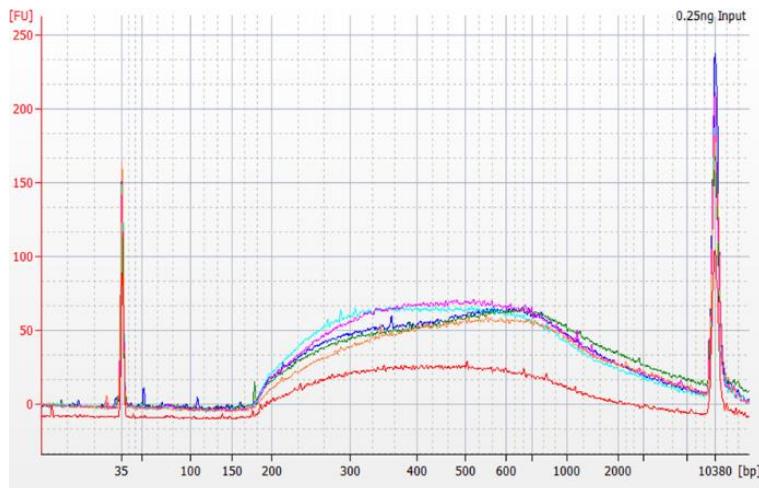
To overcome the need to measure and normalise DNA concentrations, I needed a library construction protocol which could tolerate varying input amounts and produce similar library profiles. I chose Epicentres' transposase based Nextera library construction kit for its simple workflow and ease of automation. The library Bioanalyzer electropherograms when using the manufacturer supplied buffer systems are shown in Figure 2.3. Library profiles were biased towards fragments <300 bp with very few molecules >400 bp limiting the spatial potential of the libraries.



**Figure 2.3:** Bioanalyzer electropherograms of Nextera libraries using the manufacturer provided buffers. Agilent Bioanalyzer traces of Nextera libraries constructed with the supplied HMW (A) and LMW (B) buffers.

Reproduced from the archived Epicentre User Guide.

I reworked and optimised the reaction buffers and volumes and titrated the DNA to enzyme ratios and generated the library electropherograms shown in Figure 2.4. This showed that consistent library profiles could be achieved with a range of DNA inputs from 0.25 to 2 ng. These libraries, with molecules spanning 200 bp to 1 Kbp, helped future proof the method allowing for larger molecules to be isolated as sequence read lengths increased. To maximise the multiplexing capability of the pipeline, 434 unique 9 mer barcodes were designed with a Hamming distance of 4 bp.



**Figure 2.4:** The effect of varying DNA input on library size profiles in the optimised Nextera based library construction protocol. Libraries had consistent wide insert size distributions from 200 bp to 1 Kbp when inputting 0.25 ng (red), 0.5 ng (blue), 1 ng (green), 1.5 ng (turquoise), 2 ng (pink) and 2.5 ng (orange).

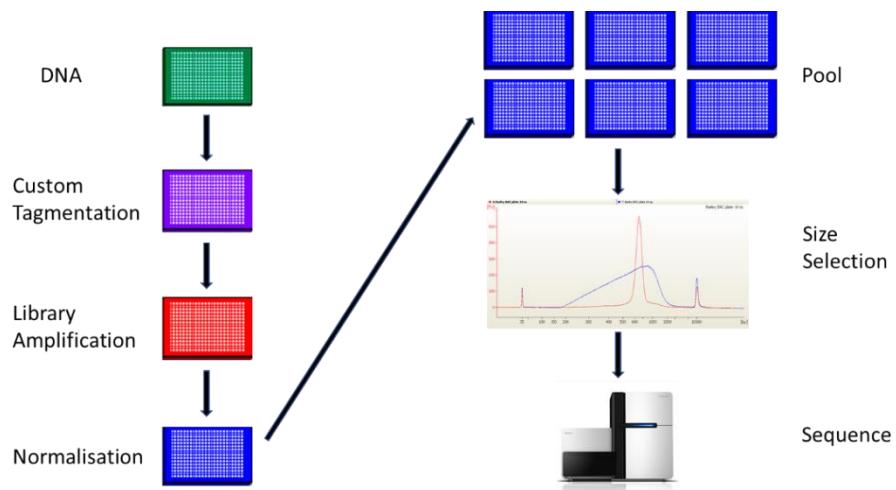
Post tagmentation, QIAGEN buffer PB was used to inactivate any remaining transposases and a bead based purification step employed ahead of library construction. Using the Epicentre Nextera kit, 384 libraries with different P7 barcodes could be multiplexed on a single HiSeq2500 lane and when Illumina acquired Epicentre, and brought out their own version of the Nextera kit, it facilitated dual indexing. Using 48 barcoded P5 adapters and 48 barcoded P7 adapters, 2,304 paired-end libraries could be pooled per lane. I normalised libraries using MagQuant beads, then pooled and concentrated them. I then size selected them on the BluePippin to recover molecules between 400 and 600 bp.

After sequencing, cross contamination was determined by looking for the presence of sequence from more than one BAC for a given barcode combination. If  $>10\%$  of the reads indicated the presence of a neighbouring clone, I cherry picked the BAC from its original plate and re-arrayed it into a new plate creating a new glycerol stock. I then repeated culturing, DNA extraction and constructed a new paired-end library. Overall  $<15\%$  of clones failed first round library construction of which  $>90\%$  was due to insufficient sequence data. Of these,  $>90\%$  passed QC, generated enough

data and could be assembled when repeated.

When the library construction pipeline was fully optimised, I could construct >1,000 libraries from a standard 24 reaction Nextera kit and an Illumina compatible paired-end library could be constructed and sequenced for <£3 FEC per BAC.

An overview of the BAC paired-end library construction, normalisation, pooling and sequencing pipeline is shown in Figure 2.5.



**Figure 2.5:** Overview of the BAC paired-end library construction and sequencing pipeline.

### 2.4.3 BAC LMP library construction

To help resolve larger repetitive DNA sequences, I constructed LMP libraries using the BluePippin to target insert sizes between 6 and 8 Kbp. These had the potential to resolve most LINEs and some smaller LTRs and DNA transposons.

I constructed LMP libraries from pools of 384 BACs and I optimised culturing to ensure that BACs were present at as even a concentration as possible. I performed higher quality DNA extractions using QIAGEN's large construct kit to try and maximise DNA quality and improve LMP library complexity. I achieved cost savings by constructing eight LMP libraries over two days, reducing DNA inputs and reaction volumes, and

the eight pools of 384 BAC LMP libraries were multiplexed and sequenced with 2x 150 bp reads on a HiSeq2500. Paired-end data was used to confirm which reads came from which BAC and this information used for scaffolding the appropriate BAC. Using my optimised protocol, I could culture, extract DNA and construction and sequence the BAC pool LMP libraries for <£2 FEC per BAC.

#### **2.4.4 Additional developments to the BAC sequencing pipeline**

Following completion of the barley BAC project, further modifications were made to improve the published paired-end library construction aspect of the pipeline. Up to 20 % of the paired-end reads generated for barley BACs were PCR duplicates. DNA losses associated with buffer PB treatment and bead-based purification meant 21 cycles of PCR were required to obtain sufficient library molecules for sequencing. This contributed toward the high duplication rate observed. I discovered that Robust 2G *Taq* polymerase tolerated the transposase buffer system and the heat denaturation at the beginning of the PCR inactivated any remaining transposase. This meant I could bypass the buffer PB treatment and the bead-based purification. I then reduced the PCR to 14 cycles and generated comparable library yields. The increased library complexity meant that I could also omit the normalisation step and although coverage was more variable, this was offset by <5 % duplication rates. Using this revised method, I could construct paired-end libraries for <£2.50 FEC per BAC.

For the revised paired-end library construction method, I combined 1  $\mu$ l of BAC DNA to 0.9  $\mu$ l of Nextera reaction buffer, 0.1  $\mu$ l of Nextera enzyme and 2  $\mu$ l of water and incubated this at 55 °C for 10 minutes. I added 2  $\mu$ l of 2.5  $\mu$ M forward and reverse primers followed by a master mix containing 5  $\mu$ l of 5x Kapa Biosystems Robust 2g reaction buffer, 0.5  $\mu$ l 10mM dNTPs, 0.125  $\mu$ l Kapa Biosystems Robust 2g *Taq* Polymerase and 10.375  $\mu$ l of water per reaction. This was then incubated at 72°C for 3 minutes follow by 95 °C for 3 minutes and then 14 cycles of 95 °C for 10 seconds, 62 °C for 20 seconds and 72 °C for 2 minutes 30 seconds. I pooled libraries by spinning the contents of the 384 well PCR plate in to a 96-pipette box tip lid in a plate centrifuge at 1,000 rpm and then pooled, concentrated and size selected the libraries as outlined in the barley genome paper.

Using both these improvements, I constructed paired-end libraries for a MTP for the long arm of CS42 chromosome 3D (3DL) consisting of 6,144 clones. This resulted in a CN50 of 13.5 Kbp, SN50 of 90 Kbp with an average of 19 contigs per BAC and first round failure rates were reduced to <5 %.

#### **2.4.5 Outputs from the BAC sequencing pipeline**

Using my paired-end and LMP approaches, a total of 17,317 barley BACs from the MTP for 2H and 0H were sequenced and assembly metrics are shown in Table 2.1. Single scaffolds were achieved for >25 % of the BACs and >75 % contained <4 scaffolds.

### **2.5 Comparing the different barley BAC sequencing strategies**

#### **2.5.1 Gene bearing versus WGS barley BACs**

Comparing the barley 2H assemblies against those generated for the barley BACs sequenced in the IBSC 2012 publication highlights the benefit of 500 bp versus 300 bp inserts for Illumina libraries. The improved physical coverage and ability to resolve more repeats provided by the larger insert libraries resulted in a near 2.5-fold improvement in CN50 and a third less contigs per BAC. I discuss the benefit of maximising paired-end library insert size in more detail in Chapter 3.

As gene rich BACs contain more unique sequence, and longer 454 reads are usually easier to assemble than shorter Illumina reads, you would expect higher CN50s for gene rich BACs sequenced on the 454. Interestingly, it is the random BACs sequenced by 454 that have the highest CN50. This helped confirm that sequencing barley BAC by BAC was a sensible decision. It suggested that many individual barley BACs were unlikely to contain multiple copies of the same repeat, so a suitable BAC by BAC strategy should significantly improve genome contiguity.

Although the pooling strategy adopted by Munoz-Amatriain *et al.* was innovative and produced a 7 Kbp improvement in CN50 over the 2H assembly, the FEC for EI to have replicated this paired-end only study would be £3.26 per BAC for library construction and £1.02 for sequencing. Significant additional costs would have been required for DNA extraction and pooling. By contrast, the ability to dual index and streamline the process using my approaches, DNA extraction, paired-end and LMP library construction, pooling and sequencing could be achieved for <£5 FEC. Including LMP data resulted in a greater than threefold increase in contiguity highlighting the benefit of my approach over the paired-end only shifted transverse design.

## **2.5.2 Comparison of the 2017 approaches**

Of the 9,061 BACs for chromosome 2H and 8,256 for 0H, useable sequence was generated for 8,969 (99 %) and 8,031 (97.3 %) respectively with 8,195 (91.4 %) and 6,714 (83.6 %) anchored within the POPSEQ map<sup>78</sup>. For the remaining barley chromosome BACs, 97.8 % produced useable sequence and 90.4 % could be anchored.

The 2H and 0H assemblies had a 15 % improvement in SN50 over the other barley chromosomes with the main difference between the two LMP strategies being the use of smaller library inserts for my approach. Although my LMP strategy would not resolve repeats >8 Kbp, it did ensure that inserts were smaller than the size of the pIndigo-BAC5 backbone so that no LMP reads spanned the vector and suggested inappropriate linkage. It is also likely that the smaller inserts were more complex, containing less duplicated reads, and would therefore be more informative. LMP library complexity is discussed in more detail in Chapter 4.

## **2.6 Summary**

As a protocol my high throughput, low cost, scalable BAC sequencing pipeline delivered. It is testament to the pipeline that after sequencing the MTPs assigned to each of the seven barley chromosomes, the IBSC chose my methods to sequence the clones that formed contigs in the physical map but couldn't be assigned to a

chromosome (0H) over the other approaches. Combined with data generated by other institutes, and using the physical map to underpin optical mapping, this led to a super scaffold N50 of 1.9 Mbp being achieved with 80.8 % of the transposable elements being resolved, the most contiguous barley genome sequenced to date. At the end of its development, my pipeline had a throughput of 9,216 BACs per day for DNA extraction and paired-end library construction and 1,536 BACs per day for LMP library construction. With sequencing to a combined average of 200x coverage, FEC was <£5 per BAC. None of the other strategies discussed in this chapter could compete in terms of throughput, cost or contiguity.

The protocol went on to successfully sequence 100,000 random wheat clones and 40,000 rye grass (*L. perenne*) clones showing that it was robust across different grass species. It was also used to generate sequence data from 96 wheat<sup>7</sup> and 96 potato<sup>10</sup> clones to help validate sequence assemblies.

Since completing the barley BACs project, I have sequenced a barley cultivar using the whole genome, amplification-free, paired-end and LMP protocols presented in Chapters 3 and 4 which resulted in a CN50 >22 Kbp and SN50 >86 Kbp. This data was generated for a cost of <£40,000 whereas sequencing the entire barley MTP using my BAC pipeline would cost >£430,000. The difference in costs whilst achieving comparable contiguity has effectively ended the need to sequence BACs as part of a genome project.

Although the DNA extraction element of the pipeline may be confined to the history books, I later refined the library construction aspect to produce Low Input, Transposase Enabled (LITE) libraries which have shown great promise for low cost resequencing projects for a variety of different size genomes and amplicons. It is tuneable to genome size and can construct an Illumina compatible library for <£5 FEC. It has supported successful GCRF grant applications helping generate sequence data for large collections of wheat, salmonella, sugar cane, red clover, tilapia and rye grass and publications on *Pseudomonas*<sup>79</sup> and yeast have recently been submitted with many more expected.

### 3 Optimising NGS paired-end library insert sizes.

In this chapter, I outline how NGS shotgun library construction has evolved over the last 9 years and describe some of the unique modifications I have made. In establishing a 454 pyrosequencing pipeline at EI, I demonstrated the benefit of size selected libraries on sequence outputs. Combining this knowledge with the benefit of paired-end libraries and the advantage of amplification-free libraries led to the development of several improvements in optimising NGS paired-end library insert sizes.

To increase the insert size and robustness of NGS paired-end libraries I developed a novel Illumina compatible, amplification-free paired-end library construction protocol, Tight, Amplification-free, Large-insert Libraries (TALL). As Illumina read length increased, I adopted the wider insert spanning, amplification-free libraries developed at the Broad Institute which fed into their DISCOVAR assembler. I later modified these to improve library characteristics and used these to sequence the European polecat (*M. putorius*). Finally, I created a hybrid of the TALL and improved DISCOVAR libraries, Size Exclusion-Amplification-free, Paired-end (SE-APE) libraries, designed to maximise spatial potential and improve the resultant *de novo* genome assemblies. Each of these paired-end libraries has underpinned development of the W<sup>2</sup>RAP *de novo* genome assembler.

By comparing my protocols against other library construction strategies, I highlight the benefits and limitations they offer. This work is supported by publications on the critical comparison of technologies in sequencing *S. verrucosum*<sup>10</sup> and the W<sup>2</sup>RAP assembler<sup>6</sup> which are both submitted as part of this thesis.

The genome assemblies and KAT plots discussed in this chapter were generated at EI by Bernardo Clavijo (diatom), Pirita Paajanen and George Kettleborough (potato) and Graham Etherington (polecat). The development of the W<sup>2</sup>RAP algorithms at EI was undertaken by Bernardo Clavijo, Gonza Garcia-Accinelli and Jon Wright.

## 3.1 NGS shotgun library construction

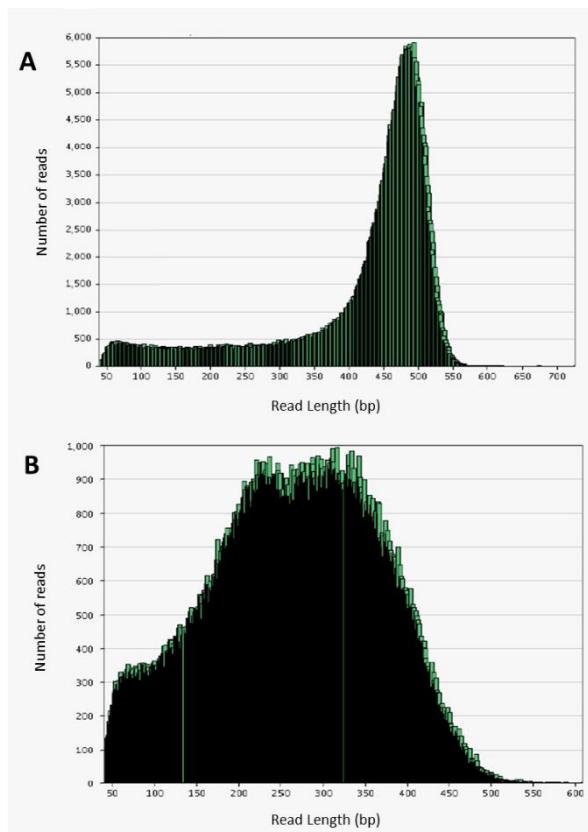
### 3.1.1 The benefits of controlling NGS library insert sizes

The 454FLX instrument could generate up to 500 Mbp from a single run and although achieving >1 million single-end reads was relatively straightforward, short read lengths often resulted in reduced yields. To investigate the effect of insert size on sequence outputs I constructed two libraries, one with molecules size selected at 600 bp +/-10 % on a Perkin Elmer LabChipXT and a second using the standard Solid Phase Reversible Immobilisation<sup>80,81</sup> (SPRI) bead based size selection recommended by 454. I sequenced these on the 454FLX and sequence outputs for these are shown in Table 3.1 and Figure 3.1.

Library	Average Read Length (bp)	Median Read Length (bp)	Average Quality
Standard	259 +/-100	268	30
Size Selected	415 +/-116	461	32

**Table 3.1:** 454 Pyrosequencing sequence outputs. Read length and quality statistics for the size selected and standard libraries.

Targeting 600 bp fragments for library construction resulted in a 60 % increase in average read length, a 72 % increase in median read length and an improvement in average quality over non-size selected libraries. This highlighted the benefits of controlling insert sizes and laid the foundation for the size selection strategies discussed later in this chapter.



**Figure 3.1:** 454 Pyrosequencing read length distribution plots. The effect of isolating molecules using the Perkin Elmer LabChipXT targeting 600 bp +/- 10 % (A) molecules and using the standard size selection method (B) ahead of library construction. These were then sequenced on the same 454FLX run and read lengths determined and plotted.

### 3.1.2 The benefits of read length and paired-end sequencing

While industry standard approaches to paired-end libraries target insert sizes <500 bp, these do not make full use of spatial potential of the Illumina instruments. A single-end 18 bp read can resolve 97 % of the *E.coli* genome and increasing this to 475 bp read resolves 99 %<sup>82</sup>. In contrast, when using the spatial information provided by a 300 bp fragment, 97.4 % of *E. coli* is resolved by an 8 bp paired-end read, an effective 11 % decrease in the information required<sup>83</sup>. Optimising paired-end library insert size and maximising the spatial information they provide has the potential to

resolve all SINEs, more LINEs and more LTRs which can result in significant improvements in contiguity. The non-overlapping reads would also reduce coverage requirements and lower costs.

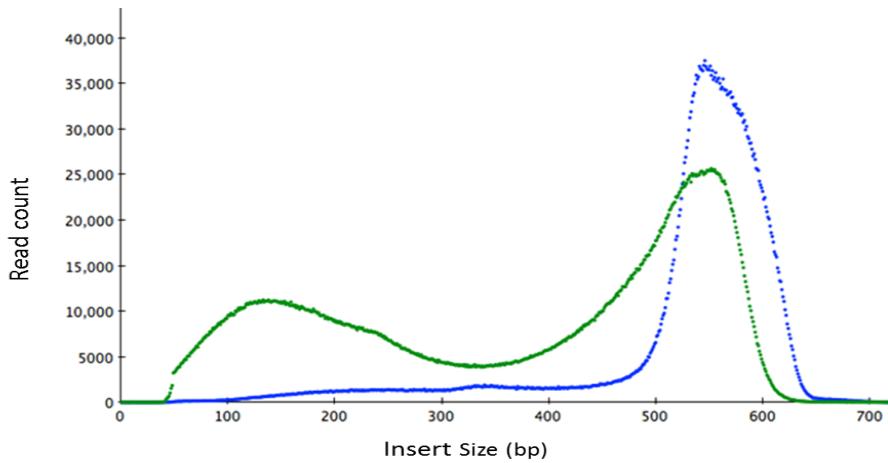
### **3.1.3 The benefit of amplification-free, paired-end libraries**

PCR was first described in 1986 to clonally amplify beta albumin and HLA-DQ alpha DNA and it went on to revolutionise the field of molecular biology<sup>84-86</sup>. When applied to whole genome shotgun libraries to enrich for viable library molecules within Illumina library construction protocols, it became apparent that not all areas of the genome were covered to the same extent.

Studies using different amplicon combinations to represent GC contents ranging from 6 to 90 % showed that there was a distinct drop in representation of molecules >50 % GC after amplification<sup>87</sup>. Reports also acknowledged the effect of amplification biases in Illumina library construction, highlighting the benefit of constructing amplification-free libraries<sup>45,46,88,89</sup>.

I also observed how amplification combined with size selection using E-gels compromised paired-end libraries. Working on the diatom *E. Huxleyi*, my EI colleague Meena Assini constructed a standard PCR amplified Illumina paired-end library, with E-gel size selection, and I constructed an amplification-free paired-end library, using the LabChipXT.

We both targeted a 600 bp insert and when the libraries were sequenced the reads were mapped back to the assembly using BWA<sup>90</sup>. The actual library insert sizes were calculated, then plotted and these are shown in Figure 3.2



**Figure 3.2:** The consequences of E-gel size selection and amplification in paired-end library construction. The difference in BWA mapped library insert sizes for amplified, E-gel size selected (green) and unamplified, LabChipXT size selected (blue) paired-end libraries.

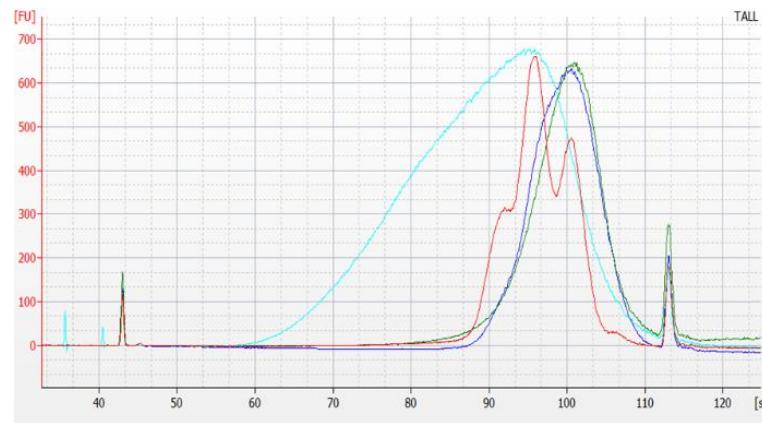
When size selecting using E-gels, some smaller DNA molecules are either retained in the collection well, or trapped amongst larger fragments. PCR preferentially amplifies smaller molecules and this phenomenon can be seen with E-gel size selected and amplified library. Figure 3.2 shows the presence of molecules with inserts <200 bp in this diatom library. By contrast, the amplification-free library has very few sequenced molecules with inserts <500 bp.

However, it is worth noting that amplification cannot be escaped completely on second generation NGS instruments. Illumina instruments require bridge amplification to generate clusters containing sufficient library molecules for fluorescent detection and 454 pyrosequencing uses emulsion PCR to coat beads with enough library molecules for signal detection. Amplification biases because of high GC content in any of these steps could result in some regions of the genome being under represented in the sequence outputs.

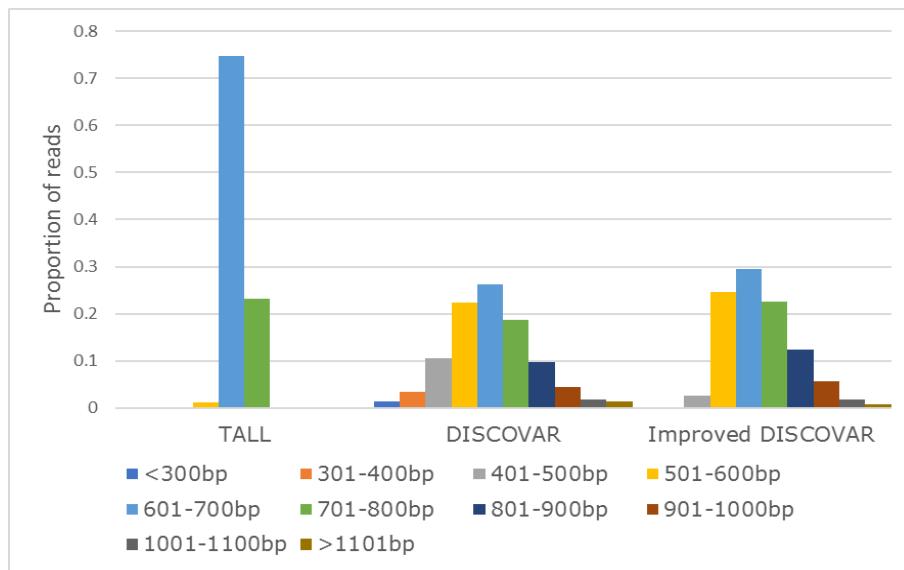
## 3.2 Evolution of paired-end library construction

### 3.2.1 Development of TALL libraries

Hypothesising that robust, amplification-free, large insert, narrow insert size distribution libraries would aid assembly, by providing proportionally more reads spanning larger repeats, I developed TALL libraries. I fragmented 3 µg of DNA and molecules 800 bp +/-10 % were isolated on the BluePippin ahead of library construction. These were sequenced with 2x 150 bp reads and library insert size determined by mapping reads back to the genome assembly using BWA. A typical TALL library Bioanalyzer electropherogram shown in Figure 3.3 and insert size distribution in Figure 3.4.



**Figure 3.3:** Paired-end library Bioanalyzer electropherograms. Bioanalyzer electropherograms for TALL (red), DISCOVAR (turquoise), Improved DISCOVAR (green) and SE-APE (blue) libraries.



**Figure 3.4:** TALL, DISCOVAR and Improved DISCOVAR library insert size distribution plots.

With a mean insert size of 690 bp, TALL libraries were significantly shorter than the 800 bp targeted. This, in part, was due to slight inaccuracies with the size selection on the BluePippin and the fact that the DNA polymerase I used for end repair has 3' to 5' exonuclease activity. As size selection was performed ahead of library construction, it would be expected that a number of molecules would have a 3' overhang which would be removed by this activity shortening some of the DNA molecules.

Although TALL libraries had a very tight insert size distribution, with >95 % of molecules in the 600-800 bp size range, they lacked molecules >800 bp limiting the spatial information the libraries could provide. These TALL libraries were used in a CS42 wheat genome project and will be discussed in Chapter 5.

TALL libraries lent themselves to investigation of the effect of wide spanning insert sizes on sequence outputs. I constructed libraries with tight insert size distributions centred on 400 bp, 600 bp and 800 bp and pooled then sequenced these on a MiSeq with 2x 250 bp reads. The number of reads, and the base pair to which at least 75 % of the reads have an expected error rate of <1 in 1,000 (Q30), are shown in Table 3.2.

Library	Number of Reads	Q30 (bp)
<b>TALL_400 bp Read 1</b>	5,377,064	219
<b>TALL_400 bp Read 2</b>	5,377,064	159
<b>TALL_600 bp Read 1</b>	7,787,062	179
<b>TALL_600 bp Read 2</b>	7,787,062	79
<b>TALL_800 bp Read 1</b>	3,922,822	149
<b>TALL_800 bp Read 2</b>	3,922,822	0

**Table 3.2:** The effect on Q30 of simultaneously sequencing libraries with different insert sizes.

Although the 800 bp insert size libraries could be clustered successfully, the reduction in quality especially for read 2 was alarming. This phenomenon is down to the Illumina software which sets quality thresholds based on the most intense fluorescence. It favours tighter clusters generated from smaller library molecules and it assigns them a higher quality score compared with larger molecules with more diffuse clusters. This indicates that in libraries that have broad insert size distributions, larger insert library molecules are less likely to pass filtering compared with smaller insert library molecules.

Illumina recommends that you spike in a PhiX control library into every sequencing lane as this helps set the quality metrics for the run. The error rates for the known sequence of the control library, along with the intensity of the signal, is assessed by the software and this helps set the Q30 threshold. As this control library has an average size of 451 bp, this data suggests that spiking this in alongside a library with a >200 bp larger insert could compromise the quality outputs for that library. For the larger insert libraries discussed later in this chapter, we did not spike in the PhiX control library so that it did not influence the Q30 scores.

### 3.2.2 Evaluation of DISCOVAR libraries

As read lengths on the Illumina HiSeq2500 increased to 2x 250 bp, scientists at the Broad Institute developed a bead based size selection, amplification-free, paired-end library construction protocol and an accompanying assembly algorithm, DISCOVAR<sup>91,92</sup>. Fragmenting 500 ng of material and targeting a 500 bp molecule, a 0.6x SPRI bead based clean-up was used to remove many of the DNA fragments <400 bp. A typical DISCOVAR library Bioanalyzer electropherogram is shown in Figure 3.3 and library insert size distribution shown in Figure 3.4.

Compared with TALL libraries, the DISCOVAR libraries had a much lower mean insert size of 570 bp. They had the advantage of larger molecules with library insert sizes ranging from 300 bp to >1.1 Kbp but <15 % of reads had inserts >800 bp, limiting the spatial information they provided. Up to 15 % of reads were <500 bp and as they would overlap on sequencing, they reduced effective coverage. This highlighted the limitations of bead based size selection to effectively remove smaller molecules during paired-end library construction.

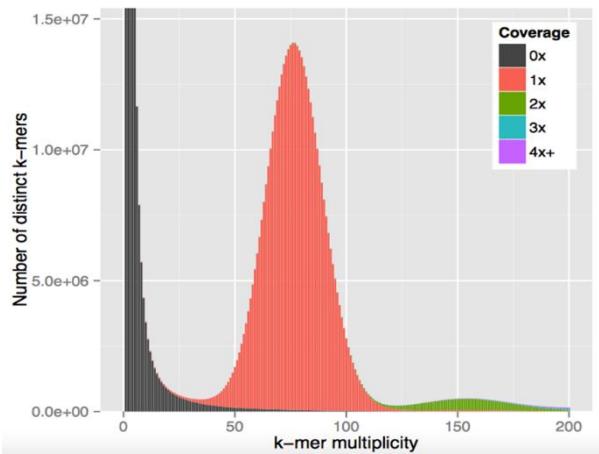
For *S. verrucosum*, I constructed both a DISCOVAR library (2x 250 bp sequence reads) and TALL library (2x 150 bp sequence reads). For further comparison a draft genome of *S. tuberosum* had been published by the Potato Genome Sequencing Consortium (PGSC)<sup>39</sup>. They generated 16 different Illumina libraries with inserts ranging between 200 and 811 bp, sequenced with 2x 100 bp reads and combined this with single-end 454 data. Metrics for these different assemblies are shown in Table 3.3.

Genome	Coverage	CN50 (Kbp)	Contigs	Total Length (Mbp)
<i>S. verrucosum</i> (TALL) <sup>10</sup>	135x	75	33,146	702
<i>S. verrucosum</i> (DISCOVAR) <sup>10</sup>	120x	77	25,216	646
<i>S. tuberosum</i> <sup>39</sup>	n/a	22.4	na	na

**Table 3.3:** Assembly metrics for *S. verrucosum* and *S. tuberosum*.

When assembling genome data, the more reads that span a repeat, the more confident an assembler can be in producing a contig that includes it. Although a library can contain some large insert library molecules, it may not generate uniform coverage across the entire genome and have sufficient reads in a given region to resolve a repeat. TALL and DISCOVAR assemblies for *S. verrucosum* had very similar CN50s, 75 versus 77 Kbp, which were >3x better than the published *S. tuberosum* assembly. However, while the DISCOVAR assembly had 30 % fewer contigs, the TALL assembly had 8.5 % more content, which is closer to the estimated genome size of 720 Mbp. This indicated that the increased average insert size TALL library (mean 650 bp versus 570 bp) allowed more repeats to be resolved and more content to be assembled for *S. verrucosum*.

This suggested the DISCOVAR assembly is collapsing content where it can't resolve some repeats. The KAT plot for the *S. verrucosum* DISCOVAR plus LMP assembly is shown in Figure 3.5. There is a second peak at 150x coverage, the green portion of the plot at twice the *k-mer* multiplicity of the main red peak, confirming that the assembled sequence in this region of the graph has twice as many reads. This indicates that *S. verrucosum* has a class/ classes of repetitive DNA sequence that only the TALL library can resolve and suggests that a larger insert, broader spanning library might have resulted in an even more contiguous assembly.



**Figure 3.5:** KAT plot of the *S. verrucosum* DISCOVAR assembly. The presence of the green peak at twice the *k-mer* multiplicity of the red peak is indicative of the DISCOVAR assembly unable to resolve repeats.

Reproduced from Paajanen *et al.*<sup>10</sup>

### 3.2.3 Improving DISCOVAR libraries

To improve the insert size distribution, enabling a greater proportion of longer molecules to be sequenced, I reworked the DISCOVAR protocol performing a less aggressive fragmentation and a more stringent bead based clean-up.

The Covaris S2 ultrasonication instrument was used to fragment the DNA. By controlling the intensity and frequency of the soundwaves the instrument produces, DNA can be fragmented between 100 and 5,000 bp. Using a duty cycle of 5 %, intensity of 3 and cycles/ burst of 200 for 40 seconds, I fragmented DNA molecules targeting an average size of 1 Kbp.

Maintaining DNA input at 500 ng, I switched to a 0.58x bead based clean-up to increase the proportion of molecules >500 bp. The yield recovered after size selection was 10 % greater than for the standard DISCOVAR size selection due to the increase in average molecule length and this indicated that sufficient material was present to construct a library suitable for sequencing.

A typical Bioanalyzer electropherogram for an improved DISCOVAR library is shown in Figure 3.3 and insert size distribution shown in Figure 3.4. Figure 3.3 confirms the increase in larger molecules over both the TALL and DISCOVAR libraries but closer inspection also reveals the presence of smaller molecules in the library providing further confirmation of the inability of bead based size selection protocols to efficiently remove smaller molecules.

I used this protocol to construct libraries for an as yet unpublished polecat *de novo* genome project. Average inserts for these improved DISCOVAR libraries was 700 bp which was larger than both TALL and standard DISCOVAR libraries. They had <6 % of reads <500 bp, reducing the number of reads that would overlap on sequencing compared with the standard DISCOVAR libraries, and >20 % were >800 bp improving the spatial information they provided.

For polecat, using the single-end reads from >50x coverage of the improved DISCOVAR library and assembling the data with the W<sup>2</sup>RAP assembler produced a CN50 of 155 Kbp. When this was scaffolded using the paired read this increased to a CN50 of 255 Kbp. Contiguity was 10-fold greater than that observed by Peng *et al.* when sequencing the ferret (*M. putorius furo*)<sup>93</sup>, a close relative of the polecat, for which they achieved a CN50 of 22 Kbp with 45x coverage. For the ferret, the use of 180 bp average insert sized paired-end libraries would have resulted in very few of the SINEs being resolved. As these are a major repetitive element in mammals, this would have contributed to the reduced contiguity. In contrast, the 700 bp average insert of the improved DISCOVAR library used in the polecat assembly would have resolved many of the SINEs, some LINEs and some smaller LTRs and this will account for much of the increase in contiguity.

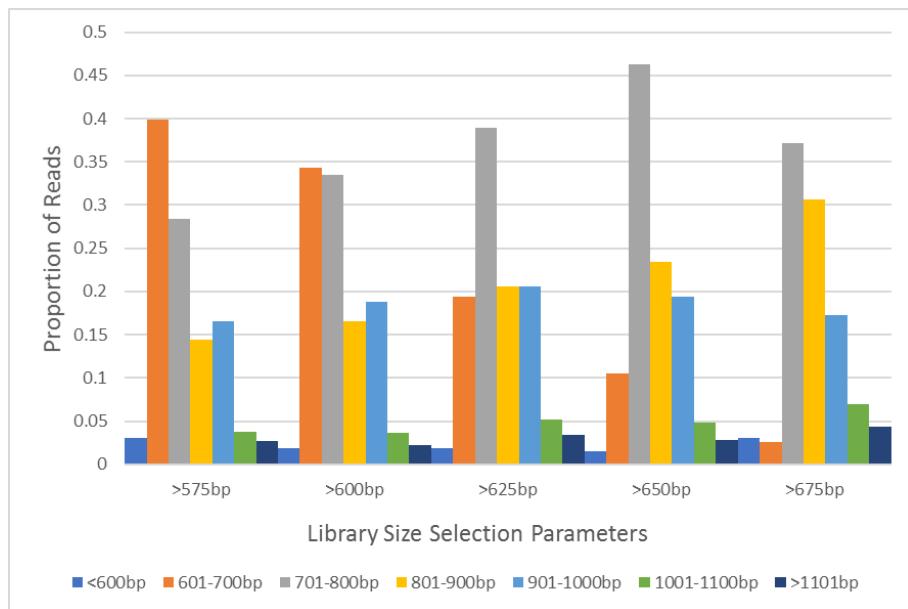
Although these improved DISCOVAR libraries generated a highly contiguous polecat assembly, the continued presence of the smaller library molecules was a concern. In my experience, these often cluster preferentially over larger molecules so would continue to limit the spatial potential of a library. This led to my development of a method to maximise the spatial information provided by a paired-end library.

### 3.2.4 Maximising spatial information in paired-end libraries

Huptas *et al.* claim to be the first to look at the effect of insert size in genome assembly in prokaryotes<sup>45</sup>. Using a double SPRI based size selection they compared different GC content and investigated a range of insert sizes. They achieved their best assemblies with average library inserts of 990 bp and 1.2 Kbp and determined sequence depths between 50 to 80x coverage proved optimal.

As the improved DISCOVAR libraries still produced many overlapping reads and fewer reads >800 bp than I had hoped for, I developed a hybrid of the TALL and improved DISCOVAR methods, SE-APE libraries. I expected that these libraries would have fewer library molecules producing overlapping sequence reads, reduced chimeric molecules, as these would be >1.1 Kbp and unlikely to be sequenced, and an increased proportion of molecules with inserts >800 bp.

I fragmented 1 µg aliquots of DNA, targeting a 1 Kbp fragmentation and then used the high pass settings on the BluePippin to exclude molecules below sizes ranging from 575 to 675 bp. I then constructed five amplification-free, paired-end libraries using this size selected material. A typical SE-APE library Bioanalyzer electropherogram with molecules <600 bp removed is shown in Figure 3.3. The BWA mapped insert size distributions for libraries with different size exclusion settings are shown in Figure 3.6.



**Figure 3.6:** The effect of different size exclusion parameters on library insert size distribution. The insert size distribution of library molecules binned into 100 bp size ranges when molecules less than the desired size are removed using the high pass setting on a BluePippin.

SE-APE libraries with molecules <600 bp removed had an average insert size of 780 bp which was significantly greater than TALL, DISCOVAR and improved DISCOVAR libraries. They spanned 600 bp to 1.2 Kbp with >40 % of reads >800 bp and <1 % of the library inserts were <500 bp, confirming they had minimal overlap between reads and maximised the spatial information they provided and coverage generated.

In broad spanning libraries, molecules with larger inserts can be compromised. While Huptas *et al.* found that insert sizes up to 1.2 Kbp are optimal for assemblies, their data shows these libraries spanned 350 bp to 1.7 Kbp. They reported that many of the reads with inserts >1.2 Kbp were lost on filtering (Phred score <20) suggesting that these molecules are too long relative to the smaller molecules present. By contrast, when not spiking the PhiX control library when sequencing SE-APE libraries with molecules <600 bp removed, it is not uncommon to achieve Q30 scores of 250 bp on read 1 and 229 bp on read 2 on sequencing.

It is possible that targeting libraries with increased size exclusion settings would be beneficial and could improve contiguity further. However, from the data used to generate Figure 3.5 and Figure 3.6, when excluding molecules <625 bp DNA was 80 %, for the <650 bp 65 % and for the <675 bp 50 % of that recovered for the <600 bp SE-APE libraries. This suggested that input amounts would need to be considerably higher to ensure sufficient library molecules for sequencing and it may be prohibitory for some samples. To investigate this further would require the fragmentation and DNA input amounts to be re-optimised. An additional consideration is that, in my experience, larger molecules are notoriously more difficult to accurately quantify and cluster, adding further complications.

It could be argued that sequence read length increases from 25 to 300 bp have made as big a contribution as insert size optimisation to improved contiguity in the genomes discussed in this thesis. Longer reads allow longer *k*-mers to be used which in turn can lead to more repeats being resolved. This would result in fewer edges in the DBG and longer contigs but this can be read depth dependent. However, Huptas *et al.* noted there was no benefit in sequencing beyond 189 bp, as Illumina instruments introduced too many errors after this point, and it has been empirically calculated that increasing the insert size increases the physical coverage and reduces the length of read required. With a 5 Kbp insert, an 18 bp paired read can be used to unambiguously map the *E. coli* genome. Minimal improvement is also observed in assembly contiguity when increasing paired-end read length from 35 bp for *E. coli* and 60 bp for *S. cerevisiae* with a 300 bp insert library<sup>94</sup>.

### 3.3 Summary

Amplification-free, paired-end libraries will continue to be constructed and sequenced as the Illumina instruments are the most accessible of the current NGS platforms. As the technology has improved and read lengths increased, I have evolved and improved paired-end library construction to maximise the quality and quantity of the data generated.

With the minimum requirement of 1 µg of DNA >10 Kbp, I have shown that the SE-APE protocol provides an innovative library construction solution for *de novo*

genome projects. It overcomes problems associated with targeting specific insert libraries seen in TALL libraries, reduces the number of reads that overlap increasing the effective and physical coverage and reduces the cost of sequencing. It maximises the spatial potential of Illumina sequencers and reduces the number of chimeric molecules that will be sequenced.

Reads generated using this protocol provide a great resource for validating assemblies using KAT plots and the libraries have underpinned the development of the W<sup>2</sup>RAP genome assembler. The application of SE-APE libraries in *de novo* genome projects will be discussed in more detail in Chapter 5.

## 4 Enhancing LMP library characteristics

In recent years there have been numerous publications of genome assembly projects and of new software algorithms aimed at improving contiguity using NGS sequence data, yet there have been surprisingly few looking to optimise library construction, especially for LMPs. As sequencing accessibility for many laboratories is limited to short read sequencers and many genomes assembly projects require greater contiguity than can be achieved by paired-end reads alone, this created an opportunity to optimise LMP library construction and develop a more robust method.

My LMP library construction publication<sup>5</sup> describes a novel, robust approach to constructing LMP libraries and is submitted as part of this thesis. In this chapter I discuss the benefits of my innovative approach in constructing large insert size and highly complex LMP libraries with reduced input requirements and tight insert size distributions. I introduce further improvements to the protocol and highlight the advantages and limitations compared with previously published LMP library construction methods.

The LMP insert size distribution and duplication statistics discussed in this chapter were calculated at EI by Gonza Garcia-Accinelli and Jon Wright.

### 4.1 Established LMP library construction strategies

#### 4.1.1 Different approaches to constructing LMP libraries

Ditags, or LMPs as we now refer to them, were first described in 2006<sup>95,96</sup> and NGS equipment manufacturers soon released their own library construction protocols. Methods were based on cre-lox (454) and intramolecular ligation (ABI SOLiD and Illumina) and more recently transposase methods (Illumina) have been released, which remove the need for physical fragmentation, simplifying the LMP library construction process.

Targeting 3 Kbp inserts, Park *et al.* compared each LMP approach, including their own homebrew method, based on a hybrid of the ABI and Illumina methods, and targeted 1.5 million reads per library<sup>97</sup>. The Illumina intramolecular ligation protocol produced the highest proportion of unique, true LMP reads (85 %) and 454 the least (45 %) with their homebrew protocol averaging 80 % and the Nextera protocol 75 %.

They went on to construct LMP libraries for seven mice strains targeting 3 and 6kb inserts using their own homebrew method, inputting 10 and 20 µg of DNA respectively. Size selection was performed using a BluePippin and recoveries averaged 11.4 % and 13.6 % of starting material and averages of 80.2 % and 81.7 % of reads were determined unique, true LMPs. Mean insert sizes were 3.7 and 6.6 kbp and these spanned 2 to 5 Kbp and 4 to 8 Kbp respectively. Library characteristics for these are shown in Table 4.1.

Although DNA inputs were high, considering the insert sizes they were targeting, the percentage of unique true LMP reads was impressive. The nick translation method they adopted is technically the best LMP library construction method. By ligating biotinylated adapters which are then used to walk out from via a nick translation step after circularisation, it ensures that the junction molecule sits in the middle of the final library molecule, so every read has the potential to be a true LMP. By comparison, as Nextera utilises random fragmentation after circularisation, the junction molecule can be at any point in the final library molecule. When targeting final libraries with average 400 bp inserts and needing a minimum of 25 bp either side of the junction molecule for the read to be informative, theoretically only 87.5 % of all reads can be true LMPs.

Genome project	Insert Size (Kbp)	PCR Cycles	Number of Reads (million)	% Unique, True LMPs
<b>Mouse Strains<sup>97</sup></b>	3	10	12.3-22.9	71.1-88.1
	6	10	12.7-20.7	69.7-91.4
<b>Rat<sup>98</sup></b>	3	14	17.7	85.8
	5	18/ 13	11.9/ 16.7	40.2/ 83.8
	8	14/ 13	20.8/ 11.8	37.5/ 89.8
	15	21/ 21	31.2/ 11.6	3.2/ 10.3
	20	14	13.3	44.3
	25	17	56.9	1.9
<b><i>P. Picta</i><sup>99</sup></b>	1-6	10/ 10	15/ 15	59.9/ 65.4
	11-18	10	15	68.1
<b>Wheat-CS42<sup>7</sup></b>	9	10	432	48.5
	11.3	12	404	44.9
<b>WLA</b>	9.5	10	165	69.2
	11.5	10	365	62.5
	14.6	10	354	57.4

**Table 4.1:** LMP library characteristics for different genome projects.

#### 4.1.2 Investigating the benefit of multiple insert size LMPs

In optimising the scaffolding of the rat genome, van Heesch *et al.* used the ABI SOLiD LMP construction protocol with 100 µg of input material and cut bands out of agarose gels to construct LMP libraries with inserts ranging from 3 to 25 Kbp<sup>98</sup>. Characteristics for the LMP libraries constructed in this study are shown in Table 4.1. Their libraries suffered from inaccurate insert sizes and wide size distributions. The target 10 to 14 Kbp insert library was shown to be 8 Kbp and spanned 4 to 12 Kbp when mapped back to the assembly. They also needed to perform up to 21 PCR cycles to get

sufficient library molecules for some insert sizes which resulted in some very low complexity libraries.

Sequencing and integrating multiple different insert size LMP libraries benefited the rat genome assembly. Scaffolding with a 15 Kbp insert library achieved a SN50 of 163 Kbp. This was improved to 522 Kbp by using a combination of 5 and 25 Kbp insert libraries and in incorporating data from all the LMPs, this rose to 1.28 Mbp.

An improvement in contiguity by adding multiple different insert LMP library data was also observed in assembling the *S. tuberosum* genome<sup>43</sup>. Adding successive insert size LMP library data up to 10 Kbp effectively doubled SN50 at each step and in adding a 20 Kbp insert LMP the SN50 more than trebled to 1.30 Mbp.

#### **4.1.3 Reducing costs and improving Nextera LMP outputs**

In sequencing the *P. picta* genome, Tatsumi *et al.* reported an optimised method to reduce costs and improve outputs. They prepared their own reaction buffer and by switching the strand displacement and size selection steps, increased the capacity of the standard Nextera LMP library kit threefold to 36 reactions<sup>99</sup>. They also used four 50 cycle TruSeq Rapid SBS v1 kits allowing them to sequence 2x 171 bp reads to further to reduce costs.

They constructed LMPs with insert sizes ranging from 1 to 6 Kbp and 11 to 18 Kbp and library characteristics for these are shown in Table 4.1. Assembling the genome with the higher complexity 1 to 6 Kbp insert library, compared with the lower complexity library, increased SN50 by 20 %. Adding the 11 to 18 Kbp insert library increased it 36-fold to 1.81 Mbp.

They highlighted the benefit of increasing the read length from 100 bp to 171 bp, and achieved an improvement in the percentage of true LMPs from 59 % to 65 % when they reduced the insert size of their final library to between 400 and 700 bp. As the Illumina adapters account for 130 bp of this, their library insert sizes ranged between 270 and 570 bp. Therefore, not all reads would overlap when using 2x 171 bp reads

and they may not have identified an adapter junction in every library molecule they sequenced.

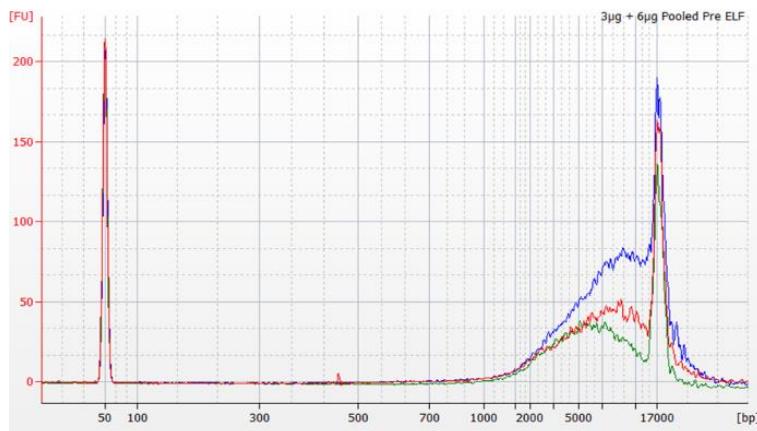
## **4.2 Development of a unique LMP library construction protocol**

My LMP paper describes a new way to think about library construction with the ability to construct up to twelve different insert size libraries at the same time using the SageELF. Many of the genome assembly projects discussed in this thesis used multiple LMP libraries, with different insert sizes, and my method streamlines the process, saving both time and money and requires proportionally less input material.

### **4.2.1 The benefit of controlling LMP insert size and distribution**

LMPs provide the spatial information to be able to scaffold across repeat sequences smaller than the library insert size. Accurately controlling the span and insert size of LMP libraries has multiple benefits. It simplifies scaffolding reducing the number of non-determined bases helping improve contiguity and minimises redundancy when producing different insert size libraries.

Using wheat CS42 DNA, I optimised the ratio of DNA to transposase enzyme to increase the insert size of the fragmented DNA so that more molecules were in the desired 8 to 12 Kbp target insert size range. I performed two fragmentation reactions, one with 6 µg and the second with 3 µg and the effect on fragmentation is shown in Figure 4.1. The amount of DNA I retrieved after SageELF based size selection for each of the twelve fractions is shown in Table 4.2. A total of 22.2 % of starting material was recovered across all the fractions. For wheat, 100 ng of material represents >5,000 copies of the genome. This yield has the potential to generate highly complex LMP libraries and was recovered for all but the smallest and largest fraction highlighting the accuracy of the fragmentation optimisation. Because of this improved recovery post-size selection, I reduced the number of PCR cycles from the recommended 10-15 down to 8-12.



**Figure 4.1:** The effect of increasing DNA concentration on transposase mediated fragmentation. Fragmentation patterns for tagmented samples with 3  $\mu$ g (green) and 6  $\mu$ g (red) DNA added and then when pooled ahead of size selection (blue).

Reproduced from Heavens *et al.*<sup>5</sup>

Average insert size and insert size distributions for all twelve fraction LMP libraries calculated using BWA, are shown in Table 4.2. My LMP libraries centred on inserts of 9 Kbp and 11.3 Kbp were sequenced to a greater depth as part of our CS42 genome project and will be discussed further in Chapter 5.

Since publication of my protocol, I have constructed LMPs for a further eleven wheat lines and in targeting insert sizes of 9.5 and 12 Kbp, average insert sizes have been 9.4 and 11.9 Kbp. Four of these wheat lines had an additional LMP library sequenced targeting 15 Kbp inserts and these were shown to be an average of 14.9 Kbp.

ELF Fraction	Average Insert Size (Kbp)	Recovery Post Size Selection (ng)	Size Span (Kbp)	
			smallest	largest
<b>1</b>	(insufficient data)	53.4	-	-
<b>2</b>	14.8	169.2	12.750	17.749
<b>3</b>	11.3	245.4	9.966	13.327
<b>4</b>	9.0	261	7.988	10.021
<b>5</b>	7.3	181	6.459	8.356
<b>6</b>	5.9	248.4	5.148	6.716
<b>7</b>	4.8	153	4.107	5.551
<b>8</b>	3.8	204	3.261	4.445
<b>9</b>	3.2	184.8	2.601	3.618
<b>10</b>	2.4	120	1.972	2.854
<b>11</b>	1.9	109.2	1.520	2.290
<b>12</b>	1.4	75	1.110	1.780

**Table 4.2:** DNA recovered post size selection and library characteristics for ELF based CS42 LMP libraries.

Reproduced from Heavens *et al.*<sup>5</sup>

Of the methods discussed in this chapter, automated size selection outperformed manual size selection with the SageELF being the most reliable at targeting insert sizes and it also had the benefit of producing narrower spanning libraries. The average insert size span across the twelve fractions was -15 to +17 % of the targeted size and it was tightest in the 9 Kbp insert library spanning +/-11 %. By contrast, the 8 Kbp rat library spanned +/-50 %, the 6 Kbp insert mouse LMPs spanned +/-33 % and the 14.5 Kbp *P. picta* LMP library spanned +/-25 % of the average insert. Tightly distributed libraries should make assembly more straightforward, and improve

contiguity, highlighting a benefit of my approach over the other protocols discussed in this chapter.

## 4.2.2 Improving LMP library complexity

Maintaining library complexity is an important attribute of a robust LMP library construction protocol. Highly complex LMP libraries have more unique molecules providing more information and require less sequencing reducing the need to construct multiple libraries with the same insert size.

To determine complexity of the libraries constructed using my approach, sequence reads were first processed through FLASH<sup>100</sup> to determine the numbers of reads that overlap and provide contiguous sequence of the final library insert. My LMP libraries are typically in the 85 to 90 % range, maximising the chances of finding the junction adapter molecule. Reads were then deduplicated to remove any identical reads and then processed using NextClip<sup>101</sup> which categorises them to determine the proportion of reads that are informative and true LMPs. Complexity and other library characteristics for the CS42 LMP libraries constructed using my approach and sequenced as part of the CS42 *de novo* genome assembly project, presented in Chapter 5, are shown in Table 4.1

Comparing the proportion of unique, true LMPs for the libraries discussed in this chapter, they range from 1.9 to 91 % although these values can be misleading as the number of reads for each of the libraries range from 11.6 to 432 million. The more a LMP library is sequenced, the more chance there is of sequencing a duplicate read. Therefore, to provide a more direct comparison of complexity with other studies, the percentage of unique, true LMPs in subsampled reads from sequencing runs for the CS42 9 Kbp insert LMP library were calculated. Results for this are shown in Table 4.3.

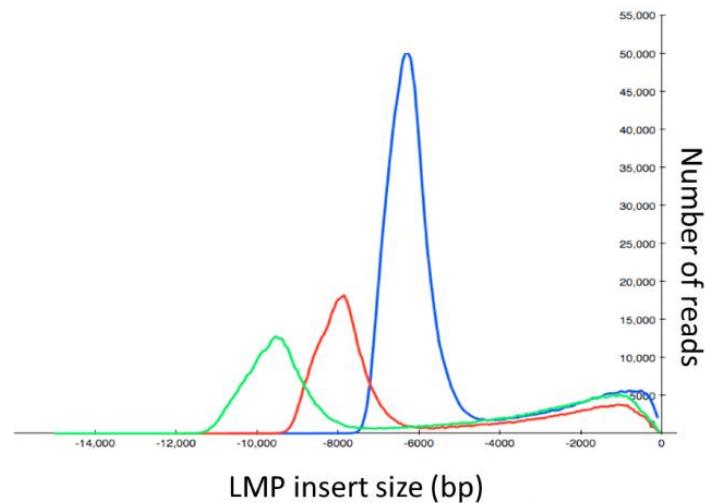
	% Unique True LMPs	
Number of reads (million)	CS42 9 Kbp insert library	WLA 9.5 Kbp insert library
<b>10</b>	65.2	74.5
<b>15</b>	65.8	73.4
<b>20</b>	65.9	72.6
<b>25</b>	66.9	72.9
<b>50</b>	63.1	70.5
<b>100</b>	60.5	67.2

**Table 4.3:** The proportion of unique, true LMPs in subsampled CS42 and WLA LMP reads.

Subsampling the CS42 9 Kbp LMP libraries clearly shows how complex libraries can be made to look if they are only sequenced to a few million reads. From the 48.5 % being true LMPs in the 432 million reads reported in Table 4.1, the value is 65 % when subsampling down to 15 million reads which is close to the values seen by Tatsumi *et al.* for their proportionally broader and smaller 6 Kbp insert library when sequenced with the same number of reads.

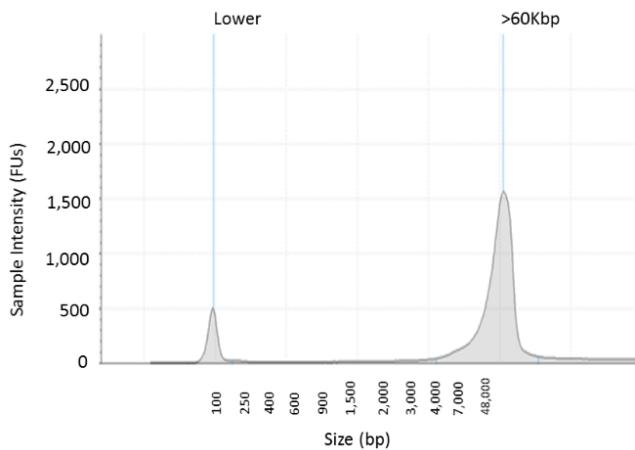
#### 4.2.3 Further improvements to my LMP protocol

Changes to my protocol were made after the presence of a small proportion of reads with shorter inserts than those targeted were observed in some libraries. BWA mapped LMP insert sizes exhibiting this phenomenon are shown in Figure 4.2. These were also observed by Park *et al.* in their LMP libraries and suggested that some smaller DNA molecules were making it through the size selection. Contaminating shorter insert reads complicates genome assembly. These make the assembly algorithms think they should be further apart than they are. To overcome this, some bioinformaticians have chosen to build algorithms to take this into account<sup>102</sup>.



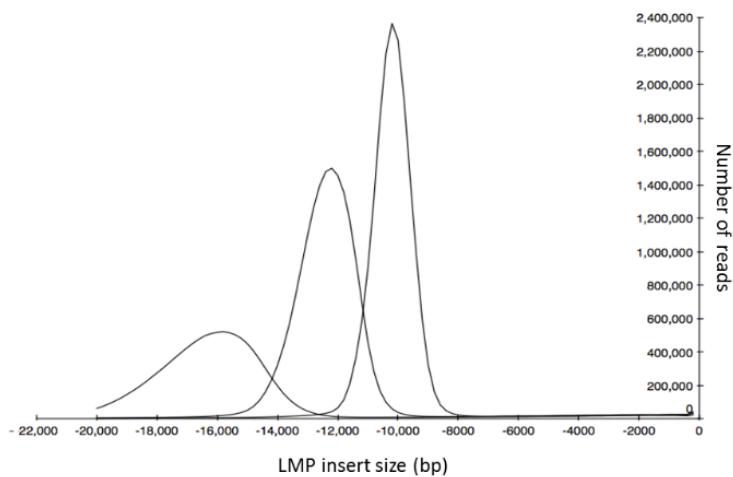
**Figure 4.2:** The presence of smaller than target insert size molecules in LMP libraries. LMP libraries with target insert sizes of 10 Kbp (green), 8 Kbp (red) and 6 Kbp (blue) showing some library molecules with insert sizes between 0 bp and 4 Kbp.

Working with a new wheat sample-Wheat Line A (WLA), I extracted DNA using a CTAB protocol<sup>103</sup>, as outlined for *S. verrucosum*<sup>10</sup>, and immediately ahead of LMP library construction. The Agilent TapeStation (Agilent, Stockport, UK) genomic tape electropherogram for this DNA is shown in Figure 4.3 and revealed the DNA to be >60 Kbp and suitable for LMP library construction.



**Figure 4.3:** Agilent TapeStation genomic tape electropherogram of freshly extracted WLA DNA.

I constructed LMP libraries targeting 15, 12 and 10 Kbp inserts and BWA mapped reads for these LMP libraries are shown in Figure 4.4. No smaller insert library molecules were observed. A possible explanation for the improved size selection could be due to lack of smaller fragments in the freshly extracted DNA which may get caught behind or trapped within the larger DNA molecules during electrophoresis. This is a similar phenomenon to that seen in the diatom work presented in Chapter 3, albeit with smaller fragments.



**Figure 4.4:** The absence of smaller than target insert size molecules in LMP libraries. Three LMP libraries constructed showing the absence of smaller than target molecules present when the final libraries are mapped back to the genome assembly using BWA.

I recovered 12 ng less, 8 ng more and 1 ng less for ELF fractions 2, 3 and 4 when using freshly extracted DNA than for the corresponding CS42 LMPs suggesting that freshly extracted DNA did not impact size selection yields. I was able to reduce the PCR cycle number down from 12 to 10 cycles for fractions 2 and 3 and the yields obtained were within 5 % of that achieved for the CS42 libraries. With two fewer PCR cycles, if the amount of starting material was the same, you would expect four times less product. Achieving these final library yields suggests that there is a higher proportion of size selected material in the freshly extracted DNA fractions that goes on to circularise and provide template for the remaining LMP library construction steps. Freshly extracted DNA will have longer molecules present and this increases the

chances of adapters being inserted the correct distance apart by the transposase for the insert sizes I was targeting and this could account for the observed increase in final library yield.

The characteristics for these WLA LMP libraries is shown in Table 4.1 and for comparison purposes the percentage of unique, true LMPs from subsampled reads for the 9.5 Kbp library shown in Table 4.3. When used for scaffolding, these libraries resulted in our best hexaploid wheat assembly to date achieving a SN50 >120 Kbp with only 80x genome coverage.

Subsampling the WLA 9.5 Kbp insert LMP reads shows the library is more complex than the comparable CS42 library at each subsampling point. The 72 % unique, true LMP reads of the 9.5 Kbp insert WLA library when subsampling down to 20 million reads is less than the average 81.7 % true LMPs seen in the 6 Kbp insert mouse LMP libraries. However, you would only expect a maximum of 87.5 % of library molecules to be deemed true LMPs in a Nextera library and this represents 82 % of them. Additional advantages of the WLA 9.5 Kbp insert LMP library is the benefit of having a 50 % increase in insert size, which provides more physical coverage, and a much tighter insert size distribution which aids scaffolding and should improve contiguity. It also only requires 9 µg of input material compared with 20 µg for the mouse 6 Kbp insert LMP library.

Since observing the improvement in complexity and absence of smaller insert libraries, this has been replicated in a further four wheat LMP libraries confirming that the modifications to my protocol are both robust and reproducible.

### 4.3 Summary

Ultimately it is a combination of input requirements, library characteristics and cost that will determine which LMP library protocols are widely adopted. An ideal LMP library would be highly complex, have the desired insert size with a tight insert size distribution, not require prohibitively high DNA input or be prohibitively expensive.

For large complex genomes, my published LMP library construction protocol combined with the subsequent improvements discussed here provides a cost effective and time efficient means to aid *de novo* genome assembly. LMP libraries prepared this way have the ideal combination of larger inserts, which provide more physical coverage, are more complex and enable more control over insert size and distribution over established protocols coupled with reduced input requirements. They also have the added benefit of constructing twelve libraries simultaneously, all of which can be sequenced if desired.

Constructing the twelve libraries using my approach can be achieved for twice the cost of a single LMP and I have shown it to be more robust than other approaches discussed in this chapter. To date, it has been used in more than twelve wheat genome assembly projects, some of which will be discussed in Chapter 5.

## 5 Improving wheat *de novo* genome assemblies

For higher eukaryotes, the prevalence and nature of repetitive DNA sequences, polyploidy and large genome sizes makes *de novo* genome assembly more challenging. In this chapter I discuss the application of my novel library construction protocols, presented earlier in this thesis, in wheat *de novo* genome assembly projects. I constructed multiple amplification-free, paired-end and LMP libraries for CS42 and this provided an opportunity to determine if my optimised methods helped improve contiguity in a repeat rich, polyploid plant species. This is supported by publication of the CS42 genome<sup>7</sup> which is submitted as part of this thesis.

I went on to construct paired-end and LMP libraries for a further five wheat lines, one tetraploid and four hexaploids, and assemblies for these have been made available through EI's Grassroots Genomics Portal<sup>104,105</sup>. These helped test the robustness of my protocols. In addition, many alternative strategies have been published in attempting to decode wheat and by comparing assembly outputs, I highlight the advantages and disadvantages of my approaches over these.

The wheat genome assemblies and KAT plots discussed in this chapter were generated at EI by Bernardo Clavijo, Gonza Garcia-Accinelli and Jon Wright.

### 5.1 Wheat

Bread wheat is an allohexaploid with an estimated genome size of 17 Gbp. It has 21 chromosomes and, because of two independent hybridisation events, it has two copies of three genomes- known as the A, B and D genomes. Over 90 % of the genome is thought to be dispersed repeats containing at least 6.5 million LTRs and 1 million DNA transposons. Strategies to sequence wheat have included decomplexing the genome by sequencing BACs, flow sorting chromosomes and sequencing the diploid ancestral progenitors. Prior to the publication of our CS42 assembly, there were at least two attempts at whole genome shotgun sequencing and recently there has been more contiguous hexaploid wheat and *A. tauschii* assemblies published.

Assembly metrics for the wheat genomes discussed in this chapter are shown in Table 5.1.

Publication	Strategy	Genome	CN50 (Kbp)	SN50 (Kbp)
Visendi <sup>106</sup>	BACs	CS42	80	106
Chromosome 3B <sup>107</sup>	BACs + FSC	CS42	-	<b>892.4</b>
Flow Sorted <sup>108</sup>	FSC	CS42	<b>1.7-8.9</b>	-
Belova <sup>109</sup>	FSC	7DS/ 7DL	2.4/0 .5	14.4/ 11.1
Helgeura <sup>110</sup>	FSC	4DS/ 4DL	<b>1.1/ 0.8</b>	<b>5.5/ 3.9</b>
Brenchley <sup>111</sup>	WGS	CS42	0.884	-
Chapman <sup>112</sup>	WGS	Synthetic Line W7984	6.7	25
Ling <sup>113</sup>	Progenitor WGS	<i>T. uratu</i>	3.42	63.6
Jia <sup>114</sup>	Progenitor WGS	<i>A. tauschii</i>	4.51	58.0
Zhao <sup>115</sup>	Progenitor WGS	<i>A. tauschii</i>	50.3	6,830
Zimin <sup>116</sup>	WGS	CS42	232.6	-
Clavijo <sup>7</sup>	WGS	CS42	16.5	83.9
Grassroots Genomics <sup>104,105</sup>	WGS	Cadenza	16.0	103.8
	WGS	Paragon	16.5	84.4
	WGS	Kronos	20.0	155.8
	WGS	Robigus	16.8	86.4
	WGS	Claire	17.0	72.1

**Table 5.1:** Assembly metrics for different wheat based genome assemblies.

## 5.2 Decomplexing the wheat genome

### 5.2.1 Sequencing wheat BACs

Visendi *et al.* published a strategy to sequence pools of wheat BACs targeting 300 bp insert paired-end libraries and 6-10 Kbp insert LMPs<sup>106</sup>. They determined the optimal paired-end coverage to be between 450 and 900x and they equimolar pooled four non-overlapping BACs prior to library construction. They sequenced 96 pools at a time and used BAC End Sequencing (BES) to attribute contigs to BACs and achieved a SN50 >17 % larger than we achieved for the CS42 3DL MTP using the BAC sequencing pipeline discussed in Chapter 2.

Scientists in the International Wheat Genome Sequencing Consortium (IWGSC) took this a stage further and sequenced a MTP of BACs of wheat chromosome 3B<sup>107</sup>. Creating 922 pools from 8,453 clones from the wheat MTP they sequenced 8 Kbp insert LMP libraries to an average of 36x coverage on 454 pyrosequencing instruments. Augmenting the data with BES, they filled gaps and error corrected using Illumina reads from chromosome 3B flow sorted material and integrated the size information from BAC fingerprint data.

These approaches to sequence BACs are not cheap. If a wheat MTP BAC library was available, it would cost >£1 million to extract DNA, construct paired-end and LMP libraries and generate sequence data for all these wheat BACs using the pipeline presented in Chapter 2. It would cost significantly more for the extra 2.5 to 4.5x sequence coverage required by Visendi *et al.* The effort by the IWGSC in achieving what is considered a gold standard assembly for wheat chromosome 3B was admirable, especially for a repeat rich, polyploid plant. However, the library construction consumable cost for 922 LMP libraries would be >£100,000 and take over 100-person days to complete and that does not include sequencing or DNA extraction. There would also be additional costs for the BES and individual chromosome isolation and library construction and sequencing of the flow sorted material. Based on these figures, BAC approaches to sequence the wheat genome are simply not viable.

## 5.2.2 Sequencing flow sorted wheat chromosomes

Attention turned to flow sorting and isolating individual chromosome arms and sequencing these using NGS technology. The IWGSC isolated the long and short arms of all 21 wheat chromosomes<sup>108</sup> and set about sequencing them. CN50s above 10 Kbp proved elusive so in an attempt to improve contiguity, construction of LMPs was attempted by Belova *et al.* for chromosome 7BS and 7BL<sup>109</sup> and Helguera *et al.* for 4DS and 4DL<sup>110</sup>. The need to MDA treat sorted material to provide sufficient DNA for processing resulted in maximum LMP insert sizes <5 Kbp.

This strategy offered the potential to resolve homeologous genes but in failing to achieve CN50s >9 Kbp and SN50s >15 Kbp, it did not help improve contiguity significantly. BUSCO v2 analysis for this assembly is shown in Table 5.2.

<b>Gene Status</b>	<b>BUSCO v2</b>			<b>BUSCO v3.0.2</b>	
	<b>Chromosome Survey</b>	<b>EI</b>	<b>NRgene</b>	<b>EI</b>	<b>Zimin <i>et al.</i></b>
<b>Complete</b>	828	914	921	1,411	1,415
<b>Duplicated</b>	628	873	899	1,285	1,254
<b>Fragmented</b>	56	22	15	8	4
<b>Missing</b>	72	20	20	21	21

**Table 5.2:** BUSCO analysis for five different CS42 wheat genome projects.

Although the chromosome survey assembly lacked contiguity, with a CN50 <9 Kbp and 128 single copy ortholog genes either fragmented or missing, the data generated proved useful. It allowed much more contiguous wheat assemblies, including our own, to have their scaffolds chromosomal location confirmed.

### 5.2.3 Sequencing wheat progenitors

Prior to the release of our wheat assembly, two of the three progenitors of wheat had been sequenced using Illumina only approaches- *T. urartu*<sup>113</sup>, the A genome progenitor, and *A. tauschii*<sup>114</sup>, the D genome progenitor. As both are diploids, their genomes are less complex and theoretically easier to assemble. The strategies adopted included targeting multiple libraries with inserts ranging from 200 to 700 bp by cutting bands out of agarose gels and libraries were sequenced on HiSeq2000s with 2x 114 bp reads. LMPs with 2, 5, 10 and 20 Kbp inserts were added and 454 pyrosequencing reads used for error correcting to further improve the assembly. Although both assemblies achieved SN50s >50 Kbp, they constructed >20 paired-end libraries and >15 LMP libraries for each genome indicating they had problems with library complexity.

In 2017, a more contiguous *A. tauschii* assembly was published<sup>115</sup>. Zhao *et al.* targeted amplification-free, paired-end libraries with a 400 bp insert to generate 76x coverage using 2x 250 bp Illumina reads and then stitched these together to form continuous sequence reads before assembly. They constructed five different LMP libraries with insert sizes ranging from 2 to 40 Kbp, generating a combined 110x coverage and then added 11x coverage of 20 Kbp insert PacBio libraries.

## 5.3 WGS wheat genome project strategies

Brenchley *et al.* reported a WGS assembly for a hexaploid wheat using a combination of single-end 454FLX and FLX+ reads<sup>111</sup>. Chapman *et al.* improved contiguity in wheat when sequencing a synthetic wheat line, rather than CS42, and targeted paired-end libraries with inserts of 250, 500 and 800 bp sequenced on the Illumina HiSeq2500 with 2x 150 bp (250 and 500 bp inserts) and 2x 250 bp reads (800 bp inserts)<sup>112</sup>. They added sequence data from two LMPs with 1.5 Kbp and 4 Kbp inserts but these assemblies suffered from the inability to generate spatial information >5 Kbp resulting in SN50s <25 Kbp. With wheat LTRs >7 Kbp being the major repetitive sequence, any strategy which cannot resolve these would not achieve highly contiguous assemblies.

At the Plant and Animal Genome conference in 2015, NRgene presented their *DeNovoMagic* assembly pipeline<sup>117</sup>. Although little is known about the algorithms used to generate assemblies, or whether they used any additional data, they constructed two amplification-free paired-end libraries (460 and 800 bp inserts) and three LMP (3.5, 6 and 9 Kbp insert) libraries for CS42. They sequenced these to a total of 230x coverage and achieved a SN50 of 28.9 Mbp. BUSCO v2 analysis for this assembly is shown in Table 5.2.

Since publication of the TGAC CS42 assembly, Zimin *et al.* have produced an even more contiguous assembly<sup>116</sup>. They used a combination of 65x coverage of PCR amplified, 400 bp average insert paired-end library with 2x 150 bp reads combined with 36x coverage of 10 Kbp average insert PacBio libraries run on 1,100 SMRT cells to generate their assembly. Interestingly, they used my CS42 libraries to confirm the absence of 31 mers in their different assemblies as a QC measure to verify and validate assembly completeness. This revealed that the PacBio only assembly to be the worst and that the Illumina reads were needed to error correct to achieve the best assembly. BUSCO v3.0.2 analysis for this assembly is shown in Table 5.2.

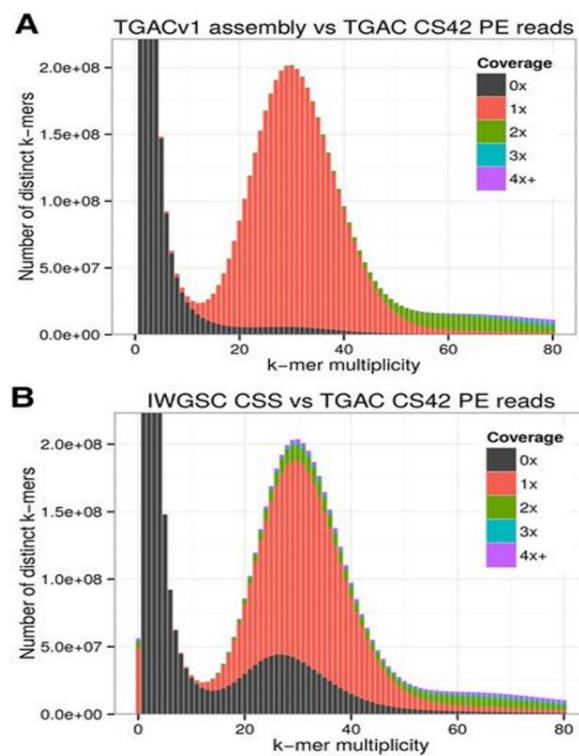
## 5.4 TGAC wheat genome assemblies

### 5.4.1 CS42 assembly

We sequenced a combination of different amplification-free paired-end and LMP libraries to generate our CS42 assembly. I constructed standard DISCOVAR paired-end libraries to generate >60x coverage and these were used for the initial contigging with the W<sup>2</sup>RAP assembler. I also constructed TALL libraries which were used to generate >30x coverage and these were used to scaffold the DISCOVAR assembly. My EI colleagues constructed four LMPs following a standard Nextera LMP library construction protocol, with size selection on a BluePippin, and I constructed two libraries with inserts of 9 and 11.3 Kbp using the published LMP protocol discussed in Chapter 4. In total the LMP libraries generated >53x coverage and these were used

in a final scaffolding step to produce the published assembly. BUSCO v2 and v3.0.2 analysis for this assembly is shown in Table 5.2.

The KAT plots for the TGAC and the IWSGC chromosome survey assemblies are shown in Figure 5.1. For the chromosome survey assembly, the KAT plot reveals that there were a significant number of reads not in the assembly, as shown by the black peak under the main red peak. It also has multiple duplications within the assembly as shown by the differently coloured peaks above the main peak and has some *k-mers* in the assembly but not in the reads, as characterised by the red portion of the plot along the y-axis. The TGAC KAT plot has fewer reads absent in the assembly and looks much cleaner. There are no misassembled sequences or duplications within the assembly but there is some evidence of true duplications that have not been resolved, as judged by the green region of the plot to the right of the main red peak.



**Figure 5.1:** KAT plots for the TGAC (A) and the IWSGC chromosome survey (B) CS42 wheat assemblies.

Reproduced from Clavijo *et al.*<sup>7</sup>

Each of three sets of chromosomes in wheat should have their own copy of a majority of the SCOs and therefore many of the orthologs will appear duplicated. BUSCO analysis reveals that the chromosome survey assembly failed to identify many of the homeologs that it was designed to resolve. Of the 828 SCOs that were deemed complete, it only identified 75 % of them as duplicated. It also failed to assemble >7.5 % SCOs and >5 % of them were fragmented. By comparison, 95 % of the 914 genes identified in our CS42 assembly had duplicates, and <3 % were fragmented with 2 % missing. The number of missing genes in our assembly was consistent with both the NRgene assembly when using BUSCO v2 and with the Zimin *et al.* assembly when using BUSCI v3.0.2, suggesting that these orthologs are not present in wheat.

The Zimin *et al.* assembly had 10 % more content and more than double the contiguity than our CS42 assembly. To generate the 1,110 SMRT cells of data would take >6 months if capturing continuous 4-hour movies. When adding in the Illumina data, the total sequencing cost for this assembly would be >£500k. BUSCO analysis reveals it only identifies four more complete SCOs but it had thirty-one less duplicated genes suggesting that it was unable to resolve as many of the homeologs as our assembly. For many, the extra cost and time will fail to justify the improved contiguity achieved by this approach.

The contiguity in terms of CN50 achieved by the NRgene assembly, by comparison to our CS42 assembly, is also very impressive. It would cost more than twice that of our approach to generate the sequence data and reports suggest that their assembly costs are considerable. It only identifies seven more complete and twenty-six more duplicated genes, but as the assembly algorithms aren't available for scrutiny, it is difficult at present to recommend this as the best approach for *de novo* genome assembly projects.

#### **5.4.2 Additional wheat line genome projects**

For the five subsequent wheat genome projects presented in this thesis, I constructed only one SE-APE library and two LMP libraries with 9.5 and 12 Kbp inserts. For the paired-end assembly, the sequence coverage requirements for the hexaploids was reduced to 55x, and for the combined LMPs 27.5x. This approach achieved similar

contiguity to that seen in CS42 and sequencing costs for the paired-end library were 60 % and LMP library 50 % of that to achieve the CS42 assembly. The further advances that I made, in improving the complexity of larger insert LMP discussed in Chapter 4, have resulted in a revised optimal strategy for sequencing wheat at EI. Our preferred wheat genome project recipe currently includes sequencing three LMP libraries with 9.5, 12 and 15 Kbp inserts each to 8.5x coverage. Using this approach, paired-end and LMP library construction and sequencing can be completed for wheat in under two weeks on a single HiSeq2500 for <£70k and can achieve SN50s >100 Kbp.

The assembled content for the TGAC CS42 genome and the four hexaploid wheat lines sequenced using my SE-APE plus two LMP library strategy are shown in Table 5.3. On average, my improved library construction protocols helped increase the assembled content for a wheat line by almost 1 Gbp (>7 % of the genome) over that achieved for CS42 confirming the benefit of the extra spatial information my methods provided. With both Cadenza and Paragon having >15 Gbp of assembled content, this is nearly 90 % of the estimated genome size and close to the 15.3 Gbp Zimin *et al.* achieved with a five times more expensive strategy which takes over twelve times longer to generate the data.

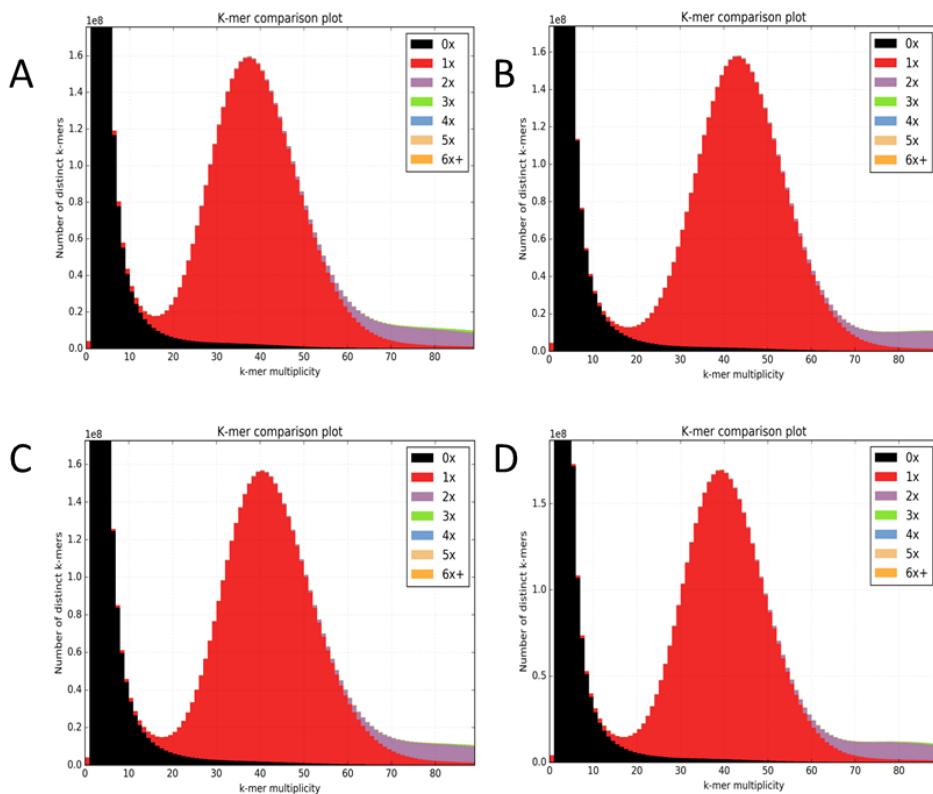
<b>Wheat line</b>	<b>Number of contigs &gt;1 Kbp (million)</b>	<b>Number of scaffolds &gt;1 Kbp (million)</b>	<b>Assembled content (Gbp)</b>
<b>CS42</b>	17.59	15.51	13.94
<b>Cadenza</b>	19.66	17.93	15.01
<b>Paragon</b>	20.21	18.31	15.11
<b>Robigus</b>	20.95	19.11	14.88
<b>Claire</b>	19.53	17.87	14.60

**Table 5.3:** Assembled content and contig/ scaffold number for five hexaploid wheat lines sequenced at EI.

The increase in assembled content in the four new wheat assemblies corresponds with an increase in number of contigs with no significant increase in CN50 (<3 % higher for the four new wheat line assemblies). The four additional wheat lines have an average of 14 % more contigs and 18 % more scaffolds than for CS42. We also

saw this effect in the *S. verrucosum* assembly discussed in Chapter 3. Both projects used different genome assemblers suggesting the improved spatial information provided by the paired-end libraries is making a significant contribution to the increase in assembled content.

The KAT plots for the four new hexaploid wheat line assemblies are shown in Figure 5.2. They confirm that the genomes are more complete than for CS42, with fewer reads missing from the assemblies. There still are some duplicated sequences to the right of the main red peak for each of the wheat lines and these will be repeats that cannot be resolved due to the limitations of the insert sizes of the SE-APE and LMP libraries I constructed. It will be some of these repeats that Zimin *et al.* were able to resolve using the longer, more continuous PacBio reads that led to them generating 279,430 contigs and a CN50 over twice that we achieved using an Illumina only approach.



**Figure 5.2:** KAT plots for the four new hexaploid wheat lines assembled at EI. KAT plots for Cadenza (A), Paragon (B), Robigus (C) and Claire (D).

With similar levels of paired-end coverage, the tetraploid wheat Kronos CN50 was up to 25 % bigger than the hexaploid wheat lines and in producing nearly twice the LMP coverage, the SN50 was up to 115 % greater. Total assembled content was 10.73 Gbp which is nearly 95 % of the estimated 11.33 Gbp genome size. Analysing the assembly, it was possible to investigate the benefit of additional LMP library data in improving genome contiguity. Both the 9.2 and 11.3 Kbp insert Kronos LMP libraries had >480,000 sequence reads generated and a total of 57.3 % and 52.9 % of the reads respectively were unique, true LMPs showing they were both highly complex libraries. When subsampling the data down to the comparable 27.5x coverage across both LMP libraries generated for the hexaploid lines, the SN50 was only 88.50 Kbp. This clearly highlights the benefit of additional LMP reads in improving genome contiguity if the library is complex enough. This suggested that we could obtain even more contiguous wheat assemblies if we sequenced LMPs to greater depth. However, careful consideration needs to be given to the cost of this relative to the number of unique reads that will be generated.

## 5.4 Summary

Illumina sequencers continue to generate the bulk of the sequence data in many *de novo* eukaryotic genome projects. In producing highly contiguous assemblies for multiple wheat lines, my paired-end and LMP library construction protocols have been shown to provide a robust, streamlined and versatile two-step library construction solution and they remove the need to decomplex genomes. BUSCO analysis suggests that the assemblies generated using these protocols identify and resolve much of the gene space and with total library construction costs <£2k and DNA requirement of 10 µg of DNA molecules >45 Kbp, these values are within the capabilities of most scientists, making the protocols accessible to all.

As of 24th January 2018, the TGAC CS42 assembly had been accessed >9,500 times and had >4,500 BLAST searches on the EI server and in 2016 alone, the EBI hosted assembly had >55,000 BLAST searches, highlighting the importance and value of this assembly as a resource to the scientific community. If we are truly to enter the pangenomic era for complex genomes such as wheat, then low cost, rapid turnaround

protocols such as those presented in this thesis will be required and the data made available to the widest possible audience.

The protocols I developed have gone on to underpin a successful million-pound grant application to generate reference genomes for the wheat MAGIC population which are the founder wheat lines responsible for >80 % of genetics in modern British farmed wheat varieties. Once completed, these will provide a useful resource for wheat breeders. Highly contiguous assemblies have also been achieved for a variety of other species ranging from fish to butterflies and mammals to trees, pointing to the global suitability of these methods. As more genomes using these protocols get published, they will gain wider appeal in the scientific community. For some, alternative or complementary strategies such as those discussed in Paajanen *et al*<sup>10</sup>. will dictate how future genome projects will be completed. This will be discussed in more detail in Chapter 6.

## 6 Discussion

Genome projects have made significant advances since publication of the completed PhiX genome sequence in 1978. In 2009, when TGAC was established, draft prokaryotic genomes could be published in the Journal of Bacteriology with a CN50 >10 Kbp. Today, while it is straightforward to achieve single contig assemblies for prokaryotes using the long-read, single molecule sequencers discussed later in this chapter<sup>59,118-121</sup>, such have been the technological advances many scientists can be seen proclaiming on social media when they achieve genome assemblies with CN50s >1 Mbp.

One of the biggest challenges currently facing genomic scientists today is what constitutes a finished genome. This leads onto the question, do they need to be finished? First published in 2001, it took a further two years before scientists claimed that the human genome was 99.9 % complete and even today new regions are being resolved and errors in the original sequence corrected. By contrast, some scientists were happy to publish based on resolving 21 %, 48 %, 78 % and then 90 % of the 17 Gbp wheat genome.

In a 2015 meeting at the Smithsonian Institute, a group of eminent scientists proposed that all known plants and animal species should be sequenced under the banner of the Earth BioGenome Project (EBP)<sup>122</sup>. They suggested a hierarchical strategy whereby a single member of the 9,000+ eukaryotic families would be sequenced to the highest possible standard. This would be followed by generating a less contiguous genome for a single species from the 150,000+ eukaryotic genera and then a less detailed sequence of the remaining species. They argued, that as the technology currently exists to complete genomes, then this should be the target for the eukaryotic family genome projects. This would then save having to re-sequence as new, more advanced technologies emerged. Whether this approach will be practical and cost effective remains to be seen and this might not be their biggest challenge. Identifying and collecting samples for DNA extraction will take considerable time, effort and cost.

Ultimately, it will be dependent upon what biological questions are being asked that will dictate what strategy is adopted. For some this will mean just identifying the gene space, for others it will be gene order and for some it will be a handful of genes that may be responsible for the production of a secondary metabolite that may be a useful antibiotic or antifungal. There is also a growing trend not to sequence individual reference genomes but to produce genomes for multiple lines which can then be compared and probed with different biological questions.

This then leads on to what is science prepared to pay? The EBP proposed a budget of \$500 million for the eukaryotic family genome projects suggesting they think it is possible to sequence a complete genome for \$50,000. Some of the sequencing strategies discussed in this thesis cost significantly more than this and the resultant genomes are far from complete. Sequencing the barley genome BAC by BAC at £400,000+ would cost >10x more than SE-APE plus ELF LMP based strategies with very similar contiguity. For wheat, the SE-APE plus ELF LMP at £75,000 is 6.5-fold cheaper and can be completed on one HiSeq2500 inside two weeks rather than the 6 months using the Zimin *et al.* approach needing both a HiSeq2500 and PacBio instrument. With sequencing costs continuing to drop and sequence outputs increasing, genome project strategies will continually evolve and so will the cost to complete them.

## 6.1 Short read sequencing

At the heart of this thesis is the development of novel, robust protocols to resolve genomic repeat structures and improve contiguity within the limits of a short-read Illumina sequencer. These protocols can deliver highly contiguous *de novo* assemblies within a reasonable budget and time frame.

Although BAC by BAC sequencing was appropriate for Barley in 2012, the development of methods discussed in Chapters 3 and 4 have made this WGS pipeline redundant. BACs will continue to be sequenced as they can help validate assemblies. For both the *S. verrucosum* and wheat CS42 papers, fully sequenced BACs were used to validate the assemblies suggesting they may still have a role in genome assembly.

As most scientific communities have access to BAC libraries, it is possible that they could be used to help resolve difficult to assemble regions in finishing genomes.

Other vector based sequencing methods such as Fosill<sup>11,123</sup> and ShARC<sup>93</sup>, can be used to provide spatial information if the insert size of the clone is known, and they may help provide long range sequence information in genome assembly projects in the same way that the BES was used by the IWGSC with wheat chromosome 3B. For ferret, it was the addition of ShARC data that helped achieve a SN50 of 9.3 Mbp (an increase of 200 % over the CN50) highlighting its potential.

With Illumina stating that the HiSeq2500 instrument will no longer be developed, read lengths will not increase on these platforms, and the SE-APE protocol is unlikely to be improved upon as a global, single library solution to generate paired-end sequence data for *de novo* genome projects. It could be argued that if many closely related genomes were to be sequenced, then deconstructing a SE-APE library by using the SageELF to isolate different fractions and then optimising the ratio of these to match the repeat content of the genome could be beneficial, but this could take considerable time so may not be cost effective. SE-APE libraries maximise the length of molecule that can reliably be clustered on a flow cell and with its broad and controllable insert size distribution, allows repeat structures <1 Kbp to be resolved. The ability to use this amplification-free data to assess the quality of assemblies through KAT plots will ensure that this method remains popular.

Of the novel protocols presented in this thesis, ELF based LMPs offer the best potential for further optimisation although ultimately these may give way to alternative long-range and linked-read technologies discussed later in this chapter. Cassette run times have been increased from 4 hours to 8 hours, allowing fragments up to 50 Kbp to be recovered. This should make it relatively straightforward to rework the protocol targeting larger insert LMP libraries. These would provide more physical coverage than those constructed to date and would not need to be highly complex. Only 2 % of the 56.9 million 25 Kbp insert LMP reads used in the rat genome assembly were unique, true LMPs, yet they were integral to them achieving their best SN50s.

Another advantage of well characterised LMP insert sizes is that they can be useful for validating an assembly. In producing an *A. alpina* genome, the insert sizes of LMP

libraries were used as a QC measure to confirm the accuracy of a PacBio plus Bionano assembly<sup>124</sup>. Misassembled regions were detected when only one read from multiple LMPs mapped within a contig or when discrepancies occurred between the insert size and mapping distance of the reads. They can also be useful to identify structural variants<sup>125-127</sup> such as large deletions and chromosomal rearrangements. However, reviewers on the wheat CS42 genome publication thought it necessary to confirm translocations using PCR of nullisomics, rather than trust the hundreds of different LMP reads from two independent LMP libraries confirming the chromosomal break points.

## 6.2 Long range spatial information

It is unlikely that scientists will settle on one single sequencing strategy to give optimal contiguity in *de novo* genome projects and in many cases hybrid approaches involving multiple technologies will be employed. Several complementary strategies are currently available to scientists which generate long range spatial information to help genome scaffolding.

### 6.2.1 Optical mapping

Systems such as the Bionano Saphyr were developed to create optical maps to aid genome assembly and have been used to good effect in aiding plant genome assembly<sup>8,9,124,128,129</sup>. They work on a similar principle to restriction maps and produce optical maps for molecules up to 500 Kbp in length. DNA is nicked using a restriction endonuclease chosen to target cuts on average 12 Kbp apart. Powerful microscopes are then used to determine the size in base pairs between the nicks. Molecules are then merged based on similarity of restriction patterns in the same way that a MTP of BACs would be after fingerprinting and this information used to order contigs *in silico*.

## 6.2.2 Hi-C

Hi-C is based on chromosome conformation capture and involves *in situ* crosslinking DNA based on proximity and has been used to enhance many genome projects<sup>8,9,130-133</sup>. Ligation, following restriction digest and enrichment, creates chimeric molecules that reveal sequences that were close together in their natural conformation. With sequences closer together more likely to come from the same chromosomal location rather than other chromosomes, this information can be used to guide the assembly.

Dovetail Genomics took this a stage further and developed an *in vivo* method which was used to good effect in the *S. verrucosum* assembly<sup>10</sup>. Costing £20,000+ to construct a library, this is not a cheap protocol but recently they announced the launch of a commercially available kit so it will be interesting to see how much this will cost and the uptake within the scientific community.

## 6.2.3 Single molecule sequencers

Increasing read length is the simplest means of resolving repeats and if a read can be of sufficient length that it identifies unique sequence flanking a repeat, then that repeat can be resolved. Third generation, single-molecule sequencers offer great potential in this sphere. They sequence native DNA so do not require any amplification steps and with the potential to sequence molecules >25 Kbp they could hold the key to completing genome assembly projects.

Launched in 2011, PacBio technology uses hairpin adapters ligated to DNA fragments in its real-time sequencing by synthesis method to sequence single molecules in zero mode waveguides (ZMW)<sup>134</sup>. RSII instruments can sequence molecules >20 Kbp and generate >600 Mbp per SMRT cell whilst the newer Sequel instrument can sequence molecules >10 Kbp generating >2 Gbp per SMRT cell. Recent studies producing polymerase read lengths >90 Kbp highlight the potential of these instruments<sup>121</sup>.

In 2014, ONT introduced their USB driven, handheld MinION device. Based on nanopores through which single stranded molecules of DNA can pass, the change in current detected passing through the pores can be related to the composition of the

nucleotides residing within the pore. The standard flow cells have 512 pores through which DNA molecules could pass and this technology has great potential for scalability. The company have subsequently launched a GridION capable of housing 5 standard flow cells and a PromethION which can run 48 much larger flow cells, each with >3,000 pores. Despite early issues with flow cell stability, the portable nature of this device coupled with claims that as much 10 Gbp of data can be generated on a standard flow cell and with recent claims of a 1.5 Mbp read mapping back with 90 %+ accuracy to the human genome<sup>135</sup>, this is a technology that promises much.

When launched both the PacBio and MinION had high error rates, with accuracy around 80%, leading to some scientists generating complimentary Illumina data in addition to the long-range sequence. Some used this to error correct the reads ahead of assembly while others used it in KAT plots to validate the assemblies.

These single molecule, long-read technologies have the capability to sequence prokaryotes and assemble them into a single contig for <£500 and these are routinely used for such genome projects. Zimin *et al.* used PacBio to aid assembly of wheat but they needed to run 1,000+ SMRT cells and add Illumina sequence data to achieve their most contiguous assembly. This will make this approach prohibitively expensive and time consuming for most large, complex, eukaryotic *de novo* genome projects.

Recent publication of an *Arabidopsis* assembly based on data from a single nanopore flow cell achieved a CN50 of 12.3 Mbp with only 62 contigs<sup>136</sup>. By comparison, we achieved an assembly with a CN50 of 8.6 Mbp with 54 contigs for the *A. columbia* ecotype when using CANU and Nanopolish<sup>59</sup> to assemble MinION data. Both these assemblies were achieved for a FEC <£1,000 and with the manufacturers making ambitious claims about the future potential of these instruments, their ability to generate low cost, highly contiguous assemblies could revolutionise future genome projects.

An additional advantage of single molecule sequencers is their ability to detect nucleotide modifications such as base methylation without the need to manipulate the DNA using methods such as bisulphite treatment<sup>137-140</sup>. In the case of PacBio, the modified bases alter the time in which the fluorophore can be detected in the ZMW compared with unmodified bases. For MinION, the modified bases have a slightly

different structure so they subtly alter the current passing through the nanopore. Being able to identify modified bases can play an important part in genome assembly. In polyploid species where many homeologous genes are present, these may have different modification patterns and identifying these could be a means of distinguishing them and help resolve different paths within a DBG.

#### **6.2.4 Improving assemblies through analysis of the gene space**

In early NGS based prokaryotic genome projects, identifying open reading frames (ORFs) could be used to help improve assemblies. With very little non-coding sequence, up to 85 % of a prokaryotic genome is unique. Searching for the ends of contigs for partial ORFs can result in the ability to order contigs without the need to construct a LMP library. Gaps can be closed using PCR followed by Sanger sequencing and this strategy was used to sequence the *C. botulinum* strain submitted as part of this thesis. Today, annotation of genomes and some biology can be required to publish a genome so scientists have returned to RNA data to complement genome assemblies.

In eukaryotes, exons are interspersed by introns in the DNA sequence but spliced out in mRNA and the order of exons in the transcripts can help validate assemblies. Traditional methods such as RNAseq<sup>141</sup> have been replaced by IsoSeq<sup>142,143</sup> to catalogue all transcript isoforms. RNA is isolated from a variety of different tissues and full length cDNAs synthesised which are then converted into PacBio or MinION compatible libraries and sequenced. The reads have the potential to help with annotation and, as the exons must appear in the correct order within the genome, they can help verify the assembly and in polyploid species this can help guide the correct path in DBGs.

## 6.3 Linked Reads

Assembly of haploid organisms is straightforward but as ploidy increases, the presence of SNPs can cause problems. When SNPs are detected they create bubbles in DBGs. When they are further apart than the spatial information provided by the sequence data, it becomes impossible to phase them. Being able to phase SNPs helps resolve paths within a DBG and improves contiguity.

### 6.3.1 Using standard paired-end and LMP libraries

Sequence data from each of the NGS library types constructed using the methods presented in this thesis can phase SNPs but each has its limitations. BACs being haploid provide the ability to phase SNPs across the entirety of their insert. In some cases, this can be >200 Kbp. Theoretically, it should be possible to pool non-overlapping BACs from a MTP and sequence them to phase all the SNPs. The downside of this would be the difficulty in ensuring that all the BACs were equimolar pooled.

Both SE-APE and ELF LMPs are limited by the read and insert length. Delaneau *et al.* reported that using a mixture of 300 bp, 500 bp and 1 Kbp inserts sequenced with 2x 100 bp reads, 70 % of known heterozygous SNPs <1 Kbp apart could be phased<sup>144</sup>. As expected, most of the SNPs that couldn't be phased fell between 600 and 800 bp apart. This was the region that their strategy didn't cover. With SE-APE libraries sequenced with 2x 250 bp reads and spanning 600 bp to 1.2 Kbp, it is theoretically possible to phase SNPs up to 1 Kbp apart if sufficient coverage is generated. For ELF LMPs the overlap between libraries from each fraction and the linked nature of the reads would suggest that by sequencing all twelve fractions with sufficient coverage would phase SNPs between the smallest and largest inserts.

As not all communities have access to BAC libraries and the cost to sequence all 12 fractions from the ELF based LMP library construction protocol are considerable, these methods are unlikely to be adopted.

### 6.3.2 Introducing the 10x Genomics Chromium

Long reads such as those generated by the single molecule sequencers have the potential to resolve SNPs in the same way they resolve large repeats, but the most interesting development in this field is 10x Genomics Chromium. It partitions DNA molecules into thousands of micelles containing individually barcoded gel beads and resultant libraries can be sequenced on an Illumina sequencer.

Sequence reads sharing the same barcode can be grouped together as coming from the same micelle, and potentially from the same molecule, and this information used to phase SNPs and can be used to improve genome contiguity. Using this technology to generate sequence for *S. verrucosum* to complement the DISCOVAR plus LMP assembly increased the SN50 >5-fold to 4.7 Mbp.

The principle of the technology is very similar to that when sequencing individual BACs. The chances of two molecules entering the same micelle with the same repeat is low, so the technology should simplify *de novo* genome assembly for complex genomes. Input requirements are low at 1 ng for a 3 Gbp genome, but optimal conditions require DNA molecules >50 Kbp and the presence of any small molecules can reduce efficiency. These smaller molecules tend to occur at a much greater copy number than larger molecules and can occupy a large proportion of micelles complicating assembly.

The size distribution of molecules can hinder trying to optimise algorithms to assemble 10x data. In some cases, molecules can be as short as 10 Kbp and in others >150 Kbp. As molecules entering the micelles are not sequenced in their entirety, determining the spatial information the data provides can be problematic. Libraries also require an amplification step and this can introduce biases so these need to be considered when assembling the data. Work is currently underway for wheat and it will be interesting to see what results it achieves and how many libraries need to be constructed to achieve optimal contiguity.

## 6.4 DNA integrity

As sequence read lengths increase, there is an ever-growing demand for longer and longer DNA molecules. Accurately determining DNA molecule length can be problematic as devices such as the Agilent TapeStation can be affected by the amount of DNA loaded and the most reliable method, Pulse Field Gel Electrophoresis<sup>145</sup> (PFGE) is cumbersome, time consuming taking up to 16 hours to run and requires >200 ng DNA. The newer Advanced Analytical Femto Pulse<sup>146</sup> shows promise with sub ng input requirements and the ability to determine molecular weight up to 200 Kbp. Run times are only 70 minutes for up to 11 samples, but it will take time to see how robust the system is.

For *Arabidopsis*, our best MinION results were achieved after growing seedlings for 10 days post germination, followed by 48 hours in the dark to deplete starch levels and then constructing libraries immediately after DNA extraction. Our best LMP outputs were also achieved when using freshly extracted DNA. Whether this is practical or not remains to be seen, but must be a consideration for all *de novo* genome assembly projects.

To maximise outputs and make the best use of technologies available, improving extraction protocols to increase DNA molecule length is an area of science that will require significant investment in the coming months and years.

## 6.5 Future strategies for *de novo* genome projects

With the cost of assembling a genome continuing to drop, we are entering the pangenomic era for even complex genomes such as wheat<sup>147</sup>. The library construction costs and DNA requirements for those protocols discussed in this chapter and undertaken at EI are shown in Table 6.1. These are going to be pivotal in deciding which strategies get adopted by the wider scientific community.

Library Type	Cost of Library Construction (£)	DNA requirements
<b>SE-APE</b>	250	1 µg >10 Kbp
<b>LMP</b>	1,500	9 µg >45 Kbp
<b>PacBio</b>	345	10 µg >60 Kbp
<b>10x Chromium</b>	900	1 ng >50 Kbp

**Table 6.1:** The current cost and DNA requirements for different NGS library construction methods at EI.

For many, it will be cost that dictates which approach to use and for others the inability to extract HMW DNA that will prove problematic. Decisions also need to be made about how many libraries need to be constructed and sometimes technologies are not suitable. Many are optimised for human genomes and it is not always straightforward to construct libraries and then interrogate the data for alternative genome sizes and different levels of complexity. A good example is the 10X Genomics Chromium platform. It was designed as part of a pipeline to construct Illumina ready libraries for 3 Gbp genomes. Some early adopters struggled extracting DNA >50 Kbp and protocols have been reworked to allow genomes <3 Gbp to be processed. For genomes >3 Gbp, multiple libraries need to be constructed and we are currently evaluating using up to six libraries for wheat. This starts to add significant costs to a project and any increases in contiguity will need to help justify this expense.

For those wanting to produce multiple highly contiguous *de novo* genome assemblies, the methods presented in this thesis provide a good starting point in terms of cost and contiguity, especially to those with NGS access limited to Illumina instruments. For a 3 Gbp diploid, mammalian genome using the library construction approaches presented in this thesis, and sequencing to a combined 50x coverage across paired-end and LMP libraries, would cost <£10,000 and the data could be generated inside a week. While this may sound expensive, for orphan genomes where there is not a reference or closely related species to compare the outputs against, the necessity to generate amplification free data and maximise the spatial information as provided by these libraries supports this approach. The comparable cost to sequence using PacBio with 30x genome coverage, with average 10 Kbp reads, would be >£65,000 and take

two weeks. I expect that both would produce SN50s >1 Mbp and identify >90 % of the single copy orthologs complete with <1 % fragmented upon BUSCO analysis. Whether either of these strategies will be chosen for any aspect of the EBP is open to debate.

Many scientists, however, will want to add complementary technologies to improve contiguity further and each will have their own strategy based on the repeat content of their genome of interest. In sequencing potato, there was not one single approach resulting in the best contiguity, indicating that a combination of different techniques would be required for optimal assembly. In the future, if PacBio and MinION deliver on their projections, these platforms will be popular. The MinION has the capacity for very long reads. It has the potential to sequence an entire chromosome from start to finish if one can be isolated from a cell intact and remain suitable for sequencing. It is an exciting thought that we may be able to achieve this.

## 6.6 Summary

This thesis represents a knowledge and understanding of molecular biology that theoretically dates back 40 years and practically nearly 30 years. I have described my work on the development of several novel library construction protocols associated with improving *de novo* genome assembly using Illumina sequencing technology. I have established that these innovative protocols are relatively cheap, more robust, help generate more contiguous and accurate assemblies and assemble more content, outperforming comparable published strategies. They have evolved as technology has improved and combined, they have made a significant impact on genome contiguity in the wheat and barley genome projects presented. Over the next couple of years, I am optimistic that these protocols will continue to contribute toward numerous high-profile genome project publications.

The genomes discussed in this thesis have been made publicly available so are free to use for academic and commercial plant scientists alike and I truly hope that this will lead to new varieties of both barley and wheat being bred that will improve food security. Like all contemporary methods, the protocols will have a finite life span and already newer technologies are coming to the fore that have the potential to

supersede them, but that is the excitement of science. It is heartening to think that tasks that took years when I first started working in a genomics laboratory can now be completed in days. With the same rate of progression, the future promises much and I am proud of having played my part, however small, in the genomics revolution.

# Definitions

<b>BAC:</b>	Bacterial Artificial Chromosome
<b>BES:</b>	BAC End Sequencing
<b>BUSCO:</b>	Benchmarking Universal Single Copy Orthologs
<b>BWA:</b>	Burrows-Wheeler Alignment
<b>CN50:</b>	Contig N50
<b>CS42:</b>	Chinese Spring 42
<b>CTAB:</b>	Cetyltrimethyl ammonium bromide
<b>DBG:</b>	de Bruijn Graph
<b>DNA:</b>	Deoxyribonucleic Acid
<b>EBP:</b>	Earth BioGenome Project
<b>EI:</b>	Earlham Institute
<b>ELF:</b>	Electrophoretic Lateral Fractionator
<b>FEC:</b>	Full Economic Cost
<b>FosIII:</b>	Fosmid Library by Illumina
<b>FPC:</b>	Fingerprint Contigs
<b>FSC:</b>	Flow Sorted Chromosome
<b>FUs:</b>	Fluorescent Units
<b>HGP:</b>	Human Genome Project
<b>HMW:</b>	High Molecular Weight
<b>IBSC:</b>	International Barley Sequencing Consortium
<b>IWGSC:</b>	International Wheat Genome Sequencing Consortium
<b>KAT:</b>	<i>K</i> -mer Analysis Tool
<b>LB:</b>	Luria Broth
<b>LINE:</b>	Long Interspersed Nuclear Element
<b>LITE:</b>	Low Input, Transposase Enabled

<b>LMP:</b>	Long Mate Pair
<b>LMW:</b>	Low Molecular Weight.
<b>LTR:</b>	Long Terminal Repeat
<b>MDA:</b>	Multiple, Displacement Amplification
<b>MTP:</b>	Minimal Tile Path
<b>NGS:</b>	Next Generation Sequencing
<b>OLC:</b>	Overlap Consensus
<b>Oligo:</b>	Oligonucleotide
<b>ONT:</b>	Oxford Nanopore Technology
<b>ORF:</b>	Open Reading Frame
<b>PacBio:</b>	Pacific Biosciences
<b>PCR:</b>	Polymerase Chain Reaction
<b>PEG:</b>	Poly Ethylene Glycol
<b>PFGE:</b>	Pulse Field Gel Electrophoresis
<b>PGSC:</b>	Potato Genome Sequencing Consortium
<b>PhiX:</b>	PhiX 174 bacteriophage
<b>PNK:</b>	Phospho Nucleotide Kinase
<b>QPCR:</b>	Quantitative Polymerase Chain Reaction
<b>RNA:</b>	Ribonucleic acid
<b>RSII:</b>	Real-time, Sequencer II
<b>SBS:</b>	Sequencing By Synthesis
<b>SCO:</b>	Single Copy Ortholog
<b>SE-APE:</b>	Size Exclusion-Amplification-free Paired-end
<b>ShARC</b>	Shearing And Recircularisation after Cloning
<b>SINE:</b>	Short Interspersed Nuclear Element
<b>SMRT:</b>	Single Molecule, Real Time
<b>SN50:</b>	Scaffold N50

<b>SNP:</b>	Single Nucleotide Polymorphism
<b>SPRI:</b>	Solid Phase, Reversible Immobilisation
<b>TALL:</b>	Tight, Amplification-free, Large-insert Libraries
<b>TGAC:</b>	The Genome Analysis Centre
<b>W2RAP:</b>	Wheat/ Whole-genome Robust Assembly Pipeline
<b>WGS:</b>	Whole Genome Sequencing
<b>WLA:</b>	Wheat line A
<b>ZMW:</b>	Zero Mode Wavelength

# Glossary

**3DL:** The long arm of wheat chromosome 3D.

**Blunt-end:** A DNA molecule that doesn't have a 5' or 3' overhang.

**Contig:** A continuous sequence of overlapping, merged sequence reads.

**Ct Value:** The cycle threshold during qPCR at which amplification is detected above background noise.

**FEC:** The cost including all overheads such as labour, consumables and depreciation.

**FPC:** A Fingerprint Contig is the process used to identify common restriction patterns within a BAC clone and enables the identification of BACs that share sequence which can then be ordered in a minimal tiling path.

**GC content:** The percentage of nucleotides that are either Guanine or Cytosine within a genome or given stretch of DNA.

**Genome:** All the genetic information for a given organism.

**Genome Assembly:** Piecing together genome sequence to faithfully reconstruct the genome.

**(Whole) Genome Sequencing:** The process which reveals all the sequence of a given genome.

**Genome Size:** The total number of base pairs within a genome.

**Hamming distance:** The number of differences between two strings of the same length. For a DNA sequence of 9 bp in length that has a Hamming distance of 4, then at least 4 bp will differ between the two strings.

**K-mer:** A string of nucleotides of length  $k$ .

**IsoSeq:** Next Generation Sequencing typically on the PacBio to identify all RNA isoforms.

**MTP:** A minimal tile path is the fewest number of BAC clones required to fully cover every base within a chromosome/ genome.

**N:** A nucleotide position where it hasn't been possible to determine whether it is an Adenine, Guanine, Thymine or Cytosine.

**Next Generation Sequencing:** Technologies capable of massively parallel sequencing commercialised since 2005. They include the Roche 454 pyrosequencer and the Illumina HiSeq.

**Nucleotide:** The single monomer building block of DNA.

**Paired-end read:** A pair of reads that come from either end of the same DNA molecule.

**Phasing SNPs:** The linking of two SNPs onto the same chromosomal copy.

**Polymerase Chain Reaction:** An *in vivo* process to specifically clone target loci of interest.

**Primer:** A stretch of synthetic DNA typically used in PCR to amplify a locus of interest.

**Primer walking:** The process whereby an unknown DNA molecule is inserted into a vector such as a plasmid. Sequence is first generated using a primer anchored within the vector to sequence out, into the unknown DNA molecule. This sequence is then used to design a new primer and this then used to further sequence into the DNA and the whole process repeated until the entire sequence of the original unknown DNA molecule is determined.

**Q30:** A quality score assigned to an Illumina sequencing read. It is equivalent to a sequencing error once in one thousand base pairs or 99.9 % accuracy.

**Quantitative PCR:** The real-time measurement of the products of PCR during cycling.

**Repetitive DNA:** Any DNA sequences that occurs more than once within a genome. Can range from simple dinucleotide repeats to tandem duplications of several hundred thousand base pairs.

**RNASeq:** Next Generation Sequencing of RNA.

**Sanger sequencing:** The method of sequencing DNA using chain terminating dideoxynucleotides developed by Fred Sanger.

**Scaffold:** An ordered sequence of contigs separated by gaps of known length.

**Single-end read:** A sequence read that is from the single-end of a DNA molecule.

**SNP:** A Single Nucleotide Polymorphism is the difference in sequence between two chromosomal copies at a single nucleotide position within a genome.

**Tagmentation:** The act of fragmenting DNA molecules using a transposase.

**Template:** A DNA molecule which is used to prime from to generate a complementary sequence.

**Third Generation Sequencing:** Technologies capable of sequencing single DNA molecules without the need for amplification.

## References

- 1 Barke, J. *et al.* A mixed community of actinomycetes produce multiple antibiotics for the fungus farming ant *Acromyrmex octospinosus*. *BMC biology* **8**, 109, doi:10.1186/1741-7007-8-109 (2010).
- 2 Carter, A. T. *et al.* Complete genome sequence of the proteolytic Clostridium botulinum type A5 (B3') strain H04402 065. *Journal of bacteriology* **193**, 2351-2352, doi:10.1128/jb.00072-11 (2011).
- 3 Heavens, D. *et al.* Genome sequence of the vertebrate gut symbiont *Lactobacillus reuteri* ATCC 53608. *Journal of bacteriology* **193**, 4015-4016, doi:10.1128/jb.05282-11 (2011).
- 4 Seipke, R. F. *et al.* Draft genome sequence of *Streptomyces* strain S4, a symbiont of the leaf-cutting ant *Acromyrmex octospinosus*. *Journal of bacteriology* **193**, 4270-4271, doi:10.1128/jb.05275-11 (2011).
- 5 Heavens, D., Accinelli, G. G., Clavijo, B. & Clark, M. D. A method to simultaneously construct up to 12 differently sized Illumina Nextera long mate pair libraries with reduced DNA input, time, and cost. *BioTechniques* **59**, 42-45, doi:10.2144/000114310 (2015).
- 6 Clavijo, B. *et al.* W2RAP: a pipeline for high quality, robust assemblies of large complex genomes from short read data. *bioRxiv*, doi:10.1101/110999 (2017).
- 7 Clavijo, B. J. *et al.* An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome research* **27**, 885-896, doi:10.1101/gr.217117.116 (2017).
- 8 Mascher, M. *et al.* A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**, 427-433, doi:10.1038/nature22043 (2017).
- 9 Beier, S. *et al.* Construction of a map-based reference genome sequence for barley, *Hordeum vulgare* L. *Scientific data* **4**, 170044, doi:10.1038/sdata.2017.44 (2017).
- 10 Paajanen, P. *et al.* A critical comparison of technologies for a plant genome sequencing project. *bioRxiv*, doi:10.1101/201830 (2017).
- 11 Lu, F.-H. *et al.* Independent assessment and improvement of wheat genome assemblies using Fosill jumping libraries. *bioRxiv*, doi:10.1101/219352 (2017).
- 12 <https://www.statista.com/statistics/263977/world-grain-production-by-type/>.
- 13 <https://www.statista.com/statistics/190362/total-us-wheat-production-value-from-2000/>.
- 14 <https://esa.un.org/unpd/wpp/Graphs/Probabilistic/POP/TOT/>.
- 15 Sanger, F. *et al.* The nucleotide sequence of bacteriophage phiX174. *Journal of molecular biology* **125**, 225-246 (1978).
- 16 Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science (New York, N.Y.)* **269**, 496-512 (1995).
- 17 Goffeau, A. *et al.* Life with 6000 genes. *Science (New York, N.Y.)* **274**, 546, 563-547 (1996).
- 18 Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science (New York, N.Y.)* **282**, 2012-2018 (1998).
- 19 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815, doi:10.1038/35048692 (2000).

20 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).

21 Venter, J. C. *et al.* The sequence of the human genome. *Science (New York, N.Y.)* **291**, 1304-1351, doi:10.1126/science.1058040 (2001).

22 Waterston, R. H., Lander, E. S. & Sulston, J. E. More on the sequencing of the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 3022-3024; author reply 3025-3026, doi:10.1073/pnas.0634129100 (2003).

23 Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463-5467 (1977).

24 Leamon, J. H. *et al.* A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis* **24**, 3769-3777, doi:10.1002/elps.200305646 (2003).

25 Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380, doi:10.1038/nature03959 (2005).

26 Pinard, R. *et al.* Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC genomics* **7**, 216, doi:10.1186/1471-2164-7-216 (2006).

27 Rothberg, J. M. & Leamon, J. H. The development and impact of 454 sequencing. *Nature biotechnology* **26**, 1117-1124, doi:10.1038/nbt1485 (2008).

28 Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59, doi:10.1038/nature07517 (2008).

29 Kedes, L. & Campany, G. The new date, new format, new goals and new sponsor of the Archon Genomics X PRIZE competition. *Nature genetics* **43**, 1055-1058, doi:10.1038/ng.988 (2011).

30 Kedes, L., Liu, E., Jongeneel, C. V. & Sutton, G. Judging the Archon Genomics X PRIZE for whole human genome sequencing. *Nature genetics* **43**, 175, doi:10.1038/ng0311-175 (2011).

31 Kedes, L. & Liu, E. T. The Archon Genomics X PRIZE for whole human genome sequencing. *Nature genetics* **42**, 917-918, doi:10.1038/ng1110-917 (2010).

32 H., S. The constancy of deoxyribose nucleic acid in plant nuclei. *Proceedings of the National Academy of Sciences* **36**, 643-654 (1950).

33 Hozza M., V. T., Brejová B. How Big is that Genome? Estimating Genome Size and Coverage from k-mer Abundance Spectra. *Lecture Notes in Computer Science* **9309**, doi:doi.org/10.1007/978-3-319-23826-5\_20 (2015).

34 Fay, M. F., Cowan, R. S. & Leitch, I. J. The effects of nuclear DNA content (C-value) on the quality and utility of AFLP fingerprints. *Annals of botany* **95**, 237-246, doi:10.1093/aob/mci017 (2005).

35 Jakob, S. S., Meister, A. & Blattner, F. R. The considerable genome size variation of *Hordeum* species (poaceae) is linked to phylogeny, life form, ecology, and speciation rates. *Molecular biology and evolution* **21**, 860-869, doi:10.1093/molbev/msh092 (2004).

36 Mankertz, A. & Hillenbrand, B. Analysis of transcription of Porcine circovirus type 1. *The Journal of general virology* **83**, 2743-2751, doi:10.1099/0022-1317-83-11-2743 (2002).

37 Philippe, N. *et al.* Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science (New York, N.Y.)* **341**, 281-286, doi:10.1126/science.1239181 (2013).

38 Gregory, T. R. Animal Genome Size Database. <http://www.genomesize.com>. (2018).

39 Xu, X. *et al.* Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189-195, doi:10.1038/nature10158 (2011).

40 Schatz, M. C., Delcher, A. L. & Salzberg, S. L. Assembly of large genomes using second-generation sequencing. *Genome research* **20**, 1165-1173, doi:10.1101/gr.101360.109 (2010).

41 Desai, A. *et al.* Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data. *PLoS one* **8**, e60204, doi:10.1371/journal.pone.0060204 (2013).

42 Biscotti, M. A., Olmo, E. & Heslop-Harrison, J. S. Repetitive DNA in eukaryotic genomes. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **23**, 415-420, doi:10.1007/s10577-015-9499-z (2015).

43 Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews. Genetics* **13**, 36-46, doi:10.1038/nrg3117 (2011).

44 Baucom, R. S. *et al.* Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS genetics* **5**, e1000732, doi:10.1371/journal.pgen.1000732 (2009).

45 Huptas, C., Scherer, S. & Wenning, M. Optimized Illumina PCR-free library preparation for bacterial whole genome sequencing and analysis of factors influencing de novo assembly. *BMC research notes* **9**, 269, doi:10.1186/s13104-016-2072-9 (2016).

46 Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature methods* **6**, 291-295, doi:10.1038/nmeth.1311 (2009).

47 Quail, M. A. *et al.* Optimal enzymes for amplifying sequencing libraries. *Nature methods* **9**, 10-11, doi:10.1038/nmeth.1814 (2011).

48 Caruccio, N. Preparation of next-generation sequencing libraries using Nextera technology: simultaneous DNA fragmentation and adaptor tagging by in vitro transposition. *Methods in molecular biology (Clifton, N.J.)* **733**, 241-255, doi:10.1007/978-1-61779-089-8\_17 (2011).

49 Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome biology* **11**, R119, doi:10.1186/gb-2010-11-12-r119 (2010).

50 Staden, R. A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic acids research* **8**, 3673-3694 (1980).

51 GRANGER G. SUTTON, O. W., MARK D. ADAMS, and ANTHONYR. KERLAVAGE. TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects. *GENOME SCIENCE & TECHNOLOGY* **1** (1995).

52 Huson, D. H. *et al.* Design of a compartmentalized shotgun assembler for the human genome. *Bioinformatics (Oxford, England)* **17 Suppl 1**, S132-139 (2001).

53 Simpson, J. T. *et al.* ABYSS: a parallel assembler for short read sequence data. *Genome research* **19**, 1117-1123, doi:10.1101/gr.089532.108 (2009).

54 Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**, 821-829, doi:10.1101/gr.074492.107 (2008).

55 Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18, doi:10.1186/2047-217x-1-18 (2012).

56 Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the*

United States of America **98**, 9748-9753, doi:10.1073/pnas.171285098 (2001).

57 Zhang, Y. & Waterman, M. S. An Eulerian path approach to global multiple alignment for DNA sequences. *Journal of computational biology : a journal of computational molecular cell biology* **10**, 803-819, doi:10.1089/106652703322756096 (2003).

58 Cherukuri, Y. & Janga, S. C. Benchmarking of de novo assembly algorithms for Nanopore data reveals optimal performance of OLC approaches. *BMC genomics* **17 Suppl 7**, 507, doi:10.1186/s12864-016-2895-8 (2016).

59 Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial genomics* **3**, e000132, doi:10.1099/mgen.0.000132 (2017).

60 Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. & Clavijo, B. J. KAT: A K-mer Analysis Toolkit to quality control NGS datasets and genome assemblies. *bioRxiv*, doi:10.1101/064733 (2016).

61 Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)* **31**, 3210-3212, doi:10.1093/bioinformatics/btv351 (2015).

62 Waterhouse, R. M. *et al.* BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular biology and evolution*, doi:10.1093/molbev/msx319 (2017).

63 <http://www.acgt.me/blog/2015/6/11/l50-vs-n50-thats-another-fine-mess-that-bioinformatics-got-us-into>.

64 Shizuya, H. *et al.* Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 8794-8797 (1992).

65 Coulson, A., Sulston, J., Brenner, S. & Karn, J. Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences of the United States of America* **83**, 7821-7825 (1986).

66 Mayer, K. F. *et al.* A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**, 711-716, doi:10.1038/nature11543 (2012).

67 Sanchez-Martin, J. *et al.* Rapid gene isolation in barley and wheat by mutant chromosome sequencing. *Genome biology* **17**, 221, doi:10.1186/s13059-016-1082-1 (2016).

68 Wicker, T. *et al.* 454 sequencing put to the test using the complex genome of barley. *BMC genomics* **7**, 275, doi:10.1186/1471-2164-7-275 (2006).

69 Steuernagel, B. *et al.* De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley. *BMC genomics* **10**, 547, doi:10.1186/1471-2164-10-547 (2009).

70 Lonardi, S. *et al.* Combinatorial pooling enables selective sequencing of the barley gene space. *PLoS computational biology* **9**, e1003010, doi:10.1371/journal.pcbi.1003010 (2013).

71 Thierry-Mieg, N. A new pooling strategy for high-throughput screening: the Shifted Transversal Design. *BMC bioinformatics* **7**, 28, doi:10.1186/1471-2105-7-28 (2006).

72 Munoz-Amatriain, M. *et al.* Sequencing of 15 622 gene-bearing BACs clarifies the gene-dense regions of the barley genome. *The Plant journal : for cell and molecular biology* **84**, 216-227, doi:10.1111/tpj.12959 (2015).

73 <https://www.beckman.com/reagents/genomic/dna-isolation/plasmid-purification/A37064>.

74 Gibson, U. E., Heid, C. A. & Williams, P. M. A novel method for real time quantitative RT-PCR. *Genome research* **6**, 995-1001 (1996).

75 Heid, C. A., Stevens, J., Livak, K. J. & Williams, P. M. Real time quantitative PCR. *Genome research* **6**, 986-994 (1996).

76 Randerson-Moor, J. A. *et al.* A germline deletion of p14(ARF) but not CDKN2A in a melanoma-neural system tumour syndrome family. *Human molecular genetics* **10**, 55-62 (2001).

77 <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.

78 Mascher, M. *et al.* Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *The Plant journal : for cell and molecular biology* **76**, 718-727, doi:10.1111/tpj.12319 (2013).

79 Karasov, T. L. *et al.* Stability of association between *Arabidopsis thaliana* and *Pseudomonas* pathogens over evolutionary time scales. *bioRxiv*, doi:10.1101/241760 (2018).

80 DeAngelis, M. M., Wang, D. G. & Hawkins, T. L. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic acids research* **23**, 4742-4743 (1995).

81 Hawkins, T. L., O'Connor-Morin, T., Roy, A. & Santillan, C. DNA purification and isolation using a solid-phase. *Nucleic acids research* **22**, 4543-4544 (1994).

82 Whiteford, N. *et al.* An analysis of the feasibility of short read sequencing. *Nucleic acids research* **33**, e171, doi:10.1093/nar/gni170 (2005).

83 Chikhi, R. L., D. Paired-end read length lower bounds for genome resequencing. *BMC bioinformatics* **10**(Suppl 13), doi:doi.org/10.1186/1471-2105-10-S13-O2 (2009).

84 Mullis, K. *et al.* Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor symposia on quantitative biology* **51 Pt 1**, 263-273 (1986).

85 Mullis, K. B. Target amplification for DNA analysis by the polymerase chain reaction. *Annales de biologie clinique* **48**, 579-582 (1990).

86 Saiki, R. K. *et al.* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science (New York, N.Y.)* **239**, 487-491 (1988).

87 Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome biology* **12**, R18, doi:10.1186/gb-2011-12-2-r18 (2011).

88 Aigrain, L., Gu, Y. & Quail, M. A. Quantitation of next generation sequencing library preparation protocol efficiencies using droplet digital PCR assays - a systematic comparison of DNA library preparation kits for Illumina sequencing. *BMC genomics* **17**, 458, doi:10.1186/s12864-016-2757-4 (2016).

89 Bronner, I. F., Quail, M. A., Turner, D. J. & Swerdlow, H. Improved Protocols for Illumina Sequencing. *Current protocols in human genetics* **80**, 18.12.11-42, doi:10.1002/0471142905.hg1802s80 (2014).

90 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

91 [https://software.broadinstitute.org/software/discovar/blog/?page\\_id=375](https://software.broadinstitute.org/software/discovar/blog/?page_id=375).

92 Love, R. R., Weisenfeld, N. I., Jaffe, D. B., Besansky, N. J. & Neafsey, D. E. Evaluation of DISCOVAR de novo using a mosquito sample for cost-effective short-read genome assembly. *BMC genomics* **17**, 187, doi:10.1186/s12864-016-2531-7 (2016).

93 Peng, X. *et al.* The draft genome sequence of the ferret (*Mustela putorius furo*) facilitates study of human respiratory disease. *Nature biotechnology* **32**, 1250-1255, doi:10.1038/nbt.3079 (2014).

94 Chaisson, M. J., Brinza, D. & Pevzner, P. A. De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome research* **19**, 336-346, doi:10.1101/gr.079053.108 (2009).

95 Ng, P. *et al.* Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic acids research* **34**, e84, doi:10.1093/nar/gkl444 (2006).

96 Ng, P., Wei, C. L. & Ruan, Y. Paired-end diTagging for transcriptome and genome analysis. *Current protocols in molecular biology* **Chapter 21**, Unit 21.12, doi:10.1002/0471142727.mb2112s79 (2007).

97 Park, N. *et al.* An improved approach to mate-paired library preparation for Illumina sequencing. Vol. 1 (2013).

98 van Heesch, S. *et al.* Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing. *BMC genomics* **14**, 257, doi:10.1186/1471-2164-14-257 (2013).

99 Tatsumi, K., Nishimura, O., Itomi, K., Tanegashima, C. & Kuraku, S. Optimization and cost-saving in tagmentation-based mate-pair library preparation and sequencing. *BioTechniques* **58**, 253-257, doi:10.2144/000114288 (2015).

100 Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics (Oxford, England)* **27**, 2957-2963, doi:10.1093/bioinformatics/btr507 (2011).

101 Leggett, R. M., Clavijo, B. J., Clissold, L., Clark, M. D. & Caccamo, M. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics (Oxford, England)* **30**, 566-568, doi:10.1093/bioinformatics/btt702 (2014).

102 Sahlin, K., Chikhi, R. & Arvestad, L. Assembly scaffolding with PE-contaminated mate-pair libraries. *Bioinformatics (Oxford, England)* **32**, 1925-1932, doi:10.1093/bioinformatics/btw064 (2016).

103 Clarke, J. D. Cetyltrimethyl ammonium bromide (CTAB) DNA miniprep for plant DNA isolation. *Cold Spring Harbor protocols* **2009**, pdb.prot5177, doi:10.1101/pdb.prot5177 (2009).

104 <https://grassroots.tools/>.

105 Bian, X., T. S., Davey R.P. The Grassroots life science data infrastructure (2017). <https://grassroots.tools>.

106 Visendi, P. *et al.* An efficient approach to BAC based assembly of complex genomes. *Plant methods* **12**, 2, doi:10.1186/s13007-016-0107-9 (2016).

107 Choulet, F. *et al.* Structural and functional partitioning of bread wheat chromosome 3B. *Science (New York, N.Y.)* **345**, 1249721, doi:10.1126/science.1249721 (2014).

108 A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science (New York, N.Y.)* **345**, 1251788, doi:10.1126/science.1251788 (2014).

109 Belova, T. *et al.* Integration of mate pair sequences to improve shotgun assemblies of flow-sorted chromosome arms of hexaploid wheat. *BMC genomics* **14**, 222, doi:10.1186/1471-2164-14-222 (2013).

110 Helguera, M. *et al.* New insights into the wheat chromosome 4D structure and virtual gene order, revealed by survey pyrosequencing. *Plant science : an international journal of experimental plant biology* **233**, 200-212, doi:10.1016/j.plantsci.2014.12.004 (2015).

111 Brenchley, R. *et al.* Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**, 705-710, doi:10.1038/nature11650 (2012).

112 Chapman, J. A. *et al.* A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome biology* **16**, 26, doi:10.1186/s13059-015-0582-8 (2015).

113 Ling, H. Q. *et al.* Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* **496**, 87-90, doi:10.1038/nature11997 (2013).

114 Jia, J. *et al.* *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* **496**, 91-95, doi:10.1038/nature12028 (2013).

115 Zhao, G. *et al.* The *Aegilops tauschii* genome reveals multiple impacts of transposons. *Nature plants* **3**, 946-955, doi:10.1038/s41477-017-0067-8 (2017).

116 Zimin, A. V. *et al.* The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *GigaScience* **6**, 1-7, doi:10.1093/gigascience/gix097 (2017).

117 <https://www.genomeweb.com/informatics/israeli-ag-bio-firm-nrgene-seeks-build-business-proprietary-data-analysis-platform#.WrobrowCfhg>.

118 Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods* **10**, 563-569, doi:10.1038/nmeth.2474 (2013).

119 Koren, S. *et al.* Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome biology* **14**, R101, doi:10.1186/gb-2013-14-9-r101 (2013).

120 Liao, Y. C., Lin, S. H. & Lin, H. H. Completing bacterial genome assemblies: strategy and performance comparisons. *Scientific reports* **5**, 8747, doi:10.1038/srep08747 (2015).

121 Nakano, K. *et al.* Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Human cell* **30**, 149-161, doi:10.1007/s13577-017-0168-8 (2017).

122 <https://www.economist.com/news/21735546-how-map-dna-all-known-plants-and-animal-species-earth-sequencing>.

123 Williams, L. J. *et al.* Paired-end sequencing of Fosmid libraries by Illumina. *Genome research* **22**, 2241-2249, doi:10.1101/gr.138925.112 (2012).

124 Jiao, W. B. *et al.* Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome research* **27**, 778-786, doi:10.1101/gr.213652.116 (2017).

125 Aristidou, C. *et al.* Accurate Breakpoint Mapping in Apparently Balanced Translocation Families with Discordant Phenotypes Using Whole Genome Mate-Pair Sequencing. *PloS one* **12**, e0169935, doi:10.1371/journal.pone.0169935 (2017).

126 Hanscom, C. & Talkowski, M. Design of large-insert jumping libraries for structural variant detection using Illumina sequencing. *Current protocols in human genetics* **80**, 7.22.21-29, doi:10.1002/0471142905.hg0722s80 (2014).

127 Sahlin, K., Franberg, M. & Arvestad, L. Structural Variation Detection with Read Pair Information: An Improved Null Hypothesis Reduces Bias. *Journal of computational biology : a journal of computational molecular cell biology* **24**, 581-589, doi:10.1089/cmb.2016.0124 (2017).

128 Luo, M. C. *et al.* Optical Nano-mapping and Analysis of Plant Genomes. *Methods in molecular biology (Clifton, N.J.)* **1429**, 103-117, doi:10.1007/978-1-4939-3622-9\_9 (2016).

129 Stankova, H. *et al.* BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. *Plant biotechnology journal* **14**, 1523-1531, doi:10.1111/pbi.12513 (2016).

130 Bickhart, D. M. *et al.* Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature genetics* **49**, 643-650, doi:10.1038/ng.3802 (2017).

131 Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science (New York, N.Y.)* **356**, 92-95, doi:10.1126/science.aal3327 (2017).

132 Kaplan, N. & Dekker, J. High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nature biotechnology* **31**, 1143-1147, doi:10.1038/nbt.2768 (2013).

133 Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome research* **26**, 342-350, doi:10.1101/gr.193474.115 (2016).

134 Korlach, J. *et al.* Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 1176-1181, doi:10.1073/pnas.0710982105 (2008).

135 <https://www.genengnews.com/gen-news-highlights/in-excess-of-one-million-a-milestone-in-dna-sequencing/81255302>.

136 Michael, T. P. *et al.* High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *Nature communications* **9**, 541, doi:10.1038/s41467-018-03016-2 (2018).

137 Kelleher, P., Murphy, J., Mahony, J. & van Sinderen, D. Identification of DNA Base Modifications by Means of Pacific Biosciences RS Sequencing Technology. *Methods in molecular biology (Clifton, N.J.)* **1681**, 127-137, doi:10.1007/978-1-4939-7343-9\_10 (2018).

138 Powers, J. G. *et al.* Efficient and accurate whole genome assembly and methylome profiling of *E. coli*. *BMC genomics* **14**, 675, doi:10.1186/1471-2164-14-675 (2013).

139 Rand, A. C. *et al.* Mapping DNA methylation with high-throughput nanopore sequencing. *Nature methods* **14**, 411-413, doi:10.1038/nmeth.4189 (2017).

140 Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nature methods* **14**, 407-410, doi:10.1038/nmeth.4184 (2017).

141 Cloonan, N. & Grimmond, S. M. Transcriptome content and dynamics at single-nucleotide resolution. *Genome biology* **9**, 234, doi:10.1186/gb-2008-9-9-234 (2008).

142 Singh, N. *et al.* IsoSeq analysis and functional annotation of the infratentorial ependymoma tumor tissue on PacBio RSII platform. *Meta gene* **7**, 70-75, doi:10.1016/j.mg.2015.11.004 (2016).

143 Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nature biotechnology* **31**, 1009-1014, doi:10.1038/nbt.2705 (2013).

144 Delaneau, O., Howie, B., Cox, A. J., Zagury, J. F. & Marchini, J. Haplotype estimation using sequencing reads. *American journal of human genetics* **93**, 687-696, doi:10.1016/j.ajhg.2013.09.002 (2013).

145 Schwartz, D. C. & Cantor, C. R. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37**, 67-75 (1984).

146 <https://www.aati-us.com/instruments/femto-pulse/>.

147 Montenegro, J. D. *et al.* The pangenome of hexaploid bread wheat. *The Plant journal : for cell and molecular biology* **90**, 1007-1013, doi:10.1111/tpj.13515 (2017).

## **Appendix 1: Letters of support**

Professor Matt Hutchings  
Biological Sciences  
University of East Anglia  
Norwich Research Park  
Norwich  
NR4 7TJ  
United Kingdom  
Tel: 01603 592257  
[m.hutchings@uea.ac.uk](mailto:m.hutchings@uea.ac.uk)

**Re: Darren Heavens PhD thesis: Molecular Biology Strategies To Aid Assembly In *de novo* Genome Projects**

**22<sup>nd</sup> January 2018**

Dear Sir or Madam.

I am writing to confirm that Darren Heavens played a key role in the three publications listed below. My group initiated a Capacity and Capability Challenge project with EI (then TGAC) shortly after they opened, and my research group worked closely with Darren to develop methods to generate a high-quality genome sequence for the leafcutter ant-associated bacterial species *Streptomyces* S4 using Illumina and 454 (1-2). We also worked closely with Darren to develop methods to Illumina sequence bacterial 16S rDNA metagenetic libraries isolated from plant ants (3).

1. A mixed community of actinomycetes produce multiple antibiotics for the fungus farming ant *Acromyrmex octospinosus*. Barke, J., Seipke, R.F., Gruscow, S., Heavens, D., Drou, N., Bibb, M.J., Goss, R.J.M., Yu, D.W. and Hutchings, M. I. (2010). *BMC Biology* 8:109
1. Draft genome sequence of *Streptomyces* strain S4, a symbiont of the leaf-cutting ant *Acromyrmex octospinosus*. Seipke RF, Crossman L, Drou N, Heavens D, Bibb MJ, Caccamo M, Hutchings MI. *J Bacteriol*. 2011 Aug;193(16):4270-1. doi: 10.1128/JB.05275-11. Epub 2011 Jun 17. PMID:21685285
2. Analysis of the bacterial communities associated with two ant-plant symbioses. Seipke R.F., Barke J., Heavens D., Yu D.W. and Hutchings M.I. (2013). *Microbiology Open*. 2:276-83.

Yours faithfully



Matt Hutchings



Quadram Institute Bioscience  
Norwich Research Park  
Colney  
Norwich NR4 7UA  
UK

[www.quadram.ac.uk](http://www.quadram.ac.uk)

22/01/18

Dear Colleague

Re: Complete genome sequence of the proteolytic Clostridium botulinum type A5 (B3') strain H04402 065. Carter AT, Pearson BM, Crossman LC, Drou N, **Heavens D**, Baker D, Febrer M, Caccamo M, Grant KA, Peck MW. J Bacteriol. 2011 May;193(9):2351-2. doi: 10.1128/JB.00072-11. Epub 2011 Mar 4. PMID:21378191.

Darren Heavens was instrumental in helping to bring this collaborative project between the Institute of Food Research (now renamed the Quadram Institute) and TGAC (now renamed the Earlham Institute) to fruition. He advised us on bacterial genomic DNA extraction methods and other experimental procedures and in doing so demonstrated a great depth of molecular biology knowledge. Darren undertook all aspects of the 454 NGS library construction and sequencing and provided text for the publication, so clearly earned the right to be one of the authors of this paper. I believe that this paper was either the first, or at least one of the first complete bacterial genomes to be published by TGAC.

I am very happy to endorse Darren in his bid to gain a PhD by publication.

Yours sincerely,

A. T. Carter

**ANDY CARTER**

**Quadram Institute Bioscience,  
Norwich Research Park, NR4 7UA**

T: +44 (0)1603 255000

D: +44 (0)1603 255398

[andrew.carter@quadram.ac.uk](mailto:andrew.carter@quadram.ac.uk)

[www.quadram.ac.uk](http://www.quadram.ac.uk)

Quadram Institute Bioscience is a registered charity (No. 1058499)  
and a company limited by guarantee (registered in England and Wales No. 03009972).  
VAT registration No. GB 688 8914 52



Quadram Institute Bioscience  
Norwich Research Park  
Colney  
Norwich NR4 7UA  
UK

[www.quadram.ac.uk](http://www.quadram.ac.uk)

Norwich, 23<sup>rd</sup> January 2018

To whom it may concern,

Re: Genome sequence of the vertebrate gut symbiont *Lactobacillus reuteri* ATCC 53608. Heavens D, Tailford LE, Crossman L, Jeffers F, Mackenzie DA, Caccamo M, Juge N. J Bacteriol. 2011 Aug;193(15):4015-6. doi: 10.1128/JB.05282-11. Epub 2011 May 27. PMID:21622738.

Darren Heavens was an integral part of the collaborative project between TGAC (now Earlham Institute) and IFR (now Quadram Institute Biosciences) that delivered the above publication and such was his input to the project that I felt it warranted first authorship on the paper.

His advice on DNA extraction was based on an excellent understanding of the basic principles of molecular biology and was fundamental to the successful outcome. Following DNA extraction, he undertook all aspects of 454 library construction and sequencing delivering an excellent assembly.

This work directly led to the completion of the *L. reuteri* ATCC 53608 genome and follow-up comparative genomic analyses (BMC Genomics. 2015 16:1023) underpinning a DTP PhD studentship due to start in October.

I applaud Darren's decision to undertake a PhD by publication and think that he is a very creditable candidate.

Please don't hesitate to contact me if you need more information,

Yours sincerely,

Nathalie Juge  
(QIB Research Leader)

Quadram Institute Bioscience is a registered charity (No. 1058499)  
and a company limited by guarantee (registered in England and Wales No. 03009972).  
VAT registration No. GB 688 8914 52



Department of Life Sciences  
The Natural History Museum  
Cromwell Road, London SW7 5BD, UK

tel: +44 20 7942 5677

[matt.clark@nhm.ac.uk](mailto:matt.clark@nhm.ac.uk)

**Re: Darren Heavens**

Monday, 26 February 2018

To whom it may concern,

I am writing this letter in support of Darren Heavens. Darren was the first member of my team when I started at TGAC (later renamed the Earlham Institute) and is a fountain of detailed technical knowledge and ideas. His work developing new techniques helped us secure numerous grants and publications, for an example of just one technique consider the following methods paper:

**Darren Heavens**, Gonzalo Garcia Accinelli, Bernardo Clavijo, and Matthew Derek Clark. A method to simultaneously construct up to 12 differently sized Illumina Nextera long mate pair libraries with reduced DNA input, time, and cost. *BioTechniques* 2015.

Darren helped design the initial experiments, did all of the lab work in this paper, spearheaded the publication of this study including writing the paper, and gave intellectual input through the study. This method greatly improved the data quality for scaffolding genome assemblies and was an instrumental part in our successful assembly of the Barley and Wheat genome publications and our submitted publication comparing plant genome assembly methodologies, and developing resources to benchmark Wheat genome assembly accuracy. So far these publications have been cited over 130 times, with the Wheat and Barley genomes also being by far the most used plant genomic resources at ENSEMBL (over 100,000 page impressions).

1. Martin Mascher, Heidrun Gundlach, Axel Himmelbach, Sebastian Beier, Sven O Twardziok, Thomas Wicker, Volodymyr Radchuk, Christoph Dockter, Pete E Hedley, Joanne Russell, Micha Bayer, Luke Ramsay, Hui Liu, Georg Haberer, Xiao-Qi Zhang, Qisen Zhang, Roberto A Barrero, Lin Li, Stefan Taudien, Marco Groth, Marius Felder, Alex Hastie, Hana Šimková, Helena Staňková, Jan Vrána, Saki Chan, María Muñoz-Amatriaín, Rachid Ounit, Steve Wanamaker, Daniel Bolser, Christian Colmsee, Thomas Schmutz, Lala Aliyeva-Schnorr, Stefano Grasso, Jaakko Tanskanen, Anna Chailyan, Dharanya Sampath, **Darren Heavens**, Leah Clissold, Sujie Cao, Brett Chapman, Fei Dai, Yong Han, Hua Li, Xuan Li, Chongyun Lin, John K McCooke, Cong Tan, Penghao Wang, Songbo Wang, Shuya Yin, Gaofeng Zhou, Jesse A Poland, Matthew I Bellgard, Ljudmilla Borisjuk, Andreas Houben, Jaroslav Doležel, Sarah Ayling, Stefano Lonardi, Paul Kersey, Peter Langridge, Gary J Muehlbauer, Matthew D Clark, Mario Caccamo, Alan H Schulman, Klaus FX Mayer, Matthias Platzer, Timothy J Close, Uwe Scholz, Mats Hansson, Guoping Zhang, Ilka Braumann, Manuel Spannagl, Chengdao Li, Robbie Waugh, Nils Stein. A chromosome conformation capture ordered sequence of the barley genome. *Nature* 2017 (75 citations to date)
2. Sebastian Beier, Axel Himmelbach, Christian Colmsee, Xiao-Qi Zhang, Roberto A Barrero, Qisen Zhang, Lin Li, Micha Bayer, Daniel Bolser, Stefan Taudien, Marco Groth, Marius Felder, Alex Hastie, Hana Šimková, Helena Staňková, Jan Vrána, Saki Chan, María Muñoz-Amatriaín, Rachid Ounit, Steve Wanamaker, Thomas Schmutz, Lala Aliyeva-Schnorr, Stefano Grasso, Jaakko Tanskanen, Dharanya Sampath, **Darren Heavens**, Sujie Cao, Brett Chapman, Fei Dai, Yong Han, Hua Li, Xuan Li, Chongyun Lin, John K McCooke, Cong Tan, Songbo Wang, Shuya Yin, Gaofeng

Zhou, Jesse A Poland, Matthew I Bellgard, Andreas Houben, Jaroslav Doležel, Sarah Ayling, Stefano Lonardi, Peter Langridge, Gary J Muehlbauer, Paul Kersey, Matthew D Clark, Mario Caccamo, Alan H Schulman, Timothy J Close, Mats Hansson, Guoping Zhang, Ilka Braumann, , Chengdao Li, Robbie Waugh, Uwe Scholz, Nils Stein, Martin Mascher. M.Sci Data. 2017

3. Bernardo J Clavijo, Luca Venturini, Christian Schudoma, Gonzalo Garcia Accinelli, Gemy Kaithakottil, Jonathan Wright, Philippa Borrill, George Kettleborough, **Darren Heavens**, Helen Chapman, James Lipscombe, Tom Barker, Fu-Hao Lu, Neil McKenzie, Dina Raats, Ricardo H Ramirez-Gonzalez, Aurore Coince, Ned Peel, Lawrence Percival-Alwyn, Owen Duncan, Josua Trösch, Guotai Yu, Dan M Bolser, Guy Namaati, Arnaud Kerhornou, Manuel Spannagl, Heidrun Gundlach, Georg Haberer, Robert P Davey, Christine Fosker, Federica Di Palma, Andrew L Phillips, A Harvey Millar, Paul J Kersey, Cristobal Uauy, Ksenia V Krasileva, David Swarbreck, Michael W Bevan, Matthew D Clark. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Research* 2017 (56 citations to date)
4. Fu-Hao Lu, Neil McKenzie, George Kettleborough, **Darren Heavens**, Matthew D Clark, Michael W Bevan. Independent assessment and improvement of wheat genome assemblies using Fosill jumping libraries. *BMC Genomics* (under review) & bioRxiv 2017
5. Pirita Paajanen, George Kettleborough, Elena Lopez-Girona, Michael Giolai, **Darren Heavens**, David Baker, Ashleigh Lister, Gail Wilde, Ingo Hein, Iain Macaulay, Glenn J Bryan, Matthew D Clark. A critical comparison of technologies for a plant genome sequencing project. *bioRxiv* 2017.

I consider Darren to be a very strong candidate for a PhD, and support his application wholeheartedly. I'm happy to provide further information if needed.

All the best,



Matthew D. Clark

February 23, 2018

To whom it may concern:

Re: Mr. Darren Heavens, Ph.D. by Publication

I have worked since 2012 on complex genome assembly at Earlham Institute, formerly The Genome Analysis Centre, and had the pleasure to collaborate with Darren in a number of projects. His work on developing and optimizing sequencing methods has been fundamental to my work, and that of many others. Over the years, Darren's understanding of molecular biology and laboratory techniques has resulted in a number of key library construction protocols, but three of them are particularly relevant to our shared work: a method for PCR-free paired-end large fragment size libraries (TALL), an improvement on the Illumina PCR-free method to deliver low-bias fragments with large size and, crucially, a method to simultaneously construct multiple Illumina Nextera long mate pair libraries or a range of sizes up to 18Kbp with reduced DNA input and cost.

We have generated methods to make use of Darren's PCR-free and Nextera long mate paired data generation techniques, which enabled the development of our w2rap pipeline for complex genomes. Our estimation is that almost 50% of the improvements on our wheat genome assemblies can be directly and solely attributed to the quality of the data these new techniques produced.

I would like to highlight Darren's contribution to these particular publications:

**W2RAP: a pipeline for high quality, robust assemblies of large complex genomes from short read data.** Bernardo Clavijo, Gonzalo Garcia Accinelli, Jonathan Wright, Darren Heavens, Katie Barr, Luis Yanes, Federica Di Palma. bioRxiv 110999; doi: <https://doi.org/10.1101/110999>: Darren produced all the key datasets, wrote the corresponding parts of the methods section, and generally enabled the rest of this work, participating on the manuscript's writing.

**A method to simultaneously construct up to 12 differently sized Illumina Nextera long mate pair libraries with reduced DNA input, time, and cost.** Heavens D, Accinelli GG, Clavijo B, Clark MD. *Biotechniques*. 2015 Jul 1;59(1):42-5. doi: 10.2144/000114310. eCollection 2015 Jul. PMID:26156783: This is the publication of Darren's Nextera long-mate-paired protocol. He designed the methods, did all the lab work, and wrote the publication, with minimal support from the rest of the authors who mainly contributed QC and general ideas.

**An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations.** Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, Borrill P, Kettleborough G, Heavens D, Chapman H, Lipscombe J, Barker T, Lu FH, McKenzie N, Raats D, Ramirez-

Earlham Institute  
Norwich Research Park  
Norwich,  
NR4 7UG, UK

Tel: +44 (0)1603 450001  
Fax: +44 (0)1603 450021  
[www.earlham.ac.uk](http://www.earlham.ac.uk)  
[enquiries@earlham.ac.uk](mailto:enquiries@earlham.ac.uk)



Decoding Living Systems

Gonzalez RH, Coince A, Peel N, Percival-Alwyn L, Duncan O, Trösch J, Yu G, Bolser DM, Namaati G, Kerhornou A, Spannagl M, Gundlach H, Haberer G, Davey RP, Fosker C, Palma FD, Phillips AL, Millar AH, Kersey PJ, Uauy C, Krasileva KV, Swarbreck D, Bevan MW, Clark MD. *Genome Res.* 2017 May;27(5):885-896. doi: 10.1101/gr.217117.116. PMID: 28420692: The genome assembly on this publication, which my team and I put together, is the direct result of Darren's techniques for data generation. This manuscript is the culmination of years of work at EI where his efforts have been central. He wrote all corresponding methods for library preparation of the genome assembly dataset and greatly contributed to this publication.

If you need any further information regarding this matter, please feel free to contact me.

Yours sincerely,



Bernardo J. Clavijo  
Assembly and Algorithms Development Group Leader, Earlham Institute



Earlham Institute  
Registered in England No. 06855533  
VAT Registration No. 100 1080 05  
Registered Charity No. 1136213



New Road, East Malling  
Kent ME19 6BJ  
T. +44 (0)1732 843833  
enquiries@emr.ac.uk  
@emrcomms  
www.emr.ac.uk

11 March 2018

To Whom It May Concern

**Re: Darren Heavens – PhD candidate**

I am writing this letter in support of Darren Heavens as a candidate for a PhD degree at the University of East Anglia. I met Darren for the first time at the launch of The Genome Analysis Centre (TGAC, now Earlham Institute) in July 2009. Since then I had the opportunity to work with Darren on a number of projects, initially from my role as Head of Bioinformatics at TGAC and later as Director of the organization until my departure in August 2015. This work supported several high-quality manuscripts. These publications speak on their own about the quality of Darren's work but I would also like to add that I have always been impressed with his dedication and commitment to deliver excellence in his work.

Darren's expertise in laboratory methods and new techniques is another aspect of his strength as a scientist. In collaboration with the Institute of Food Research (IFR) we generated and analysed the genomes for new strains of bacterial species (*Clostridium botulinum* and *Lactobacillus reuteri*). Darren led the work on library preparation and sequencing but he also contributed in the design of the experiments. These results were published in the following two manuscripts:

Complete genome sequence of the proteolytic *Clostridium botulinum* type A5 (B3') strain H04402 065. Carter AT, Pearson BM, Crossman LC, Drou N, **Heavens D**, Baker D, Febrer M, Caccamo M, Grant KA, Peck MW. J Bacteriol. 2011 May;193(9):2351-2. doi: 10.1128/JB.00072-11. Epub 2011 Mar 4. PMID:21378191

Genome sequence of the vertebrate gut symbiont *Lactobacillus reuteri* ATCC 53608. **Heavens D**, Taitford LE, Crossman L, Jeffers F, Mackenzie DA, Caccamo M, Juge N. J Bacteriol. 2011 Aug;193(15):4015-6. doi: 10.1128/JB.05282-11. Epub 2011 May 27. PMID:21622738

Darren leadership in the implementation of automation for DNA sequencing and library construction contributed to a fundamental achievement that helped to establish TGAC as a globally recognised research institute. This work was one of the key objectives of the organisation in my time as Director of TGAC. Indeed these developments were essential component for the work we did for the International Barley Genome Consortium leading to two high-quality publications:

Construction of a map-based reference genome sequence for barley, *Hordeum vulgare* L. Beier S, Himmelbach A, Colmsee C, Zhang XQ, Barrero RA, Zhang Q, Li L, Bayer M, Bolser D, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Staňková H, Vrána J, Chan S, Muñoz-Amatriain M, Ounit R, Wanamaker S, Schmutz T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Sampath D, **Heavens D**, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, McCooke JK, Tan C, Wang S, Yin S, Zhou



NIAB EMR, part of the NIAB group, is a charitable company limited by guarantee.  
Registered in England Registration No. 09894859. Charity Registration No. 1165055.  
Registered Office: NIAB EMR, Huntingdon Road, Cambridge, Cambridgeshire, United Kingdom, CB3 0LE, UK

G, Poland JA, Bellgard MI, Houben A, Doležel J, Ayling S, Lonardi S, Langridge P, Muehlbauer GJ, Kersey P, Clark MD, Caccamo M, Schulman AH, Platzer M, Close TJ, Hansson M, Zhang G, Braumann I, Li C, Waugh R, Scholz U, Stein N, Mascher M. Sci Data. 2017 Apr 27;4:170044. doi: 10.1038/sdata.2017.44. PMID: 28448065

A chromosome conformation capture ordered sequence of the barley genome. Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, Bayer M, Ramsay L, Liu H, Haberer G, Zhang XQ, Zhang Q, Barrero RA, Li L, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Staňková H, Vrána J, Chan S, Muñoz-Amatriain M, Ounit R, Wanamaker S, Bolser D, Colmsee C, Schmutzler T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Chailyan A, Sampath D, **Heavens D**, Clissold L, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, McCooke JK, Tan C, Wang P, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Borisjuk L, Houben A, Doležel J, Ayling S, Lonardi S, Kersey P, Langridge P, Muehlbauer GJ, Clark MD, Caccamo M, Schulman AH, Mayer KFX, Platzer M, Close TJ, Scholz U, Hansson M, Zhang G, Braumann I, Spannagl M, Li C, Waugh R, Stein N. Nature. 2017 Apr 26;544(7651):427-433. doi: 10.1038/nature22043. PMID: 28447635

From these achievements and the emphasis Darren places in delivery high-quality research I believe he is an exceptionally strong candidate for a PhD and therefore I am very happy to write this letter of support.

Yours faithfully



Professor Mario Caccamo  
Managing Director  
NIAB EMR



NIAB EMR, part of the NIAB group, is a charitable company limited by guarantee.  
Registered in England Registration No. 09894859. Charity Registration No. 1165055.  
Registered Office: NIAB EMR, Huntingdon Road, Cambridge, Cambridgeshire, United Kingdom, CB3 0LE, UK



Professor Michael W Bevan FRS  
Project Leader  
Cell and Developmental Biology Dept

March 9 2018

To whom it may concern:

**Darren Heavens**

I am writing to provide my strongest support for Darren Heavens' application for a PhD degree at UEA. I have worked with Darren for the past 10 years or so, first when he was at the John Innes Centre, and more recently at the Earlham Institute. Darren is an internationally recognised expert on DNA sequencing technology, and he has made key contributions to sequencing the wheat genome, one of the largest and most complex genomes attempted to date. The outcomes of the first stages of our joint work in wheat genome analyses are in the two papers described below:

1. Bernardo J Clavijo, Luca Venturini, Christian Schudoma, Gonzalo Garcia Accinelli, Gemy Kaithakottil, Jonathan Wright, Philippa Borrill, George Kettleborough, **Darren Heavens**, Helen Chapman, James Lipscombe, Tom Barker, Fu-Hao Lu, Neil McKenzie, Dina Raats, Ricardo H Ramirez-Gonzalez, Aurore Coince, Ned Peel, Lawrence Percival-Alwyn, Owen Duncan, Josua Trösch, Guotai Yu, Dan M Bolser, Guy Namaati, Arnaud Kerhorne, Manuel Spannagl, Heidrun Gundlach, Georg Haberer, Robert P Davey, Christine Fosker, Federica Di Palma, Andrew L Phillips, A Harvey Millar, Paul J Kersey, Cristobal Uauy, Ksenia V Krasileva, David Swarbreck, **Michael W Bevan**, Matthew D Clark. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Research* 2017 (56 citations to date)
2. Fu-Hao Lu, Neil McKenzie, George Kettleborough, **Darren Heavens**, Matthew D Clark, **Michael W Bevan**. Independent assessment and improvement of wheat genome assemblies using Fosill jumping libraries. *GigaScience* (under review) and *BioRxiv* (doi: <https://doi.org/10.1101/219352>).

A key technical challenge in wheat genomics was the generating whole genome assemblies of the most accurate sequence reads. This is essential because the wheat genome is composed of three independent genomes that are very closely related. Sequence reads needs to be very accurate in order that their genome of origin can be determined. Darren developed a PCR-free method for making representative long-read libraries. These were shown to be absolutely essential for making representative libraries that permitted accurate assemblies of each A, B and D genome. In this way, Darren directly enabled the Earlham Institute (EI), JIC and RRES to deliver a key BBSRC objective- the first complete, accurate and annotated assembly of a wheat genome. Achieving this important international objective ensured the success of a BBSRC-funded LOLA grant led by EI, and directly promoted the formation of an international collaboration in wheat comparative genomics with EI and leading international wheat centres. The major impacts include the sequencing and assembly of ten wheat genomes by EI and collaborators, placing EI as the leading wheat genomics laboratory in the world. Darren's expertise also contributed directly to the initiation of two other major collaborative projects in wheat genomics and epigenomics at JIC and EI. These wheat genomics projects have been essential foundations for BBSRC-funded Strategic Programmes at JIC, EI, and several other UK centres. For example, the Designing Future Wheat Programme is developing improved wheat lines with UK commercial breeders.

In summary Darren' long-term delivery of technological innovation has had and will continue to have major international impacts in food security

Yours faithfully

John Innes Centre is a company limited by guarantee.  
Registered in England No. 511709 Registered Charity No. 223852

*John Innes Centre is grant aided by the Biotechnology and Biological Sciences Research Council*

Director, Professor Dale Sanders FRS

Registered Office  
Norwich Research Park, Colney, Norwich NR4 7UH  
Tel: +44 (0)1603 450000 Fax: +44 (0)1603 450045

## **Appendix 2: Publications submitted**

RESEARCH ARTICLE

Open Access

# A mixed community of actinomycetes produce multiple antibiotics for the fungus farming ant *Acromyrmex octospinosus*

Jörg Barke<sup>1</sup>, Ryan F Seipke<sup>1†</sup>, Sabine Grüschorw<sup>2‡</sup>, Darren Heavens<sup>3</sup>, Nizar Drou<sup>3</sup>, Mervyn J Bibb<sup>4</sup>,  
Rebecca JM Goss<sup>2</sup>, Douglas W Yu<sup>1,5</sup>, Matthew I Hutchings<sup>1,6\*</sup>

## Abstract

**Background:** Attine ants live in an intensely studied tripartite mutualism with the fungus *Leucoagaricus gongylophorus*, which provides food to the ants, and with antibiotic-producing actinomycete bacteria. One hypothesis suggests that bacteria from the genus *Pseudonocardia* are the sole, co-evolved mutualists of attine ants and are transmitted vertically by the queens. A recent study identified a *Pseudonocardia*-produced antifungal, named dentigerumycin, associated with the lower attine *Apterostigma dentigerum* consistent with the idea that co-evolved *Pseudonocardia* make novel antibiotics. An alternative possibility is that attine ants sample actinomycete bacteria from the soil, selecting and maintaining those species that make useful antibiotics. Consistent with this idea, a *Streptomyces* species associated with the higher attine *Acromyrmex octospinosus* was recently shown to produce the well-known antifungal candidin. Candidin production is widespread in environmental isolates of *Streptomyces*, so this could either be an environmental contaminant or evidence of recruitment of useful actinomycetes from the environment. It should be noted that the two possibilities for actinomycete acquisition are not necessarily mutually exclusive.

**Results:** In order to test these possibilities we isolated bacteria from a geographically distinct population of *A. octospinosus* and identified a candidin-producing *Streptomyces* species, which suggests that they are common mutualists of attine ants, most probably recruited from the environment. We also identified a *Pseudonocardia* species in the same ant colony that produces an unusual polyene antifungal, providing evidence for co-evolution of *Pseudonocardia* with *A. octospinosus*.

**Conclusions:** Our results show that a combination of co-evolution and environmental sampling results in the diversity of actinomycete symbionts and antibiotics associated with attine ants.

## Background

Fungiculture in the insect world is practised by ants, termites, beetles and gall midges [1]. The best-characterized examples are the attine ants, which are endemic to South and Central America and to the southern USA. The ancestor of these ants evolved the ability to cultivate fungus as a food source around 50 million years ago, leading to the monophyletic tribe Attini, which number 12 genera with more than 230 species. The genera *Acromyrmex*

and *Atta* (40 species) evolved 8-12 million years ago and form a branch of the higher attines, also known as leaf-cutting ants, which are characterized by large colonies of up to several million individuals [2]. Like the other leaf-cutting ants, the well-studied species *Acromyrmex octospinosus* forms a mutualism with a single basidiomycete fungus (Agaricales: Leiotaceae: Leucocoprineae) *Leucoagaricus gongylophorus* in which they exchange food as well as protection and transport services [3].

The mutualistic fungal garden can be parasitized by a variety of other fungi [4] but the major pathogen of leaf-cutting ant fungal gardens is a necrotrophic fungus (Ascomycota: anamorphic Hypocreales) in the genus *Escovopsis* [5]. Around 25% of the gardens in Panamanian

\* Correspondence: m.hutchings@uea.ac.uk

† Contributed equally

<sup>1</sup>School of Biological Sciences, University of East Anglia, Norwich, Norwich Research Park, NR4 7TJ, UK

Full list of author information is available at the end of the article

ant colonies contain *Escovopsis* which feed on the fungal cultivar and can destroy fungal gardens, leading to the collapse of the colony [6].

There is evidence that the fungal cultivar produces antibiotics in order to defend itself [7-9] and the ant workers also defend their fungal gardens through a combination of grooming and weeding [8], production of their own antimicrobials through metapleural gland secretions [10] and the application of weedkillers. These weedkillers are natural product antimicrobials produced by symbiotic actinomycete bacteria [7,11-13]. A long-standing theory suggests that bacteria from the genus *Pseudonocardia* co-evolved with the ants and are transmitted vertically by the gynes (reproductive females) along with the fungal cultivar. However, more recently, evidence has emerged that suggests attine ants are also associated with bacteria from the actinomycete genera *Streptomyces* and *Amycolatopsis* and that antibiotic-producing actinomycetes can be horizontally acquired through male dispersal and sampling of actinomycetes from the soil [7,14].

The identities of the antifungals produced by attine ant-associated actinomycetes remain largely unknown. Only two compounds have been identified so far: a previously unknown antifungal named dentigerumycin that is produced by *Pseudonocardia* species isolated from the lower attines *Apterostigma dentigerum* and candididin, a well known antifungal that is produced by *Streptomyces* species isolated from the higher attine ants belonging to the genus *Acromyrmex* [12,13]. *Pseudonocardia* isolated from *A. octospinosus* also inhibit the growth of *Escovopsis* in bioassays, but the antifungal compounds have not been isolated or identified [12].

The aims of this work were to isolate and identify actinomycete bacteria from *A. octospinosus*, identify antifungal compounds produced by these bacteria and thereby gain insights into whether the actinomycetes (i) co-evolved with the ants, as suggested by unusual antifungal compounds produced by *Pseudonocardia* mutualists, or (ii) were acquired from the environment, as suggested by the presence of well known antifungals that are widely produced by environmental isolates. We isolated actinomycetes from three colonies of *A. octospinosus* that were collected in Trinidad, identified two *Pseudonocardia* and nine *Streptomyces* species and chose single antifungal producing *Pseudonocardia* and *Streptomyces* species isolated from the same ant colony for further analysis. The *Streptomyces* species was found to produce candididin and is closely related to the candididin-producing *Streptomyces* bacteria isolated from *A. octospinosus* in Panama [12], supporting the hypothesis that candididin-producing *Streptomyces* species are common mutualists of higher attines and are probably acquired via environmental sampling. The *Pseudonocardia* species produces an unusual

antifungal compound that is related to the clinically important polyene antifungal nystatin. The isolation of these species suggests that the diversity of actinomycetes associated with attine ants probably occurs through both co-evolution of *Pseudonocardia* with the ants and environmental sampling.

This work also takes the total number of known antifungals associated with attine ants to three, two of which are associated with *A. octospinosus*, and provides the first direct biochemical evidence that a diversity of actinomycete symbionts translates into a diversity of antifungal compounds in attine ant colonies.

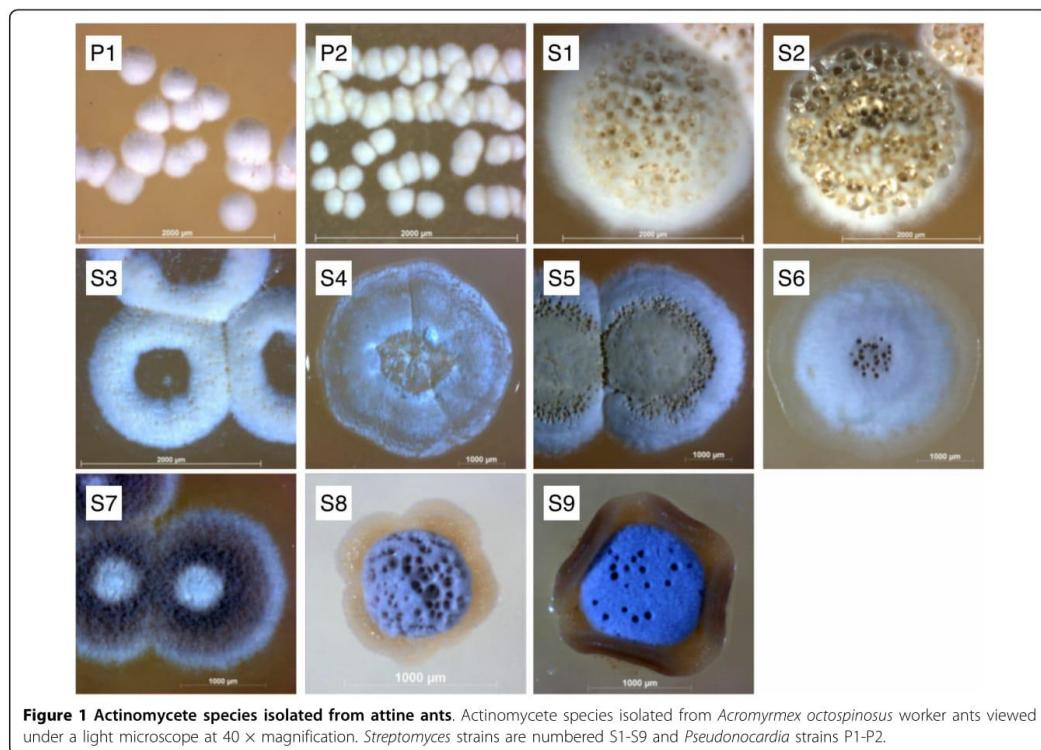
## Results

### Isolation and bioassay of actinomycetes

*A. octospinosus* ants from three colonies collected in Trinidad were either streaked directly onto HC and MS agar plates or washed in sterile water which was then spread onto the agar. Actinomycete colonies were purified by restreaking and then examined by light microscopy and identified by 16 S rDNA sequencing. Together with bacteria from other genera (*Tsukamurella* and *Nocardiopsis*) two *Pseudonocardia* (P1-P2) and nine *Streptomyces* (S1-S9) strains were isolated and identified (Figure 1, GenBank accession HM179225-HM179235). All bacterial strains were screened in bioassays against a strain of *Escovopsis weberi* isolated from an *A. octospinosus* nest and against *Candida albicans*, a human pathogen. Bioassays revealed that strains P1, S3, S4, S5 and S9 inhibit the growth of *E. weberi* when grown on MS agar (Figure 2) while P1, S3, S4 and S5 also inhibit the growth of *C. albicans* (Figure 3). The *Pseudonocardia* P1 strain has weak activity against *E. weberi* and very weak activity against *C. albicans* (Figures 2 and 3).

### *Streptomyces* S4 makes candididin

A previous study revealed that a *Streptomyces* strain isolated from *A. octospinosus* in Panama makes the polyene antifungal candididin [12] and a polymerase chain reaction (PCR) analysis of the nine *Streptomyces* and two *Pseudonocardia* strains using primers used by Haeder et al. in their study revealed that only *Streptomyces* S4 and S5 contain the candididin biosynthesis genes *fscM* and *fscP* (Additional Files 1 and 2). Candididin production was confirmed using liquid chromatography (LC) followed by tandem mass spectrometry (MS/MS) on butanol-extracted culture supernatants of *Streptomyces* S4 (Additional File 3). The *fscM* and *fscP* genes were not found in P1, S3, or S9, which suggests that they are producing antifungals not previously identified in the *A. octospinosus* mutualism. The PCR product amplified using *fscP* primers in the S9 sample was sequenced and is not *fscP*, consistent with its slightly larger size (Additional File 1).



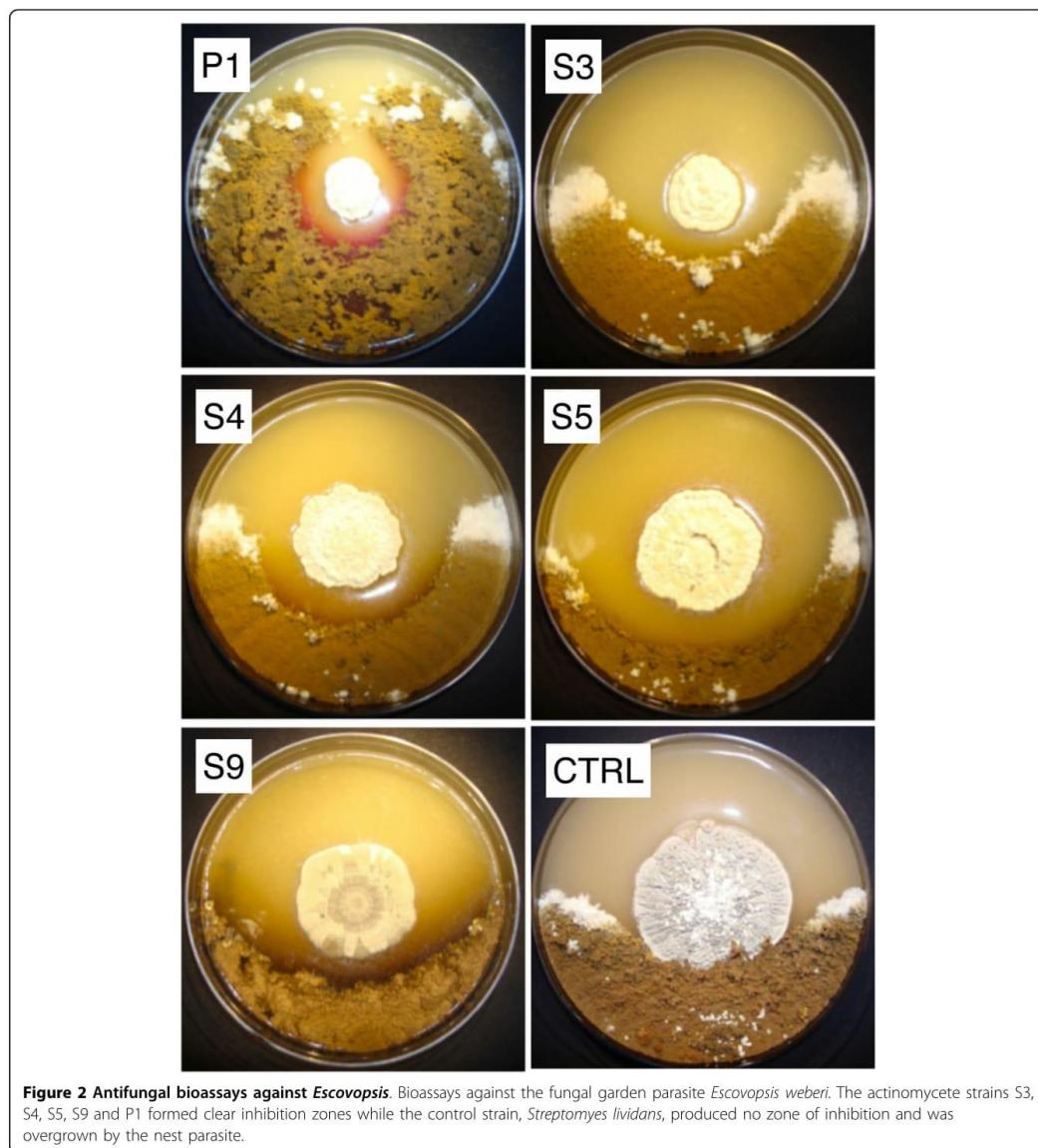
**Figure 1** Actinomycete species isolated from attine ants. Actinomycete species isolated from *Acromyrmex octospinosus* worker ants viewed under a light microscope at 40 × magnification. *Streptomyces* strains are numbered S1-S9 and *Pseudonocardia* strains P1-P2.

#### Genome scanning of *Pseudonocardia* P1

*Pseudonocardia* P1, isolated from the same ant colony as *Streptomyces* S4, produces a relatively small zone of inhibition in bioassays against *E. weberi* and a very small zone of inhibition against *C. albicans* (Figures 2 and 3). Furthermore, the antifungal activity of *Pseudonocardia* P1 was only detected on solid growth medium. This combination of factors made it difficult to purify sufficient antifungal compound(s) for analysis and identification. In order to gain further insight into the antifungal(s) produced by *Pseudonocardia* P1, we used 454-pyrosequencing to scan the genome of strain P1 (GenBank accession ADUJ00000000; Additional File 4). Analysis of the annotated contigs from this sequencing project revealed several polyketide synthase (PKS) gene fragments with > 90% amino acid identity to proteins involved in the biosynthesis of an antifungal compound named nystatin-like *Pseudonocardia* polyene (NPP) that is produced by *Pseudonocardia autotrophica* [15]. NPP is related to nystatin, a polyene antifungal that is made by *Streptomyces noursei* [16,17].

In order to determine whether or not *Pseudonocardia* P1 contains the entire biosynthetic gene cluster for a

nystatin-like compound, contigs were aligned against the characterized NPP biosynthetic gene cluster from *P. autotrophica* (see Methods and Additional File 5). The tiled contigs spanned the entire cluster, including the six PKS genes that assemble the nystatin aglycone, the non-sugar containing backbone of nystatin. Full-length coding sequences were captured for 11 genes (*nypF*, *nypH*, *nypDIII*, *nypL*, *nypN*, *nypDII*, *nypDII*, *nypE*, *nypO*, *nypRIV*, *nypM*) that are proposed to be primarily involved in the post PKS-modification of the nystatin aglycone and two new genes, *nypY* and *nypZ*, with unknown functions (Table 1) [16]. Interestingly, a second glycosyltransferase, absent in *S. noursei* and *P. autotrophica*, is present in the *nyp* gene cluster and we have named it *nypY* (Table 1). The NypY protein belongs to the same glycosyltransferase family as NypDI, however it displays only 42% amino acid identity to NypDI and is therefore unlikely to be a functionally redundant copy of NypDI. This genome analysis strongly suggested that *Pseudonocardia* P1 has the genetic capacity to produce a nystatin-like polyene antifungal. PCR screening of the *Pseudonocardia* P2 strain and the nine *Streptomyces* strains isolated in this study



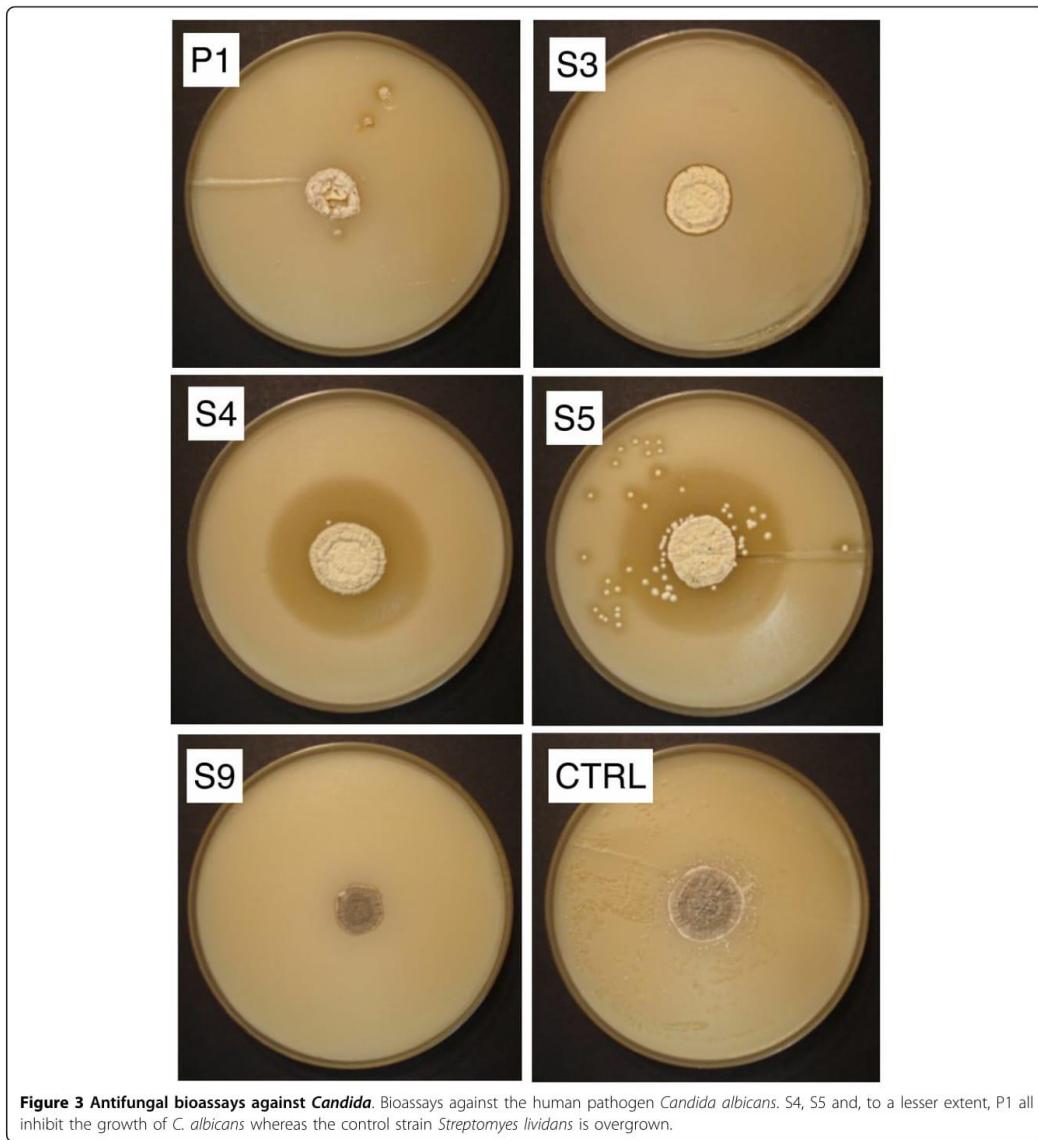
**Figure 2** Antifungal bioassays against *Escovopsis*. Bioassays against the fungal garden parasite *Escovopsis weberi*. The actinomycete strains S3, S4, S5, S9 and P1 formed clear inhibition zones while the control strain, *Streptomyces lividans*, produced no zone of inhibition and was overgrown by the nest parasite.

suggests that none of them contain biosynthetic genes for a nystatin-like antifungal (Additional File 2).

#### Identification of a nystatin-like compound in *Pseudonocardia* P1

In order to determine whether *Pseudonocardia* P1 produces a nystatin-like antifungal compound, extracts of *Pseudonocardia* P1 were analysed by LC-MS/MS and

compared to a nystatin A<sub>1</sub> standard (Figure 4). Molecular ions for nystatin A<sub>1</sub> (*m/z* 926.5) or for NPP (*m/z* 1129.6), produced by *P. autotrophica* [15] were not detected. However, a compound with a similar retention time on high-performance liquid chromatography (HPLC) to nystatin A<sub>1</sub> and with a molecular ion of *m/z* 1088.6 was identified (Figure 4a and b). This compound clearly, though somewhat concealed by the absorption



of co-eluting peaks, shows the characteristic polyene absorption bands in its ultraviolet spectrum (absorption maxima at 292, 305 and 320 nm, Figure 4e). Together with the presence of nystatin-like biosynthetic genes in *Pseudonocardia* P1, the LC-MS/MS results strongly suggested that the P1-derived extract contained a nystatin-like compound. We have tentatively named this compound nystatin P1.

The mass difference of 162 observed between nystatin P1 and nystatin A<sub>1</sub> suggested that nystatin P1 contains an additional hexose molecule. MS/MS fragmentation of the nystatin P1 ion (*m/z* 1088.6) resulted in a series of product ions that are very similar to those derived from nystatin A<sub>1</sub> (Figure 4c). All of the fragment ions corresponding to the nystatin P1 aglycone have corresponding counterparts in the nystatin A<sub>1</sub> standard (Figure 4d).

**Table 1 Nystatin P1 biosynthetic genes**

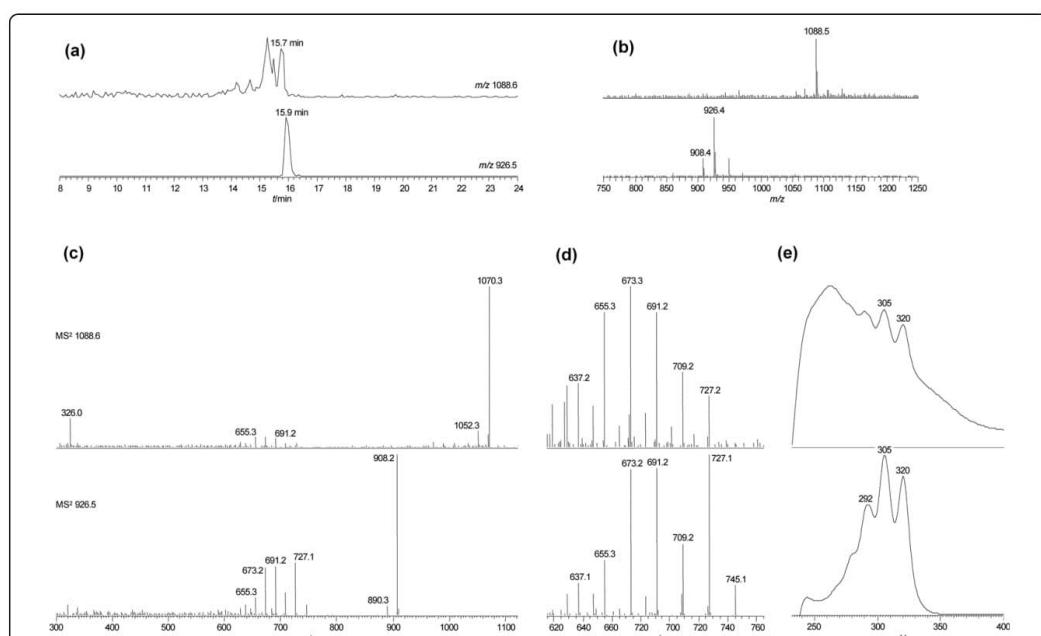
Contig ID	Pseudonocardia sp. P1 protein	Proposed function*	P. autotrophica ortholog	Identity (%)
PP100949	NypF	Phosphopantetheinyl transferase	NppF	89
PP100949	NypY	Glycosyltransferase	None†	–
PP100949	NypZ	Metallophosphoesterase	None‡	95
PP100398	NypH	ABC transporter	NppH	88
PP100398	NypDIII	dGDP-mannose-4,6-dehydratase	NppDIII	96
PP100400	NypL	P450 monooxygenase	NppL	84
PP100400	NypN	P450 monooxygenase	NppN	94
PP100400	NypDII	Aminotransferase	NppDII	96
PP100400	NypDI	Glycosyltransferase	NppDI	92
PP100821	NypE	Thioesterase	NppE	92
PP100306	NypO	Acyl-CoA decarboxylase	NppO	96
PP100306	NypRIV	LuxR transcriptional regulator	NppRIV	93
PP100306	NypM	Hypothetical protein	NppM§	82

\*Proposed function of full length nystatin P1 biosynthetic (nyp) genes present in the draft genome of *Pseudonocardia* sp. P1 (Genbank accession ADUJ000000000).

† NypY is a glycosyltransferase unique to the nystatin P1 biosynthetic gene cluster and is not orthologous to proteins in the nystatin-like *Pseudonocardia* polyene (NPP) biosynthetic gene cluster from *P. autotrophica* (AC = EU108007) or the nystatin biosynthetic gene cluster from *Streptomyces nouresi* (AC = AF263912).

‡ The nystatin P1 and NPP biosynthetic gene clusters contain a putative metallophosphoesterase downstream of *nypH* and *nppH*, respectively that is not present in the nystatin biosynthetic gene cluster from *S. nouresi*. This open reading frame was not originally annotated by Kim et al. [15] and we have therefore given the *Pseudonocardia* P1 ortholog the name of *nypZ*.

§ *nypM* encodes a hypothetical protein with high homology to *NppM*, which is annotated as a putative ferredoxin [15], however amino acid homology-based database searches failed to reveal homology to ferredoxin proteins.



**Figure 4 Identification of nystatin P1.** Identification of a nystatin-like compound by liquid chromatography-tandem mass spectrometry. For each panel the lower tier corresponds to the nystatin standard and the upper tier to the *Pseudonocardia* P1 extract. (a) Extracted ion chromatograms for  $m/z$  926.5 (nystatin standard) and  $m/z$  1088.6 (nystatin P1). (b) Mass spectra averaged across the retention times indicated in panel A. (c)  $MS^2$  analysis of the molecular ions identified above. The main mode of fragmentation is the loss of water molecules ( $m/z$  18). (d) Enlarged region of the  $MS^2$  spectra. These product ions arise from loss of the carbohydrate portion plus one to seven water molecules and most are found in the nystatin standard as well as in nystatin P1. (e) Ultraviolet spectra extracted at the retention times indicated in panel A.

These data strongly suggested that the aglycone (backbone) of nystatin A<sub>1</sub> and nystatin P1 is the same. Interestingly, the product ion with *m/z* 326 is consistent with a mycosamine-hexose disaccharide and was only observed for nystatin P1. Further fragmentation of the *m/z* 326 ion species corroborated the disaccharide nature of this moiety (Additional File 6).

The exact identity of the sugar molecules is, of course, speculative. Mycosamine is a probable component of nystatin P1 because this aminosugar is found in nystatin A<sub>1</sub> and all the necessary genes for its biosynthesis and attachment to the aglycone have been identified in *Pseudonocardia* P1 (Table 1). Glucose is frequently found as a substituent in bacterial natural products. However, other natural hexoses such as mannose or galactose are also good candidates for the second sugar substituent. The attachment of the hexose to give nystatin P1 is most likely to be executed by the glycosyltransferase NypY (see above). The presence of the disaccharide in MS/MS furthermore suggested that the nystatin P1 aglycone is substituted at one position with a mycosamine-hexose moiety rather than the two sugar molecules being attached at separate positions.

## Discussion

We isolated actinomycetes from *A. octospinosus* garden worker ants and, in a single colony of ants, identified a *Pseudonocardia* and a *Streptomyces* species that produce antifungals in laboratory culture. The *Streptomyces* species, which we named S4, contains candidin biosynthesis genes (Additional Files 1 and 2) and produces candidin (Additional File 3), consistent with a report on antifungal-producing actinomycetes associated with *A. octospinosus* [12]. The actinomycetes studied in this work were isolated from *A. octospinosus* ants collected in Trinidad, whereas the previous study used *A. octospinosus* ants collected in Panama [12]. However, despite this geographic separation, the candidin-producing *Streptomyces* strains identified in the two studies show 99% 16 S rDNA sequence identity suggesting that candidin-producing *Streptomyces* are common mutualists of *A. octospinosus*. Candidin-producing *Streptomyces* are widespread in the environment [18] and attine ants most likely acquire them selectively from the soil.

The *Pseudonocardia* species P1, isolated from the same colony as *Streptomyces* S4, showed relatively weak antifungal activity that was only observed in cultures grown on solid growth medium. This made it difficult to purify enough of the compound for analysis and identification. Using a genome scanning approach we identified a biosynthetic gene cluster for a polyene antifungal in *Pseudonocardia* P1 and then isolated and identified this antifungal using LC-MS/MS. This combined chemical and genomic approach provides a powerful tool for

identifying and isolating new antibiotics and confirmed that *Pseudonocardia* P1 produces a polyene antifungal that we have tentatively named nystatin P1. This compound is markedly different from the antifungal dentigerumycin produced by *Pseudonocardia* associated with the lower attine ant species *A. dentigerum* [13] although it is notable that both *Pseudonocardia* strains are making previously unknown antifungals, consistent with the idea that the *Pseudonocardia* mutualists co-evolved with attine ants. We did not detect any compounds in extracts from *Pseudonocardia* P1 agar plates and mycelium that matched the isotopic mass of dentigerumycin. However, since the biosynthetic gene cluster for this compound is not known, we cannot exclude the possibility that this strain also has the ability to make dentigerumycin.

Taken together, this work provides the first direct evidence that individual leaf-cutting ant colonies have access to multiple antifungals via the diversity of hosted actinomycetes and increases the number of known antifungals used by attine ants to three. This work also provides evidence to support the two current possibilities for the identity and acquisition of mutualistic bacteria, *Pseudonocardia* co-evolution, and the environmental acquisition of useful actinomycetes. This strongly suggests that both possibilities apply, at least in the attine species *A. octospinosus*. Careful experimental work will be needed in order to demonstrate that multiple compounds are in fact produced and confer benefits *in vivo* [19]. It is interesting that the only two antifungal compounds to be isolated and identified from *A. octospinosus* colonies so far are polyenes, which are active against dimorphic fungi, yeasts (*Candida*) and molds (*Escovopsis*), but which apparently do not kill the fungal cultivar [12]. The isolation of a nystatin-like polyene from a leaf-cutting ant-associated *Pseudonocardia* species in this work agrees with the report by Sen et al. [11] that some *Pseudonocardia* bacteria associated with attine ants have non-specific antibiotic properties that inhibit a range of fungi and are not targeted specifically at *Escovopsis* [11].

The advantage to the ants of deploying two antifungals is not clear. Polyene antifungals are thought to work by interacting hydrophobically with ergosterol in the fungal cell membrane and forming channels that increase membrane permeability [20], but this may not be their only mechanism of action [21], and there may therefore be some advantage to the ants in using more than one. However, as fungi do not develop resistance to polyene antifungals (at least in a clinical setting), it is unlikely that resistance is the basis for any such advantage. Nevertheless, as candidin and nystatin are not antibacterial, neither of these compounds is likely to be involved in competition amongst the bacteria for host resources. Thus, the identities of these two antifungal

compounds are consistent with the longstanding hypothesis that these actinomycete associates of leaf-cutting ants can be mutualists of the ant and the attine fungus, provided that the compounds are applied correctly by the ant [11].

### Conclusions

We used a combined genomic and chemical approach that has proven useful for the identification of a new antifungal associated with *Acromyrmex* ants, this time produced by their *Pseudonocardia* mutualist. This approach should stimulate further chemical ecology studies of insect fungiculture systems, which are widespread in nature and which are likely to use symbiotic antibiotic-producing bacteria to protect their fungal partners [1]. We also provide evidence that supports both of the possibilities proposed to explain the mutualism between actinomycetes and attine ants-co-evolution of *Pseudonocardia* with attine ants and environmental sampling by the ants of useful antibiotic-producing bacteria. We propose that these possibilities are not mutually exclusive and that both are likely to apply to both attine ants and other systems of insect fungiculture.

### Methods

#### Bacterial isolation and identification

Ants from three *A. octospinosus* colonies collected in Trinidad and Tobago were streaked onto hydrolysed chitin (HC) and mannitol plus soya flour (MS) agar plates [22,23] containing the antifungals nystatin and cycloheximide at final concentrations of 0.05 mg/mL. The remainder of the ants were washed in sterile water which was then spread onto HC and MS agar plates. Actinomycete isolates were colony purified and stored in 20% glycerol at -20°C. Genomic DNA was isolated from actinomycetes as described [23].

#### 16 S rDNA analysis

A 1000 bp fragment of the 16 S ribosomal DNA gene was PCR-amplified using the following primers: 533F 5'-GTGCCAGCMGCCGCGGTAA-3' [24] and 1492R 5'-GGTTACCTTGTACGACTT-3' [25]. The resulting PCR products were gel purified, sequenced (The Genome Analysis Centre, <http://www.tgac.bbsrc.ac.uk/>) and subsequently used to query the Green Genes database [http://greengenes.lbl.gov/cgi-bin/nph-simrank\\_interface.cgi](http://greengenes.lbl.gov/cgi-bin/nph-simrank_interface.cgi).

#### Bioassays against *Escovopsis* and *Candida*

Spores (50 µL) of each actinomycete were inoculated into 10 mL liquid TSB/YEME (1:1) [23] and grown on a shaker (260 rpm, 30°C) for three days in order to generate mycelium. The mycelium was collected by

centrifugation and resuspended in fresh TSB/YEME to yield a concentrated cell paste. The centre of an MS plate was inoculated with either 10 µL sterile TSB/YEME (negative control) or 10 µL of the concentrated cell paste and incubated for 10 days at 22°C, at which point the edge of the plate was inoculated with a small amount of mycelium of *Escovopsis weberi* (CBS 110660). The *Escovopsis* strain used in this study was obtained from CBS Fungal Biodiversity Centre <http://www.cbs.knaw.nl> and maintained on MS agar containing carbenicillin and streptomycin each at final concentrations of 0.05 mg/mL. Alternatively, *C. albicans* was inoculated into soft (0.5%) Lysogeny Broth agar, which was then used to overlay the plate containing the actinomycete.

#### 454-pyrosequencing and analysis

Genomic DNA was quantified using the Quant-it dsDNA HS Assay Kit (Invitrogen, CA, USA) and measured on a Qubit fluorometer (Invitrogen). An aliquot of 5 µg was used to generate the single stranded library for 454 pyrosequencing using the GS Titanium General Library Prep Kit according to the manufacturer's protocol (Roche, Hertfordshire, UK) except that, rather than fragmenting by nebulization, DNA was fragmented in a 100 µL volume using the Covaris-S2 ultra sonicator (K Biosciences, PA, USA) with the following settings-Mode: Frequency Sweep, Duty Cycle: 5%, Intensity: 3, Cycle Burst: 200 for two continuous cycles of 45 s. Library quality and quantity was assessed by running 1 µL of the library on a RNA PICO 6000 labchip (Agilent, CA, USA) and an emPCR titration was used to determine the optimal number of molecules per bead required to achieve the targeted 8% enrichment for the full scale emPCR. Approximately 790,000 enriched templated beads were subjected to 454 pyrosequencing on a quarter of a picotitre plate on the GS FLX sequencer (Roche) using the GS FLX Titanium Chemistry. The sequence reads were quality filtered and assembled into contigs using the Newbler Assembly v2 software (Roche).

Contigs were annotated using the Rapid Annotation Seed Technology Server [26]. Coding sequences annotated as polyketide synthases were extracted and inspected further by BlastP analysis against the National Center for Biotechnology Information non-redundant protein database, as well as Pfam [27] and non-ribosomal peptide synthetases-PKS [28]. NUCmer [29] using an 80% cutoff and the show-tiling utility were used to tile contigs to the *Pseudonocardia autotrophica* biosynthetic gene cluster for NPP [15]. Microsoft Excel was used to convert the output of the NUCmer show-tiling utility to Gene Finder Format and visualized using Artemis (release 11.22) [30].

#### LC-MS analysis

The residue obtained from butanol-extracted *Streptomyces* S4 cultures (50 mL) grown in liquid MS was redissolved in 50% aqueous methanol (0.3 mL). The samples were centrifuged at maximum speed prior to injection (5  $\mu$ L) into a Shimadzu single quadrupole LCMS-2010A mass spectrometer equipped with Prominence HPLC system. Compounds were separated on a Waters XBridge™ C18 3.5  $\mu$ m 2.1  $\times$  100 mm column using the following gradient (solvent A: 0.1 formic acid in water, solvent B: 0.1% formic acid in acetonitrile, flow rate 0.35 mL min $^{-1}$ ): 0.01-0.5 min 15% B, 0.5-14 min 15-95% B, 14-16 min 95% B, 16-16.5 min 95-15% B, 16.5-19 min 15% B. Mass spectra were acquired in positive ion mode with the capillary voltage set to 1.3 kV.

A sporulating culture of the *Pseudonocardia* P1 isolate on MS agar was extracted twice with methanol (200 mL). The solvent was removed under reduced pressure and the residue redissolved in 50% aqueous methanol (150  $\mu$ L). An authentic nystatin A<sub>1</sub> standard (Sigma-Aldrich, MO, USA) was prepared at 0.1 mg mL $^{-1}$  in 50% aqueous methanol. Immediately before LC-MS analysis, the crude extract and the standard were diluted twofold with 20% aqueous methanol and spun in a microcentrifuge at maximum speed for 4 min to remove any insoluble matter. Only the supernatant was used for injection (5  $\mu$ L). The samples were run on a Surveyor HPLC system attached to a LCQ DecaXP<sup>plus</sup> ion trap mass spectrometer (both Thermo Fisher, MA, USA). Separation was on a 100  $\times$  2 mm 3  $\mu$  Luna C18(2) column (Phenomenex) with 0.1% formic acid in water as solvent A and methanol as solvent B using the following gradient: 0-20 min 20-95% B, 20-22 min 95% B, 22-23 min 95-20% B, 23-30 min 20% B. The flow rate was set to 260  $\mu$ L min $^{-1}$  and the column temperature was maintained at 30°C. Detection was by ultraviolet (full spectra from 200-600 nm) and by positive electrospray MS using spray chamber conditions of 350°C capillary temperature, 50 units sheath gas, five units auxiliary gas, and 5.2 kV spray voltage. Targeted MS<sup>2</sup> with S4 and P1 extracts was performed with 35% collision energy and an isolation width of  $m/z$  4.0.

#### Additional material

**Additional file 1: Detecting candidin biosynthesis genes using polymerase chain reaction (PCR).** PCR analysis of antifungal producers using primers against candidin biosynthesis genes *fscM* and *fscP*. Sequence identities to Haeder et al. [12]: *fscM* gene, S4 = 100%, S5 = 99%; *fscP* gene: S4 = 98% and S5 = 98%

**Additional file 2: Streptomyces and Pseudonocardia strains identified in this study.** The *Pseudonocardia* and *Streptomyces* strains isolated in this study are listed with the *Acromyrmex octospinosus* colony they were isolated from (1,2 or 3), the accession numbers for their 16 S ribosomal DNA (rDNA) sequences, the top National Center for Biotechnology Information Blast hits for each of their 16 S rRNA sequences and the percentage identity to these BLAST hits. Also noted are the results from polymerase chain reaction testing for the candidin biosynthetic genes *fscM* and *fscP* using primers from a previous study [12] and the nystatin-like *Pseudonocardia* polyene biosynthetic gene *nppDIII* using the primer set RFS84 (CAGATCCGCTTCTACAGG) and RFS85 (CGCACCGAGTGCATCTG).

#### Additional file 3: Liquid chromatography-tandem mass spectrometry (LC-MS/MS) identification of candidin in S4 extracts.

Analysis of S4-derived extracts. Left panel (A), ultraviolet spectrum extracted at RT 8.3 min (see panel B) from the S4 extract. The absorption maxima match those previously reported for candidin D [12]. Right panel (B), LC-MS analysis of S4 extract. Ion chromatograms extracted for the molecular ion of candidin D ( $m/z$  1109.6) are shown. (C), MS<sup>2</sup> analysis of the extracted ion  $m/z$  1109.6. The fragmentation pattern of the antifungal compound from *Streptomyces* S4 perfectly matched the fragmentation of candidin as reported previously [12]. The ions highlighted in the Haeder et al. study [12] are labelled in a larger font.

#### Additional file 4: genome sequencing data for Pseudonocardia P1.

Summary of the *Pseudonocardia* sp. P1 draft genome sequence output obtained by 454 pyrosequencing

**Additional file 5: Identification of the nystatin P1 biosynthetic gene cluster.** Tiling of *Pseudonocardia* sp. P1 contigs (GenBank accession ADUJ00000000) to the NPP biosynthetic gene cluster from *P. autotrophica* (GenBank accession EU108007). \*The negative value for PP100949 denotes that the contig extends 4517 bp beyond the nystatin-like *Pseudonocardia* polyene biosynthetic gene cluster. \*\*Negative values indicate that adjacent contigs overlap.

**Additional file 6: MS<sup>3</sup> data for nystatin P1.** The spectrum shows the fragmentation data of the  $m/z$  1088  $\rightarrow$  326 ion. The most frequently observed fragmentation corresponds to loss of water:  $m/z$  308 (-1 H<sub>2</sub>O),  $m/z$  290 (-2 H<sub>2</sub>O),  $m/z$  272 (-1 H<sub>2</sub>O). The  $m/z$  146 product ion is consistent with a mycosamine sugar after loss of the hexose (mass difference 180).

#### Abbreviations

HC: hydrolyzed chitin; HPLC: high-performance liquid chromatography; MS: mannitol plus soya flour; MS/MS: tandem mass spectrometry; NPP: nystatin-like *Pseudonocardia* polyene; PCR: polymerase chain reaction; PKS: polyketide synthase.

#### Acknowledgements

This work was supported by a UEA-funded PhD studentship (JB) and an MRC Milstein award, G0801721 (MIH, RJMG and DY). MIH is a Research Councils UK Fellow. DY also received support from the Yunnan provincial government (2008A001) and the Chinese Academy of Sciences (0902281081). Genome sequencing was carried out at The Genome Analysis Centre under a Capacity and Capability Challenge project with MIH and MJB. We thank Paul Thomas for his assistance with light microscopy, Colin Kay and Lionel Hill for their assistance with LC-MS/MS, Govind Chandra for his advice on sequence analysis, Neil Gow for *Candida* strains and the Hutchings group members for their help with actinomycete culture and storage.

#### Author details

<sup>1</sup>School of Biological Sciences, University of East Anglia, Norwich, Norwich Research Park, NR4 7TJ, UK. <sup>2</sup>School of Chemistry, University of East Anglia,

Norwich, Norwich Research Park, NR4 7TJ, UK. <sup>3</sup>The Genome Analysis Centre, Norwich, Norwich Research Park, NR4 7UH, UK. <sup>4</sup>Department of Molecular Microbiology, John Innes Centre, Norwich, Norwich Research Park, NR4 7UH, UK. <sup>5</sup>State Key Laboratory of Genetic Resources, and Evolution, Ecology, Conservation and Environment Center (ECEC), Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China. <sup>6</sup>School of Medicine, Health Policy and Practice, University of East Anglia, Norwich, Norwich Research Park, NR4 7TJ, UK.

#### Authors' contributions

JB carried out the bacterial isolation and identification. RFS and SG contributed equally to this work. RFS carried out the genome sequence analysis. SG carried out the chemical isolations and identification. DH and ND sequenced and assembled the genome. MJB, DWY, RJMG and MIH conceived the study, participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Received: 29 June 2010 Accepted: 26 August 2010

Published: 26 August 2010

#### References

1. Kaltenpoth M: **Actinobacteria as mutualists: general healthcare for insects?** *Trends Microbiol* 2009, **17**:529-535.
2. Schultz TR, Brady SG: **Major evolutionary transitions in ant agriculture.** *Proc Natl Acad Sci USA* 2008, **105**:5435-5440.
3. Currie CR: **A community of ants, fungi, and bacteria: a multilateral approach to studying symbiosis.** *Annu Rev Microbiol* 2001, **55**:357-380.
4. Rodrigues A, Bacci M, Mueller UG, Ortiz A, Pagnocca FC: **Microfungal 'weeds' in the leafcutter ant symbiosis.** *Microb Ecol* 2008, **56**:604-614.
5. Reynolds HT, Currie CR: **Pathogenicity of *Escovopsis weberi*: The parasite of the attine ant-microbe symbiosis directly consumes the ant-cultivated fungus.** *Mycologia* 2004, **96**:955-959.
6. Gerardo NM, Mueller UG, Price SL, Currie CR: **Exploiting a mutualism: parasite specialization on cultivars within the fungus-growing ant symbiosis.** *Proc Biol Sci* 2004, **271**:1791-1798.
7. Currie CR, Scott JA, Summerbell RC, Malloch D: **Fungus-growing ants use antibiotic-producing bacteria to control garden parasites.** *Nature* 1999, **398**:701-705.
8. Little AEF, Murakami T, Mueller UG, Currie CR: **Defending against parasites: fungus-growing ants combine specialized behaviours and microbial symbionts to protect their fungus gardens.** *Biol Lett* 2006, **2**:12-16.
9. Wang Y, Mueller UG, Clardy J: **Antifungal diketopiperazines from symbiotic fungus of fungus growing ant *Cyphomyrmex minutus*.** *J Chem Ecol* 1999, **25**:935-941.
10. Bot ANM, Ortíz-Lechner D, Finster K, Maile R, Boomsma JJ: **Variable sensitivity of fungi and bacteria to compounds produced by the metapleural glands of leaf-cutting ants.** *Insectes Sociaux* 2002, **49**:363-370.
11. Sen R, Ishak HD, Estrada D, Dowd SE, Hong E, Mueller UG: **Generalized antifungal activity and 454-screening of Pseudonocardia and Amycolatopsis bacteria in nests of fungus-growing ants.** *Proc Natl Acad Sci USA* 2009, **106**:17805-17810.
12. Haeder S, Wirth R, Herz H, Spiteller D: **Candidicin-producing Streptomyces support leaf-cutting ants to protect their fungus garden against the pathogenic fungus Escovopsis.** *Proc Natl Acad Sci USA* 2009, **106**:4742-4746.
13. Oh DC, Poulsen M, Currie CR, Clardy J: **Dentigerumycin: a bacterial mediator of an ant-fungus symbiosis.** *Nat Chem Biol* 2009, **5**:391-393.
14. Mueller UG, Dash D, Rabeling C, Rodrigues A: **Coevolution between attine ants and actinomycete bacteria: a reevaluation.** *Evolution* 2008, **62**:2894-2912.
15. Kim B-G, Lee M-J, Seo J, Hwang Y-B, Lee M-Y, Han K, Sherman DH, Kim E-S: **Identification of functionally clustered nystatin-like biosynthetic genes in a rare actinomycetes, *Pseudonocardia autotrophica*.** *J Ind Microbiol Biotechnol* 2009, **36**:1425-1434.
16. Brautaset T, Sekurova ON, Sletta H, Ellingsen TE, Strøm AR, Valla S, Zotchev SB: **Biosynthesis of the polyene antifungal antibiotic nystatin in *Streptomyces noursei* ATCC 11455: analysis of the gene cluster and deduction of the biosynthetic pathway.** *Chem Biol* 2000, **7**:395-403.
17. Brown R, Hazen EL: **Present knowledge of nystatin, an antifungal antibiotic.** *Trans N Y Acad Sci* 1957, **19**:447-456.
18. Jørgensen H, Fjaervik E, Hakvåg S, Bruheim P, Bredholt H, Klinkenberg G, Ellingsen TE, Zotchev SB: **Candidicin biosynthesis gene cluster is widely distributed among Streptomyces spp. isolated from the sediments and the neuston layer of the Trondheim fjord, Norway.** *Appl Environmental Microbiol* 2009, **75**:3296-3303.
19. Kroiss J, Kaltenpoth M, Schneider B, Schwinger M-G, Hertweck C, Maddula RK, Strohm E, Svatopluk A: **Symbiotic streptomycetes provide antibiotic combination prophylaxis for wasp offspring.** *Nat Chem Biol* 2010, **6**:261-263.
20. de Kruifji B, Demel RA: **Polyene antibiotic-sterol interactions in membranes of *Acholeplasma laidlawii* cells and lecithin liposomes. 3. Molecular structure of the polyene antibiotic-cholesterol complexes.** *Biochim Biophys Acta* 1974, **339**:57-70.
21. Van Leeuwen MR, Golovina EA, Dijksterhuis J: **The polyene antimycotics nystatin and filipin disrupt the plasma membrane, whereas natamycin inhibits endocytosis in germinating conidia of *Penicillium discolor*.** *J Appl Microbiol* 2009, **106**:1908-1918.
22. Hsu SC, Lockwood JL: **Powdered chitin agar as a selective medium for enumeration of actinomycetes in water and soil.** *Appl Microbiol* 1975, **29**:422-426.
23. Kieser T, Bibb MJ, Buttner MJ, Chater KF, Hopwood DA: **Practical Streptomyces Genetics.** Norwich: John Innes Foundation 2000.
24. Hugenholtz P, Goebel BM, Pace NR: **Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity.** *J Bacteriol* 1998, **180**:4765-4774.
25. Lane J: **16S/23 S rRNA sequencing.** *Nucleic Acid Techniques In Bacterial Systematics* NY: John Wiley and Sons 1991, 115-175.
26. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, et al: **The RAST Server: rapid annotations using subsystems technology.** *BMC Genomics* 2008, **9**:75.
27. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, et al: **The Pfam protein families database.** *Nucleic Acids Res* 2010, **38**:D211-D222.
28. Jenke-Kodama H, Dittmann E: **Bioinformatic perspectives on NRPS/PKS megasynthases: advances and challenges.** *Nat Prod Rep* 2009, **26**:874-883.
29. Kurtz B, Phillipi A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5**:R12.
30. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16**:944-945.

doi:10.1186/1741-7007-8-109

**Cite this article as:** Barke et al.: A mixed community of actinomycetes produce multiple antibiotics for the fungus farming ant *Acromyrmex octospinosus*. *BMC Biology* 2010 **8**:109.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
www.biomedcentral.com/submit



## GENOME ANNOUNCEMENTS

### Complete Genome Sequence of the Proteolytic *Clostridium botulinum* Type A5 (B3') Strain H04402 065<sup>†</sup>

Andrew T. Carter,<sup>1\*</sup> Bruce M. Pearson,<sup>1</sup> Lisa C. Crossman,<sup>2</sup> Nizar Drou,<sup>2</sup> Darren Heavens,<sup>2</sup> David Baker,<sup>2</sup> Melanie Febrer,<sup>2</sup> Mario Caccamo,<sup>2</sup> Kathie A. Grant,<sup>3</sup> and Michael W. Peck<sup>1</sup>

*Institute of Food Research, Norwich, United Kingdom<sup>1</sup>; The Genome Analysis Centre, Norwich, United Kingdom<sup>2</sup>; and HPA Microbiological Services Division, London, United Kingdom<sup>3</sup>*

Received 14 January 2011/Accepted 11 February 2011

**H04402 065 is one of a very small group of strains of proteolytic *Clostridium botulinum* that form type A5 neurotoxin. Here, we report the complete 3.9-Mb genome sequence and annotation of strain H04402 065, which was isolated from a botulism patient in the United Kingdom in 2004.**

Proteolytic *Clostridium botulinum* neurotoxin causes food-borne, infant, and wound botulism. Three types of neurotoxin (A, B, and F) are formed by this organism, but type A presents the greatest bioterrorism threat (10). Five subtypes are known (A1 to A5), but to date, only genomes of strains forming subtypes A1 to A4 are published. Strains of proteolytic *C. botulinum* isolated from 4 of 40 wound botulism cases from the United Kingdom in 2004 (1) were closely related by whole-genome analysis, and each carried an identical subtype A5 gene with a standard *ha* neurotoxin gene cluster, plus a truncated type B3 neurotoxin gene (4, 5). The same arrangement was seen in a strain of proteolytic *C. botulinum* from a Californian infant botulism case (7). Here, we report the fully assembled complete genome sequence and annotation of *C. botulinum* type A5 (B3') strain H04402 065.

Genomic DNA was sequenced using Roche 454 and Illumina GA2 platforms. The former generated a 61.85-Mb sequence (16× coverage) with contigs assembled using Newbler. Illumina sequencing generated 4,709 Mb (1,068× coverage) with paired-end lane-generated contigs assembled with ABySS (13). Close proteolytic *C. botulinum* genome synteny (11) enabled contig mapping to strains Kyoto (NC\_012563) and Langeland (NC\_009699). Misassemblies were recognized by dot matrix comparison (DNAMAN version 5.1.5; Lynnon Corporation). Gaps were closed by sequencing PCR products from the same DNA. Roche 454 reads, sensitive to homopolymeric tracts, introduced nearly 1,500 sequence ambiguities. These were corrected by comparison with Illumina contigs, which are unaffected by these tracts. Protein-coding regions were predicted using Glimmer (6) and GeneMark (3) with manual curation using Artemis (12). Automatic annotation using RAST, preserving gene calls (2), was complemented with manual annotation of interesting regions highlighted by compari-

sons with other *C. botulinum* genomes by using Artemis, including InterPro domains, TMHMM, and SigP analyses.

*C. botulinum* strain H04402 065 has a circular chromosomal genome of 3,919,740 bp with a 28.2% G+C content and no plasmids. Totals of 3,719 coding sequences, 72 tRNA genes, and 9 complete rRNA loci were identified. The coding density was 0.94 genes/kb, with an average gene length of 854 bp. Double reciprocal orthologue plots identified strains Kyoto (type A2) and CDC657 (type Ba4) as close relatives, but H04402 065 shows synteny with all proteolytic *C. botulinum* genomes.

Chromosomal neurotoxin gene clusters are found at one of three sites (9), and that of strain H04402 065 resides in the *oppA-brnQ* operon, as with some other type A and B strains (9). No other neurotoxin cluster genes were found. Strain H04402 065 contains only three complete spore germinant receptor operons, as with other type A strains (11). Comparative genomics of proteolytic *C. botulinum* showed that the flagellar glycosylation island (FGI) is a genetically heterogeneous part of the genome (5). Interestingly, that of strain H04402 065 includes five genes with 90% identity to *Clostridium tetani* aminophosphonate metabolic pathway genes CTC1698, CTC1699, CTC1700, CTC1704, and CTC1705 (8).

**Nucleotide sequence accession number.** The complete genome sequence of strain H04402 065 has been deposited in EMBL/GenBank under accession number FR773526.

This work was supported by the Institute Strategic Programme Grant of the BBSRC.

#### REFERENCES

1. Akbulut, D., et al. 2005. Wound botulism in injectors of drugs: upsurge in cases in England during 2004. Euro Surveill. **10**:172–174.
2. Aziz, R. K., et al. 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9**:75.
3. Besemer, J., A. Lomsadze, and M. Borodovsky. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. **29**:2607–2618.
4. Carter, A. T., et al. 2010. Further characterization of proteolytic *Clostridium botulinum* type A5 reveals that neurotoxin formation is unaffected by loss of the *cmtR* (*boR*) promoter sigma factor binding site. J. Clin. Microbiol. **48**: 1012–1013.

\* Corresponding author. Mailing address: Institute of Food Research, Norwich, United Kingdom. Phone: 44-1603-255398. Fax: 44-1603-507723. E-mail: andrew.carter@bbsrc.ac.uk.

† Published ahead of print on 4 March 2011.

5. Carter, A. T., et al. 2009. Independent evolution of neurotoxin and flagellar genetic loci in proteolytic *Clostridium botulinum*. *BMC Genomics* **10**:115.
6. Delcher, A. L., K. A. Bratke, E. C. Powers, and S. L. Salzberg. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**:673–679.
7. Dover, N., J. R. Barash, and S. S. Arnon. 2009. Novel *Clostridium botulinum* toxin gene arrangement with subtype A5 and partial subtype B3 botulinum neurotoxin genes. *J. Clin. Microbiol.* **47**:2349–2350.
8. Fox, E. A., and G. L. Mendz. 2006. Phosphonate degradation in microorganisms. *Enzyme Microb. Technol.* **40**:145–150.
9. Hill, K. K., et al. 2009. Recombination and insertion events involving the botulinum neurotoxin complex genes in *Clostridium botulinum* types A, B, E and F and *Clostridium butyricum* type E strains. *BMC Biol.* **7**:66.
10. Peck, M. W. 2009. Biology and genomic analysis of *Clostridium botulinum*. *Adv. Microb. Physiol.* **55**:183–320.
11. Peck, M. W., S. C. Stringer, and A. T. Carter. 2011. *Clostridium botulinum* in the post-genomic era. *Food Microbiol.* **28**:183–191.
12. Rutherford, K., et al. 2000. Artemis: sequence visualization and annotation. *Bioinformatics* **16**:944–945.
13. Simpson, J. T., et al. 2009. ABYSS: a parallel assembler for short read sequence data. *Genome Res.* **19**:1117–1123.

## Draft Genome Sequence of *Streptomyces* Strain S4, a Symbiont of the Leaf-Cutting Ant *Acromyrmex octospinosus*<sup>†</sup>

Ryan F. Seipke,<sup>1</sup> Lisa Crossman,<sup>2</sup> Nizar Drou,<sup>2</sup> Darren Heavens,<sup>2</sup> Mervyn J. Bibb,<sup>3</sup> Mario Caccamo,<sup>2</sup> and Matthew I. Hutchings<sup>1\*</sup>

School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, United Kingdom<sup>1</sup>; The Genome Analysis Centre, Norwich Research Park, Norwich NR4 7UH, United Kingdom<sup>2</sup>; and Department of Molecular Microbiology, John Innes Centre, Norwich Research Park, Norwich NR4 7UH, United Kingdom<sup>3</sup>

Received 11 May 2011/Accepted 6 June 2011

*Streptomyces* spp. are common symbionts of the leaf-cutting ant species *Acromyrmex octospinosus*, which feeds on basidiomycete fungus leaf matter and harvests the lipid- and carbohydrate-rich gongylidia as a food source. *A. octospinosus* and other ant genera use antifungal compounds produced by *Streptomyces* spp. and other actinomycetes in order to help defend their fungal gardens from parasitic fungi. Herein, we report the draft genome sequence of *Streptomyces* strain S4, an antifungal-producing symbiont of *A. octospinosus*.

The well-studied leaf-cutting ant species *Acromyrmex octospinosus* forms a mutualism with a single basidiomycete fungus, *Leucoagaricus gongylophorus*, in which the ants exchange food as well as protection and transport services with the fungus (4). The fungal garden can be parasitized by a variety of microorganisms (6, 10, 11). The ants groom and weed their garden to remove parasites and produce antifungal secretions from their metaplural glands (8). The ants also host *Amycolatopsis*, *Pseudonocardia*, and *Streptomyces* exosymbionts (5, 7, 9, 14). These symbionts produce antifungal compounds that are thought to be applied as weed killers by the ants (3, 5, 7, 9, 13, 14). Both the chemical identities of these antifungal compounds and the means by which symbionts are selected have been the subject of several recent studies (3, 9, 14). One of these studies demonstrated that genome sequencing of ant symbionts can aid the identification of antifungal compounds that may be important in this mutualism and could also help us understand how the selection of leaf-cutting ant symbionts occurs (2, 3).

A combination of shotgun, 3-kbp and 8-kbp paired-end libraries were constructed to sequence the *Streptomyces* strain S4 genome on the GS FLX sequencer (Roche) using the GS FLX Titanium series chemistry kit, generating >335 Mbp of sequence. Reads were assembled using the gsAssembly version 2.3 software (Roche), generating 12 scaffolds containing 211 large contigs (>500 bp) spanning 7.47 Mbp of sequence, which is within the size range reported for other streptomycetes. The genome was shown to consist of one linear chromosome; one linear plasmid, pS4L1; and one circular plasmid, pS4C1. These were annotated using the Rapid Annotation Subsystem Technology (RAST) server, and the predicted open reading frames were manually inspected and the annotation was adjusted using Artemis, release 12 (1, 12).

The *Streptomyces* S4 genome, as with other streptomycetes,

contains multiple biosynthetic gene clusters coding for known and predicted bioactive secondary metabolites. Notably, *Streptomyces* S4 contains a biosynthetic gene cluster that directs the biosynthesis of the antifungal candidin, which was proposed to be an antifungal used by *A. octospinosus* to protect the fungal garden and was previously demonstrated to be produced by *Streptomyces* strain S4 (3, 7). The genome of *Streptomyces* strain S4 is also predicted to make mannopeptimycin-like and gramicidin-like antibacterial compounds as well as biosynthetic gene clusters predicted to encode anticancer compounds similar to fredericamycin and kendomycin as well as four cryptic biosynthetic gene clusters whose products are unknown. The presence of multiple biosynthetic gene clusters makes *Streptomyces* strain S4 an attractive symbiont and could possibly explain the isolation of a taxonomically very similar strain that produces candidin from *A. octospinosus* in Panama (7). The biosynthesis of predicted antibacterials also has implications for the bacterial community present on the ant cuticle, suggesting that leaf-cutting ant symbionts may be involved in determining which bacterial species compose the ant microbiome (2).

**Nucleotide sequence accession number.** The genome sequence has been deposited in DDBJ/EMBL/GenBank under the accession number CADY00000000. The version described in this paper is the first version, CADY01000000.

We thank Govind Chandra for advice concerning bioinformatics analysis and genome sequencing as well as members of the Hutchings laboratory for their support.

The sequencing was supported by The Genome Analysis Centre Capacity and Capability Challenge Programme (project number CCC-1-12). R.F.S. was supported through an MRC Milstein award, G0801721, awarded to M.I.H.

### REFERENCES

1. Aziz, R. K., et al. 2008. The RAST server: rapid annotations using subsystems technology. *BMC Genomics* 9:75.
2. Barke, J., R. F. Seipke, D. W. Yu, and M. I. Hutchings. 2011. A mutualistic microbiome: how do fungus-growing ants select their antibiotic-producing bacteria. *Commun. Integr. Biol.* 4:41–43.
3. Barke, J., et al. 2010. A mixed community of actinomycetes produce multiple antibiotics for the fungus farming ant *Acromyrmex octospinosus*. *BMC Biol.* 8:109.

\* Corresponding author. Mailing address: School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, United Kingdom. Phone: 44 01603 592257. Fax: 44 01603 592250. E-mail: m.hutchings@uea.ac.uk.

† Published ahead of print on 17 June 2011.

4. Currie, C. R. 2001. A community of ants, fungi, and bacteria: a multilateral approach to studying symbiosis. *Annu. Rev. Microbiol.* **55**:357–380.
5. Currie, C. R., J. A. Scott, R. C. Summerbell, and D. Malloch. 1999. Fungus-growing ants use antibiotic-producing bacteria to control garden parasites. *Nature* **398**:701–705.
6. Gerardo, N. M., U. G. Mueller, S. L. Price, and C. R. Currie. 2004. Exploiting a mutualism: parasite specialization on cultivars within the fungus-growing ant symbiosis. *Proc. Royal Soc. Biol. Sci.* **271**:1791–1798.
7. Haeder, S., R. Wirth, H. Herz, and D. Spitteler. 2009. Candinicin-producing *Streptomyces* support leaf-cutting ants to protect their fungus garden against the pathogenic fungus *Escovopsis*. *Proc. Natl. Acad. Sci. U. S. A.* **106**:4742–4746.
8. Little, A. E. F., T. Murakami, U. G. Mueller, and C. R. Currie. 2006. Defending against parasites: fungus-growing ants combine specialized behaviours and microbial symbionts to protect their fungus gardens. *Biol. Lett.* **2**:12–16.
9. Oh, D. C., M. Poulsen, C. R. Currie, and J. Clardy. 2009. Dentigerumycin: a bacterial mediator of an ant-fungus symbiosis. *Nat. Chem. Biol.* **5**:391–393.
10. Reynolds, H. T., and C. R. Currie. 2004. Pathogenicity of *Escovopsis weberi*: the parasite of the attine ant-microbe symbiosis directly consumes the ant-cultivated fungus. *Mycologia* **96**:955–959.
11. Rodrigues, A., M. Bacci, U. G. Mueller, A. Ortiz, and F. C. Pagnocca. 2008. Microfungal ‘weeds’ in the leafcutter ant symbiosis. *Microb. Ecol.* **56**:604–614.
12. Rutherford, K., et al. 2000. Artemis: sequence visualization and annotation. *Bioinformatics* **2000** **16**:944–945.
13. Schoenian, I., et al. 2011. Chemical basis of the synergism and antagonism in microbial communities in the nests of leaf-cutting ants. *Proc. Natl. Acad. U. S. A.* **108**:1955–1960.
14. Sen, R., et al. 2009. Generalized antifungal activity and 454-screening of *Pseudonocardia* and *Amycolatopsis* bacteria in nests of fungus-growing ants. *Proc. Natl. Acad. Sci.* **106**:17805–17810.

## GENOME ANNOUNCEMENTS

### Genome Sequence of the Vertebrate Gut Symbiont *Lactobacillus reuteri* ATCC 53608<sup>†</sup>

Darren Heavens,<sup>1,†</sup> Louise E. Tailford,<sup>2,†</sup> Lisa Crossman,<sup>1,†</sup> Faye Jeffers,<sup>2</sup>  
Donald A. MacKenzie,<sup>2</sup> Mario Caccamo,<sup>1</sup> and Nathalie Juge<sup>2,\*</sup>

The Genome Analysis Centre, Norwich Research Park, Norwich NR4 7UH, United Kingdom,<sup>1</sup> and Institute of Food Research, Norwich Research Park, Norwich NR4 7UA, United Kingdom<sup>2</sup>

Received 12 May 2011/Accepted 18 May 2011

*Lactobacillus reuteri*, inhabiting the gastrointestinal tracts of a range of vertebrates, is a true symbiont with effects established as beneficial to the host. Here we describe the draft genome of *L. reuteri* ATCC 53608, isolated from a pig. The genome sequence provides important insights into the evolutionary changes underlying host specialization.

The Gram-positive bacterium *Lactobacillus reuteri* is an excellent model organism to study the evolutionary strategy of a vertebrate gut symbiont, as this species inhabits the gastrointestinal tracts of mammals as diverse as humans, pigs, mice, and rats, as well as different species of birds (11). Population genetics, using amplified fragment length polymorphism and multilocus sequence analysis; genomic; and experimental approaches using lactobacillus-free and germfree mouse models revealed that host-specific subpopulations exist among members of the species *L. reuteri* (9). Furthermore, several trials have shown that *L. reuteri* confers health benefits on humans and animals, and strains of this species have been shown to modulate the host immune system (11). Efforts to understand the mechanism by which *L. reuteri* strains have remained restricted to particular hosts are ongoing (5).

To further investigate the genomic basis for host adaptation of *L. reuteri* to the gut, we have determined the genome sequence of the pig isolate *L. reuteri* ATCC 53608 (8). Genomic DNA was isolated using a modified form of the method of Oh and colleagues (9) and used to generate in excess of 365 Mbp of sequence from a combination of shotgun and 3-kbp paired-end libraries (220 Mbp and 145 Mbp, respectively) on the 454 GS FLX sequencer (Roche) using the Titanium Chemistry. Reads passing the default filter settings were assembled using gsAssembly V2.3 software (Roche) and generated 13 scaffolds containing 99 large contigs (>500 bp) and spanning 1.96 Mbp of sequence. The genome of *L. reuteri* ATCC 53608 is 1,969,869 bp in length and has an average G+C content of 38.4%. Automatic gene prediction was performed using Glimmer3 and GeneMark software (2, 3). Annotation was transferred from the related strain *L. reuteri* JCM 1112<sup>T</sup>. Unique

regions were manually annotated using Artemis (10), augmented with InterPro (6), TMHMM (transmembrane prediction using hidden Markov models) (7), and SignalP domains (4). A total of 2,024 protein-coding sequences were predicted, with a coding percentage of 88.7%. The coding density was 1.03 genes per kb, with an average gene length of 863 bp. The genome contains six predicted copies of the rRNA genes. Comparative genomics of ATCC 53608 with genome sequence available for the *L. reuteri* 100-23 and DSM 20016<sup>T</sup>/JCM 1112<sup>T</sup> strains isolated from rats and humans (5), respectively, revealed approximately 500 ATCC 53608-specific genes, whereas 1,335 genes are present in all four strains. Genome analysis also revealed the presence of a putative prophage or plasmid of 137,391 bp with flanking resolvase/integrase and transposase genes. ATCC 53608 lacks the 10.2-kb native plasmid pLUL631 described in original isolate 1063 (1) but harbors one small plasmid of 9,003 bp. Detailed analysis of the assembled ATCC 53608 genome will help to predict the competitiveness of *L. reuteri* strains *in vivo* and to provide a context for the rational selection of probiotic strains.

**Nucleotide sequence accession numbers.** This genome sequencing project has been deposited at DDBJ/EMBL/GenBank under accession number CACS00000000. The version described in this paper is CACS02000000. The 138 contigs contained in the genome have been deposited under accession numbers CACS02000001 to CACS02000138. The 13 fully annotated scaffolds built from the contigs have been deposited under accession numbers FR854361 to FR854373.

This work was supported by the Biotechnology and Biological Sciences Research Council.

We thank Robert Davey (The Genome Analysis Centre) for his help with the submission of the genome sequence.

#### REFERENCES

1. Ahrné, S., G. Molin, and L. Axellson. 1992. Transformation of *Lactobacillus reuteri* with electroporation: studies on the erythromycin resistance plasmid pLUL631. *Curr. Microbiol.* 24:199–205.
2. Besemer, J., A. Lomsadze, and M. Borodovsky. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Impli-

\* Corresponding author. Mailing address: Institute of Food Research, Norwich Research Park, NR4 7UA Norwich, United Kingdom. Phone: 44 (0)1603255068. Fax: 44 (0)1603507723. E-mail: nathalie.juge@bbsrc.ac.uk.

† The first three authors contributed equally to this work.

Published ahead of print on 27 May 2011.

cations for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **29**:2607–2618.

3. **Delcher, A. L., K. A. Bratke, E. C. Powers, and S. L. Salzberg.** 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**:673–679.
4. **Emanuelsson, O., S. Brunak, G. von Heijne, and H. Nielsen.** 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* **2**:953–971.
5. **Frese, S. A., et al.** 2011. The evolution of host specialization in the vertebrate gut symbiont *Lactobacillus reuteri*. *PLoS Genet.* **7**:e1001314.
6. **Hunter, S., et al.** 2009. InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**:D211–D215.
7. **Krogh, A., B. Larsson, G. von Heijne, and E. L. L. Sonnhammer.** 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**:567–580.
8. **MacKenzie, D. A., et al.** 2010. Strain-specific diversity of mucus-binding proteins in the adhesion and aggregation properties of *Lactobacillus reuteri*. *Microbiology* **156**:3368–3378.
9. **Oh, P. L., et al.** 2010. Diversification of the gut symbiont *Lactobacillus reuteri* as a result of host-driven evolution. *ISME J.* **4**:377–387.
10. **Rutherford, K., et al.** 2000. Artemis: sequence visualization and annotation. *Bioinformatics* **16**:944–945.
11. **Walter, J., R. A. Britton, and S. Roos.** 2011. Microbes and Health Sackler Colloquium: host-microbial symbiosis in the vertebrate gastrointestinal tract and the *Lactobacillus reuteri* paradigm. *Proc. Natl. Acad. Sci. U. S. A.* **108**:4645–5462.

# Benchmarks

## A method to simultaneously construct up to 12 differently sized Illumina Nextera long mate pair libraries with reduced DNA input, time, and cost

Darren Heavens, Gonzalo Garcia Accinelli, Bernardo Clavijo, and Matthew Derek Clark

*The Genome Analysis Centre (TGAC), Norwich Research Park, Norwich, UK*

*BioTechniques* 59:42-45 (July 2015) doi 10.2144/000114310

Keywords: long mate pair (LMP) construction; insert sizes; genome assembly; wheat

Supplementary material for this article is available at [www.BioTechniques.com/article/114310](http://www.BioTechniques.com/article/114310).

Long mate pair (LMP) or “jump” libraries are invaluable for producing contiguous genome assemblies and assessing structural variation. However the consistent production of high quality (low duplication rate, accurately sized) LMP libraries has proven problematic in many genome projects. Input DNA length and quantity are key issues that can affect success. Here we demonstrate how 12 libraries covering a wide range of jump sizes can be constructed from <10 µg of DNA, thus ensuring production of the best LMP libraries from a given DNA sample. Finally, we demonstrate the accuracy of the insert sizes by mapping reads from each library back to an existing assembly.

Standard paired-end next-generation sequencing projects can produce long continuous sections of sequence (contigs), but these alone lack the long-range information required to produce single contig assemblies of even bacterial chromosomes (1). Assemblies based on paired-end data alone are unable to resolve repeated sequences that are bigger than the insert size of the library (typically ~500 bp). The genomes of some higher eukaryotes can consist of >80% repeated sequences (2), and this can result in highly fragmented genome assemblies containing many

thousands or even millions of small contigs.

In order to increase assembly contiguity, many projects use long mate pair (LMP) libraries to jump over repeated sequences to connect contigs, a process known as scaffolding (3). Depending on the quantity and quality of the available input DNA it is possible to generate LMP libraries with insert sizes ranging from 1.5 kb to 40 kb. High quality assemblies typically use multiple LMP libraries of different insert sizes, which is costly in terms of input DNA quantity, time, and money. LMP libraries are also notori-

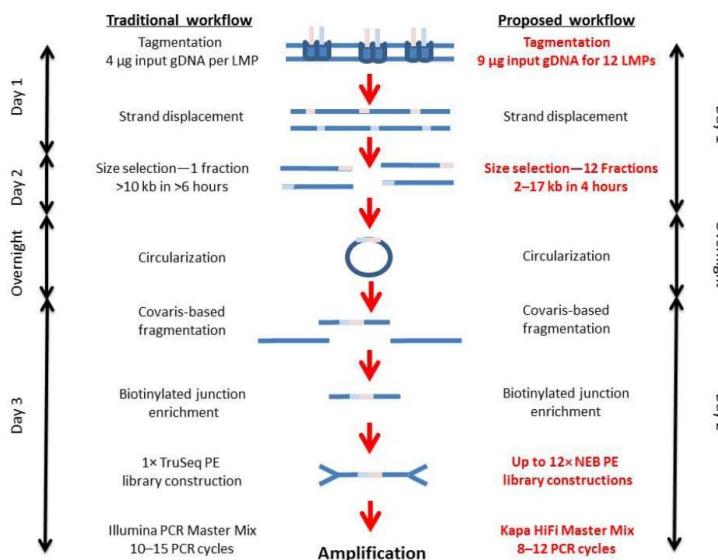
ously difficult to make, especially for the larger insert sizes.

Using the Illumina Nextera Mate Pair Sample Preparation Kit (Illumina, San Diego, CA), libraries can be constructed from as little as 1 µg of genomic DNA (gDNA) using the Nextera transposase to fragment DNA and tag the molecules with known sequences (a process known as fragmentation). However, these libraries tend to have a broad insert size which can range from 1 kb to 12 kb (Supplementary Figure S2). As a result, many labs employ gel-based size selection to generate specific insert sizes that can be supplied to the scaffolding algorithm, thereby simplifying the scaffolding step. Semi-automated gel approaches such as BluePippin (Sage Science, Beverly, MA) improve this process but limit throughput to four libraries at a time and use more input DNA. Constructing 4 LMP libraries, could require >18 µg of DNA, and if insert sizes >10 kb are targeted, each size selection run would last longer than 6 h, meaning that library construction could take up to 3 days to complete (Figure 1). Furthermore, in our experience it is hard to predict how a specific DNA sample will perform in a fragmentation reaction, so more than one reaction is often needed to obtain a specific size. Finally, there can be 10%–20% variance between the targeted and recovered DNA size on a BluePippin.

We optimized the Nextera based LMP Library Construction kit to maximize fragmentation across the largest possible size range using the minimum amount of input material. Using gDNA isolated from the bread wheat (*Triticum aestivum*) variety Chinese Spring 42, we performed just 2 Gel Plus fragmentation reactions and subsequent strand displacements to construct 12 LMP libraries. This allows us to construct 60 LMP libraries from 5 samples using a 10-reaction kit. As fragment size in a Nextera reaction is controlled by the ratio of DNA and Nextera enzyme, one reaction was performed with 3 µg of input DNA, and another with 6 µg. The two Nextera reactions were then pooled post strand displacement, and the range of fragment sizes confirmed by analyzing the profiles on

### METHOD SUMMARY

We present a method to simultaneously size select and construct up to 12 long mate pair (LMP) libraries at a time and then map the generated reads back to the available assembled sequences to accurately calculate insert sizes. These calculations can then be used to determine which libraries to sequence to greater depth and to use the accurate insert size information in de novo genome assemblies to improve outputs.



**Figure 1. Nextera-based long mate pair (LMP) workflow.** The traditional LMP workflow compared with our proposed workflow with differences between the two workflows highlighted in red.

an Agilent BioAnalyser 12000 chip (Agilent, Stockport, UK) (Supplementary Figure S1). By using 2 independent fragmentation reactions, we ensured the material entering size selection ranged from 1.5 kb to >17 kb with a good distribution, allowing us to construct LMP libraries from a wide range of insert sizes.

Size selection was performed on a Sage Science Electrophoretic Lateral Fractionator (SageELF), which is unique in its ability to simultaneously isolate 12 different discrete size fractions from a single sample loading. The pooled, strand-displaced reactions were loaded onto a 0.75% cassette, which was configured to separate the sample for 3 h 30 min and then elute 12 fractions over 35 min. Post size selection, the size of each of the 12 isolated fractions was measured on an Agilent BioAnalyser Chip 12000 (Figure 2A and Table 1), and the yield was determined using a High Sensitivity Qubit Assay (Thermo Fisher, Cambridge, UK) (Supplementary Table S1).

We loaded 5 µg of DNA onto the SageELF and recovered >2 µg across the 12 fractions, which represents >40% of the starting material. Fraction 5 encompassed an important LMP target insert size of 8 kb (a very common transposon in wheat is ~7 kb). For this size, we managed to recover >180 ng of material (Supplementary Table S1). To compare this against our standard

approach, we fragmented and strand displaced 4 µg of the same wheat gDNA, and confirmed the fragmentation profile on a 12000 BioAnalyser Chip (Supplementary Figure S1). After targeting an 8 kb (7.4–8.6 kb) size selection on a BluePippin, with the improved recovery protocol we recovered only 56 ng of material. When we ran this out on a 12000 BioAnalyser Chip, it estimated the fragments to be centered on 9.5 kb and spanning 8.0–10.5 kb (data not shown), which illustrates the problem with targeting specific insert sizes. For the comparable SageELF fraction (Fraction 4) we recovered

261 ng of material centered on 9.5 kb and spanning 8.5–10.6 kb (data not shown) highlighting that size selection is not only tighter, but we also observed significantly higher recoveries when using the SageELF.

Circularization of the SageELF fractions was then performed overnight at 30°C, followed by exonuclease (Illumina) treatment at 37°C for 30 m, incubation at 70°C for 30 m to denature the enzyme, and then addition of Stop Ligation buffer (Illumina). Circularized fragments were then sheared on a Covaris S2 (Covaris, Woburn, MA), targeting a 450 bp shear, and then library molecules containing the biotinylated junction adapter were bound to M280 streptavidin-coated beads (Thermo Fisher). Fragmented molecules from each of the 12 size-selected fractions were end repaired and A-tailed using the relevant NEB modules (NEB, Hitchin, UK) and then Illumina TruSeq adaptors (Illumina) were ligated (each size fraction received a different index) with NEB Blunt T/A ligase (NEB).

We used Kapa HiFi polymerase (Kapa Biosystems, London, UK) for its improved performance, especially in GC rich regions, instead of the Illumina PCR master mix (4). Post size selection, we calculated the copy number of each fraction based on the predicted size from the SageELF and the yield to measure the library complexity. For samples with a copy number  $>3.75 \times 10^{10}$  we performed 8 PCR cycles, for samples with a copy number between  $2 \times 10^{10}$  and  $3.75 \times 10^{10}$ , 10 cycles were performed, and for samples with a copy number  $<2 \times 10^{10}$ , 12 cycles were performed. The library molecules were amplified directly from the

**Table 1. Sizes of long mate pair (LMP) inserts for each fraction as determined by the SageELF, BioAnalyser, and mapping reads back to the wheat chromosome 3B assembly.**

Fraction	ELF library size (kb)	BA 12000 library size (kb)	Mapped insert size (kb)
1	16.18	Not determined	Insufficient data
2	13.31	Not determined	14.8
3	11.74	12.52	11.3
4	9.81	9.24	9.0
5	8.00	8.03	7.3
6	6.46	6.68	5.9
7	5.16	5.37	4.8
8	4.28	4.31	3.8
9	3.70	3.46	3.2
10	2.93	2.66	2.4
11	2.22	2.16	1.9
12	1.71	1.67	1.4

# BLOW UP THE BARRIERS TO YOUR NEXT-GEN SEQUENCING



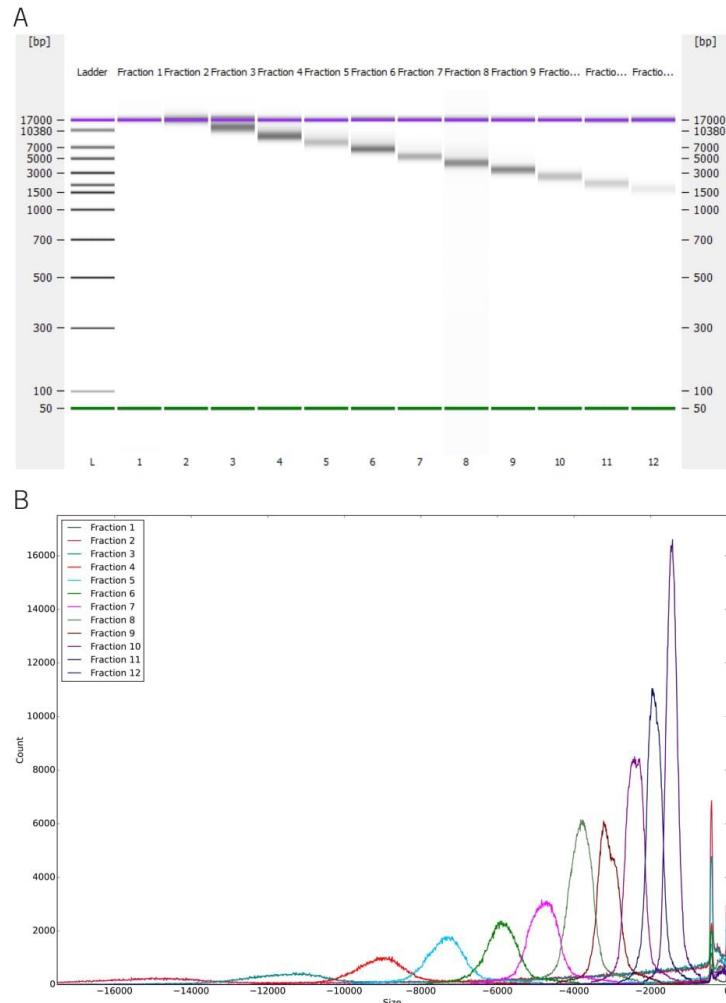
Next-gen sample QC is now hassle free.

**FULLY AUTOMATED  
FRAGMENT ANALYZER™  
DOES IT ALL.**

- Assesses quality and quantity (size and concentration)
- Resolves fragments from 25 bp to 5,000 bp
- Sizes fragments up to 20,000 bp for PacBio sequencers
- Also analyzes gDNA and RNA

**No chips. No tapes. No compromises.**  
More at [AATI-US.COM](http://AATI-US.COM)





**Figure 2. BioAnalyser images of DNA post size selection and size distribution of BWA mapped reads.** (A) SageELF size-selected fractions were analyzed to estimate fragment length prior to circularization. (B) NextClip filtered reads from each size-selected fraction library were aligned against the wheat chromosome 3B assembly and the number of reads vs. insert size plotted.

streptavidin beads using Kapa HiFi and the Illumina primer cocktail (Illumina). We aimed to maintain library complexity and reduce PCR duplication rates while generating sufficient material for multiple HiSeq runs (Supplementary Table S1).

Post amplification, a CleanPCR (GC Biotech, Alphen aan den Rijn, The Netherlands) bead clean-up was carried out, and the final library was eluted in 20  $\mu$ l resuspension buffer (Illumina). Library quality controls were performed by running an Agilent BioAnalyser High Sensitivity chip, and the DNA concentrations

were measured using the High Sensitivity Qubit assay (Supplementary Table S1). Equimolar amounts of LMP libraries from fractions 2–12 were then pooled, with the library from fraction 1 spiked in at one-tenth the concentration of the others due to it being relatively weaker (Supplementary Table S1). The 12 pooled libraries were size selected on a BluePippin to ensure that all library fragments would have insert sizes between 370 and 470 bp (maximizing usable mate pairs) and then quantified using the Kapa qPCR Illumina Quantification kit.

To validate the pool and accurately determine the insert size of each LMP, the pools were run on a MiSeq (Illumina) with 2 × 300 bp reads. Sequence data were screened via a primary analysis pipeline to demultiplex reads based on library indexes and to determine basic run metrics, including duplication rate, GC content, and the presence of over-represented sequences (5). The data were then processed through NextClip (6) to classify LMP reads. Those deemed as true mate pairs, based on the presence of the Nextera junction sequence within the reads with sufficient sequence either side, were then mapped using BWA-mem (7) to the bread wheat (*Triticum aestivum*) variety Chinese Spring 42 chromosome 3B reference sequence (8) using default parameters, and the insert size for each library determined and plotted (Figure 2B and Table 1).

Using the SageELF streamlines the library construction process, allowing LMP libraries >10 kb to be constructed in under 2 days with <10 µg input material. For many genome projects, multiple insert size LMP libraries are required, and the ability to construct up to 12 discretely sized libraries for a combined reagent cost of \$1270 compared with the reagent cost of \$715 for a single insert size LMP library highlights the potential cost savings. We also observe significant improvements with increased yield and tighter size selection than when using the BluePippin, especially when looking to construct LMP libraries with insert sizes >10 kb.

Accurately determining the size and span of the inserts for mate pair libraries simplifies the scaffolding problem, enabling the assembly of longer, more precise sequences with fewer non-determined bases (runs of N bases), empowering all subsequent downstream analysis. Although the BioAnalyser and SageELF both estimate the size of fraction 5 to be 8 kb, mapping the sequence data back to the wheat chromosome 3B assembly suggested that the size is in fact 7.2 kb (Table 1). This demonstrates the benefit of this approach both in terms of accuracy in determining insert size and also the ability to sequence slightly larger or slightly smaller insert libraries without having to repeat the whole process if one library isn't deemed suitable. It also gives the flexibility of running all 12 libraries if desired.

## Author contributions

D.H. wrote the manuscript and carried out the experiments. G.G.A. and B.C. analyzed the sequence data. D.H., B.C., and M.D.C. had the original idea and designed the study. D.H., G.G.A., B.C., and M.D.C. edited the manuscript. B.C. and M.D.C. supervised the study.

## Acknowledgments

Wheat gDNA was provided by Neil Mckenzie and Mike Bevan, John Innes Centre. Library quantification, the MiSeq Sequencing, and Primary Analysis Pipeline were run by the Platforms and Pipeline Team at TGAC. This work was supported by a BBSRC Triticeae Genomics for Sustainable Agriculture Grant, BB/J003743/1, and a BBSRC National Capability Grant, BB/J010375/1.

## Competing interests

The authors declare no competing interests.

## References

1. Magoc, T., S. Pabinger, S. Canzar, X. Liu, Q. Su, D. Puu, L.J. Tallon, S.L. Salzberg. 2013. GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics*. 29:1718-1725.
2. Treangen, T.J. and S.L. Salzberg. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*. 13:36-46.
3. Nagarajan, N. and M. Pop. 2013. Sequence assembly demystified. *Nat Rev Genet*. 14:157-167.
4. Quail M.A., T.D. Otto, Y. Gu, S.R. Harris, T.F. Skelly, J.A. McQuillan, H.P. Swerdlow, S.O. Oyola 2011. Optimal enzymes for amplifying sequencing libraries. *Nat Methods*. 9:10-11.
5. Leggett, R.M., R.H. Ramirez-Gonzalez, B.J. Clavijo, D. Waite, and R.P. Davey. 2013. Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. *Front Genet*. 4:288.
6. Leggett, R.M., B.J. Clavijo, L. Cissold, M.D. Clark, M. Caccamo. 2014. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics*. 30:566-569.
7. Li, H. and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* 25:1754-1760.
8. Choulet, F., A. Albert, S. Theil, N. Glover, V. Barbe, J. Daron, L. Pingault, P. Sourdille, et al. 2014. Structural and functional partitioning of bread wheat chromosome 3B. *Science*. 345:1249721.

Received 20 February 2015; accepted 13 April 2015.

Address correspondence to Darren Heavens, The Genome Analysis Centre (TGAC), Norwich Research Park, Norwich, UK, NR4 7UH. E-mail: darren.heavens@tgac.ac.uk

To purchase reprints of this article, contact: [biotechniques@fosterprinting.com](mailto:biotechniques@fosterprinting.com)

 KAPA BIOSYSTEMS

## Introducing KAPA HYPER PLUS

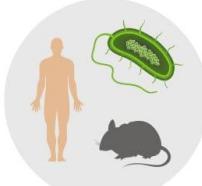
Single-tube DNA  
fragmentation and library  
preparation in 2.5 hours



Speed of fragmentation



Quality of  
mechanical shearing



Flexible sample types  
and input amounts



Reduced  
sequencing costs

Visit  
[kapabiosystems.com/hyperplus](http://kapabiosystems.com/hyperplus)  
to request a trial kit

1    **Supplementary Material For:**

2    **A method to simultaneously construct up to 12 differently sized Illumina**

3    **Nextera long mate pair libraries with reduced DNA input, time, and cost**

4    Darren Heavens, Gonzalo Garcia Accinelli, Bernardo Clavijo, and Matthew Derek Clark

5    *The Genome Analysis Centre (TGAC), Norwich Research Park, Norwich, UK*

6    *BioTechniques* 59:42-45 (July 2015) doi 10.2144/000114310

7    For the Tagmentation reactions 3 $\mu$ g and 6 $\mu$ g of Genomic DNA was prepared in 308 $\mu$ l of water and

8    then placed in the heat block set at 55 $^{\circ}$ C for 6 minutes to equilibrate. Then 80 $\mu$ l 5x Tagment Buffer

9    Mate Pair (Illumina, San Diego, USA) added followed by12 $\mu$ l Mate Pair Tagmentation Enzyme

10   (Illumina) and the reaction gently vortexed to mix. This was then incubated for 30 minutes at 55 $^{\circ}$ C,

11   100 $\mu$ l of Neutralize Tagment Buffer (Illumina) added and then incubated at room temperature for 5

12   minutes. A 1x volume bead clean-up was performed with CleanPCR beads (GC Biotech, Alphen aan

13   den Rijn, The Netherlands) and the DNA eluted in 170 $\mu$ l of Nuclease free Water (Qiagen, Manchester,

14   UK). A 1 $\mu$ l aliquot was run on a BioAnalyser 1200 chip (Agilent, Stockport, UK) (Supplementary

15   Material Figure 1) and DNA concentration determined using a Qubit HS Assay (Thermo Fisher,

16   Cambridge, UK).

17   Strand Displacement was performed by combining 162 $\mu$ l of tagmented DNA, 20 $\mu$ l 10x Strand

18   Displacement Buffer (Illumina), 8 $\mu$ l dNTPs (Illumina) and 10 $\mu$ l Strand Displacement Polymerase

19   (Illumina). This was then incubated at room temperature for 30 minutes. A 0.75x volume bead clean-

20   up was performed with CleanPCR beads and the DNA eluted in 16 $\mu$ l of Nuclease free Water and the

21   eluted DNA from the 3 $\mu$ g and 6 $\mu$ g reactions pooled. A 1 $\mu$ l aliquot was diluted 1:6 and run on a

22   BioAnalyser 1200 chip (Supplementary Material Figure 1) and DNA concentration determined using a

23   Qubit HS Assay.

24   Size selection was performed on a Sage Science ELF (Sage Science, Beverly, USA). The 30 $\mu$ l in each

25   of collection wells was replaced with fresh buffer and the collection and elution current checked prior

26   to loading the sample. To 30 $\mu$ l of the pooled Strand Displaced reaction 10 $\mu$ l of loading solution was

27   added and then loaded onto a 0.75% Cassette which was configured to separate the sample for 3

Page 1 of 4

28 hours 30 minutes and then eluting each fraction for 35 minutes. Post size selection, the 30 $\mu$ l from  
29 each of the 12 collection wells was recovered and the size isolated in each fraction estimated on  
30 12000 BioAnalyser Chip and DNA concentration determined using a Qubit HS Assay (Supplementary  
31 Material Table 1).

32 **Supplementary Material Table 1.** Experimental data recorded during the construction of LMPs.

Fraction	Yield Post Size Selection (ng)	Copy Number Entering Circularisation	PCR Cycle Number	Final Library Yield (ng)
1	53.4	2.54E+09	12	54.8
2	169.2	1.18E+10	12	346
3	245.4	1.94E+10	12	548
4	261	2.46E+10	10	744
5	181	2.08E+10	10	872
6	248.4	3.56E+10	10	808
7	153	2.59E+10	10	1420
8	204	4.42E+10	8	1040
9	184.8	4.63E+10	8	1568
10	120	3.79E+10	8	1512
11	109.2	4.56E+10	8	1716
12	75	4.06E+10	8	1280

33

34 Circularisation was performed by combining 30 $\mu$ l of size fractionated DNA, 12.5 $\mu$ l of 10x  
35 circularisation buffer (Illumina), 3 $\mu$ l Circularisation Enzyme (Illumina) and 85 $\mu$ l nuclease free water.  
36 These were then incubated at 30 $^{\circ}$ C overnight. Linear DNA was digested by adding 3.75 $\mu$ l Exonuclease  
37 (Illumina) and incubating at 37 $^{\circ}$ C for 30 minutes followed by 70 $^{\circ}$ C for 30 minutes to denature the  
38 enzyme and 5 $\mu$ l of stop ligation (Illumina) added. During exonuclease treatment 240 $\mu$ l of M280  
39 Dynabeads (Thermo Fisher) were prepared by washing twice with 600 $\mu$ l Bead Bind Buffer (Illumina)  
40 before resuspending in 1560 $\mu$ l Bead Bind Buffer. Circularised DNA was then sheared in a 130 $\mu$ l  
41 volume on a Covaris S2 (Covaris, Massachusetts, USA) for 2 cycles of 37seconds with a duty cycle of  
42 10%, cycles per burst of 200 and intensity of 4.

43 To 130 $\mu$ l fragmented DNA 130 $\mu$ l of washed M280 beads was added, mixed and then placed on a lab  
44 rotator at room temperature for 20 minutes. Library molecules bound to M280 beads were then  
45 washed four times with 200 $\mu$ l Bead Washer Buffer (Illumina) and twice with 200 $\mu$ l Resuspension  
46 Buffer (Illumina).

47 A master mix containing 1105 $\mu$ l nuclease free water, 130 $\mu$ l 10x End Repair Reaction Buffer (NEB,  
48 Hitchin, UK) and 65 $\mu$ l end repair enzyme mix (NEB) was prepared and 100 $\mu$ l added to each tube,  
49 mixed with the beads and incubated at room temperature for 30 minutes. End repaired library  
50 molecules bound to M280 beads were then washed four times with 200 $\mu$ l Bead Washer Buffer and  
51 twice with 200 $\mu$ l Resuspension Buffer.

52 A master mix containing 325 $\mu$ l nuclease free water, 39 $\mu$ l A Tailing 10x Reaction Buffer (NEB) and 26 $\mu$ l  
53 A tailing enzyme mix (NEB) was prepared and 30 $\mu$ l added to each tube, mixed with the beads and  
54 incubated at 37°C for 30 minutes. To the A tailed library molecules 1 $\mu$ l of the appropriate Illumina  
55 Index adapter (Illumina) was added and mixed then 31 $\mu$ l of Blunt/ TA ligase (NEB) added and  
56 incubated at 30oC for 10m. Post incubation 5 $\mu$ l of stop ligation added and then the adapter ligated  
57 library molecules bound to M280 beads were then washed four times with 200 $\mu$ l Bead Washer Buffer  
58 and twice with 200 $\mu$ l Resuspension Buffer.

59 A master mix containing 240 $\mu$ l nuclease free water, 300 $\mu$ l 2x Kapa HiFi (Kapa Biosystems, London, UK  
60 ) and 60 $\mu$ l Illumina Primer Cocktail (Illumina) was prepared and 50 $\mu$ l added to each tube, mixed with  
61 the beads and the contents, including beads, transferred to a 200 $\mu$ l PCR tube. Each sample was then  
62 subjected to amplification on a Veriti Thermal Cycler (Thermo Fisher) with the following conditions:-  
63 98°C for 3 minutes, 8, 10 or 12 cycles of PCR depending upon copy number entering circularisation  
64 (Supplementary Material Table 1) of 98°C for 10 seconds, 60°C for 30 seconds, 72°C for 30 seconds  
65 followed by 72°C for 5 minutes and Hold at 4°C.

66 Post amplification the PCR tubes were placed on a magnetic plate, the beads allowed to pellet and  
67 then 45 $\mu$ l of the PCR transferred to a 2ml Lobind Eppendorf Tube. To this 31.5 $\mu$ l beads of CleanPCR  
68 beads were added to precipitate the DNA, the beads washed twice with 70% ethanol and the final  
69 library eluted in 20 $\mu$ l resuspension buffer. Library QC was performed by running a 1 $\mu$ l aliquot on a  
70 High Sensitivity BioAnalyser chip (Agilent) and the DNA concentration measured using the High  
71 Sensitivity Qubit (Supplementary Material Table 1). Each library was then equimolar pooled (except  
72 for the largest insert library which was considerably weaker than the others which was at 10%  
73 concentration) based on DNA concentration and CleanPCR beads used to concentrate the sample  
74 down to 30 $\mu$ l.

75 The pooled library was then subjected to size selection on a Blue Pippin (Sage). The 40 $\mu$ l in each of  
76 collection wells was replaced with fresh buffer and the separation and elution current checked prior to  
77 loading the sample. To 30 $\mu$ l of the pooled library 10 $\mu$ l of R2 marker solution was added and then  
78 loaded onto a 1.5% Cassette. The Blue Pippin was configured to size select between 600 and 700bp  
79 and run for 50 minutes. Post size selection, the 40 $\mu$ l from the collection wells was recovered and the  
80 size isolated estimated on High Sensitvity BioAnalyser Chip and DNA concentration determined using  
81 a Qubit HS Assay. The quantification of the pool was determined by the Kapa qPCR Illumina  
82 quantification kit (Kapa) with the pool run at 10pM on a MiSeq (Illumina) with 2x300bp reads. The  
83 run clustered at 880k cluster per mm<sup>2</sup>, generating 11.3Gbp of sequence.

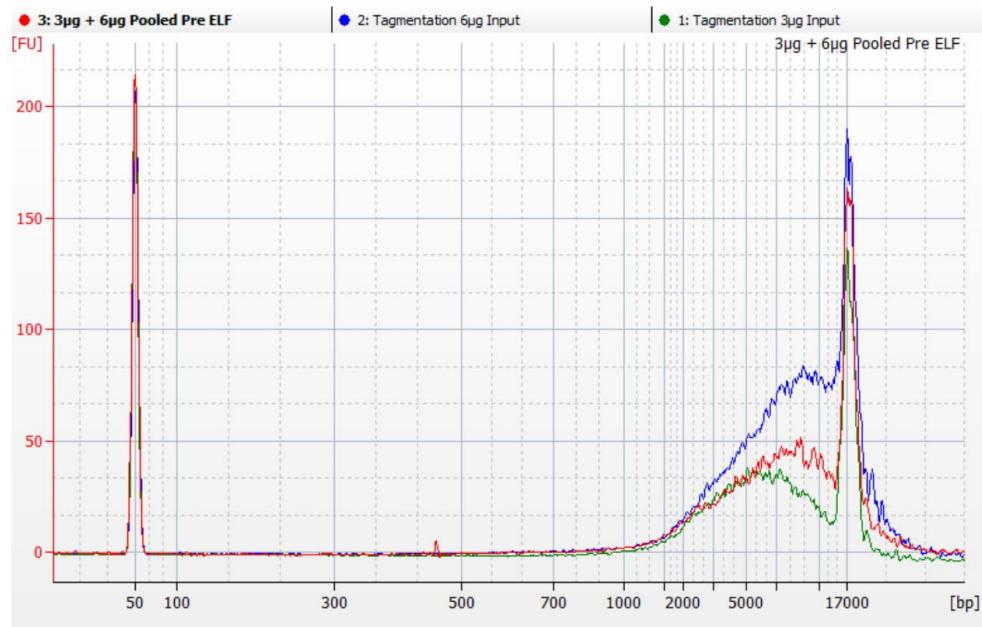
84 Reads generated were then processed through NextClip which takes LMP FASTA reads and looks to  
85 categorise them into four groups based on the presence of the Nextera adapter junction sequence.  
86 Category A pairs contain the adaptor in both reads, Category B pairs contain the adaptor in only read  
87 2, Category C pairs contain the adaptor in only read 1, Category D pairs do not contain the adaptor in  
88 either read. NextClip also uses a k-mer-based approach to estimate the PCR duplication rate while  
89 reads are examined. Filtered reads in categories A, B and C were then mapped back to the Wheat  
90 Chromosome 3B reference using BWA mem with the default parameters. This uses the reference  
91 sequence and measures from the leftmost to the rightmost aligned bases within the reads to  
92 determine the insert size.

93

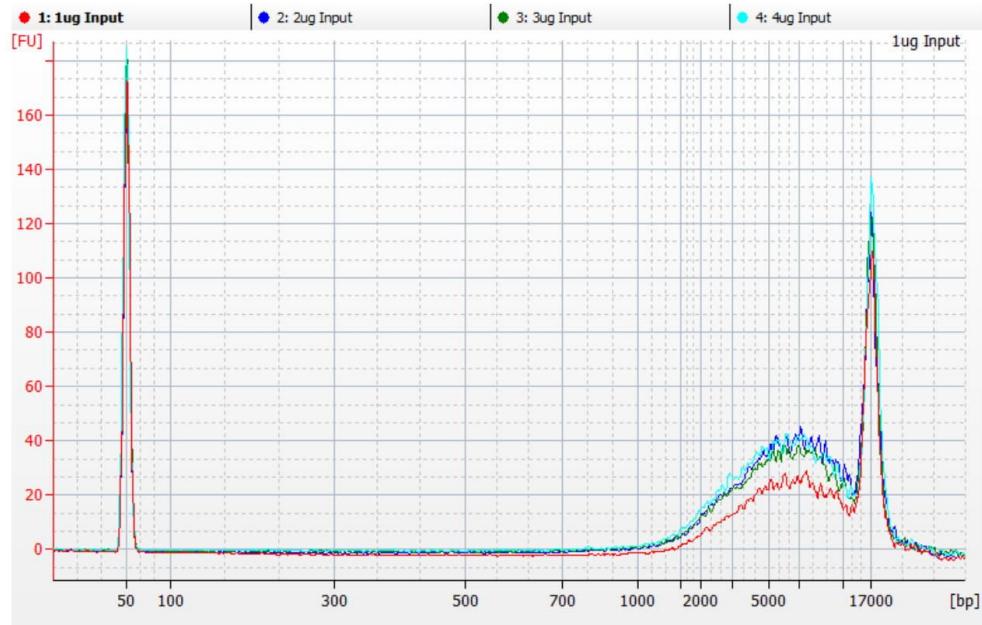
94 **Supplementary Material Figure Legends**

95 **Supplementary Material Figure 1. BioAnalyser Images of DNA pre size selection.** 3 $\mu$ g input  
96 (green) and 6 $\mu$ g input (red) tagmented DNA were pooled post strand displacement (blue).

97 **Supplementary Material Figure 2. BioAnalyser Images of Tagmented DNA with different**  
98 **DNA inputs.** Distribution of tagmented DNA fragments with 1 $\mu$ g input (red), 2 $\mu$ g input (blue), 3 $\mu$ g  
99 input (green) and 4 $\mu$ g input (turquoise) in 100 $\mu$ l, 200 $\mu$ l, 300 $\mu$ l and 400 $\mu$ l reaction volumes  
100 respectively



Supplementary Figure S1



Supplementary Figure S2

# W2RAP: a pipeline for high quality, robust assemblies of large complex genomes from short read data

Bernardo J. Clavijo<sup>1\*</sup>, Gonzalo Garcia Accinelli<sup>1</sup>, Jonathan Wright<sup>1</sup>, Darren Heavens<sup>1</sup>, Katie Barr<sup>1</sup>, Luis Yanes<sup>1</sup>, Federica Di-Palma<sup>1</sup>

*1 Earlham Institute, Norwich Research Park, UK.*

February 23, 2017

## Abstract

Producing high-quality whole-genome shotgun *de novo* assemblies from plant and animal species with large and complex genomes using low-cost short read sequencing technologies remains a challenge. But when the right sequencing data, with appropriate quality control, is assembled using approaches focused on robustness of the process rather than maximization of a single metric such as the usual contiguity estimators, good quality assemblies with informative value for comparative analyses can be produced. Here we present a complete method described from data generation and qc all the way up to scaffold of complex genomes using Illumina short reads and its application to data from plants and human datasets. We show how to use the w2rap pipeline following a metric-guided approach to produce cost-effective assemblies. The assemblies are highly accurate, provide good coverage of the genome and show good short range contiguity. Our pipeline has already enabled the rapid, cost-effective generation of *de novo* genome assemblies from large, polyploid crop species with a focus on comparative genomics.

**Availability:** w2rap is available under MIT license, with some sub-components under GPL-licenses. A ready-to-run docker with all software pre-requisites and example data is also available.

<http://github.com/bioinfologics/w2rap>  
<http://github.com/bioinfologics/w2rap-contigger>

---

\*to whom correspondence should be addressed

## 1 Introduction

Generation of a high quality genome assembly is a crucial first-step towards understanding the biology of an organism. It establishes a complete catalogue of genes and provides the foundation for characterising the genetic variation within a species and how this variation impacts gene function and phenotypic variation. Over the last 10 years, many methods have been described to address this problem and the genomes of many organisms have been published. Yet, for plant and animal species which often have large and complex genomes, assembly remains a fundamental challenge.

Genome assemblies generated from massively parallel short-read technologies such as Illumina are highly accurate at the nucleotide level and relatively inexpensive to generate, but remain highly fragmented due to complex repeat content and varying degrees of polymorphism and ploidy. While solutions such as ALLPATHS-LG (Gnerre et al., 2010), have revolutionised the field enabling the sequencing and assembly of many mammalian genomes and providing a foundation for large scale comparative analysis and lineage-specific evolutionary analysis, they require a precise recipe of input libraries coupled to a fixed set of algorithm parameters which are not suitable for larger, complex genomes.

Here we present a pipeline called w2rap (Wheat/Whole-genome Robust Assembly Pipeline) to rapidly generate high-quality, low-cost, robust assemblies from genomes with different levels of complexity. Our approach uses Illumina PCR-free paired end (PE) 250bp reads for contig construction with the w2rap-contigger, an improved algorithm based on DISCOVAR *denovo* (Weisenfeld et al., 2014) (Love et al., 2016), and Nextera long mate-pair (LMP) libraries for long-range scaffolding with SOAPdenovo2 (Luo et al., 2012). W2rap encompasses a full data processing workflow from raw reads to scaffolds, and crucially allows the user to fine tune the algorithmic parameters making draft assembly generation an iterative process adaptable to diverse genome complexities and data. We also show in our w2rap test *A. thaliana* dataset that tuning assembly to enhance accuracy produces more contiguous assemblies at lower computational cost, as the downstream analysis problems become easier.

We demonstrate our approach here by applying it to a datasets from *Arabidopsis thaliana* and show how it performs in line with state of the art approaches in standard *Homo sapiens* data. We have already used it to assemble the hexaploid, highly repetitive, 17Gbp *Triticum aestivum* (bread wheat) genome which generated highly accurate scaffolds in agreement with the existing single chromosome reference sequence (Clavijo et al., 2016). Our results maintain the completeness and accuracy achieved by DISCOVAR *denovo* coupled with reduced memory usage, and processing time. Most importantly, increased accuracy and contiguity are achieved by enhanced parameterisation of the algorithms, improved repeat resolution, and the systematic use of LMP data via SOAPdenovo2 scaffolding. This method makes it possible to consistently generate high-quality draft assemblies for large, complex genomes at low cost.

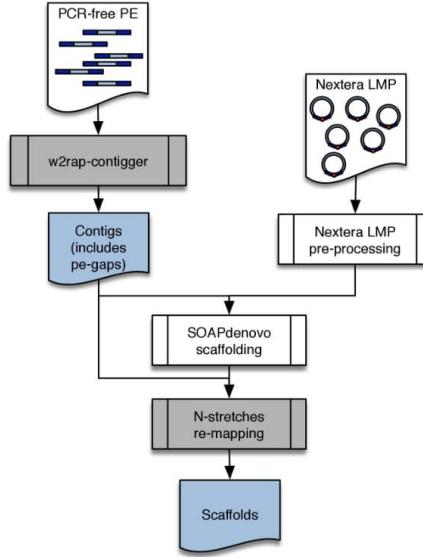


Figure 1: w2rap assembly workflow. Contigs are produced by running the w2rap-contigger on PCR-free 2x250bp Illumina data. Nextera Long Mate Paired reads are then pre-processed with the w2rap scripts and used to scaffold the contigs using the SOAPdenovo2 scaffolder included in w2rap. A script then re-introduces N-runs from the original contigs displaced during SOAPdenovo2 scaffolding.

## 2 Results

### 2.1 Data generation

W2rap uses a combination of Illumina Paired End (PE) and Long Mate Pair (LMP) reads. We recommend PCR-free PE libraries, using fragment sizes of about 700bp for optimal short-repeat resolution in the w2rap-contigger stage. Nextera LMP libraries (Heavens et al., 2015) are used to provide precise and cost-effective longer range information.

Whilst different combinations of coverages and library sizes can be used depending on genome complexity, repeat structure, and other characteristics, we

recommend a minimum coverage of 30x PE and up to 100x for highly heterozygous genomes. For LMP libraries, a minimum coverage of 30x raw reads per library is recommended, with up to 50x being routinely used in internal projects. For plant genomes that often contain high levels of LTR-retrotransposons, it is important to include a LMP library longer than 8Kbp to span these highly repetitive blocks. We have successfully used a combination of libraries of 8Kbp, 10Kbp and 14Kbp for complex genomes such as wheat. As a general rule, when trying to iteratively improve an assembly by sequencing a range of LMP libraries, we recommend getting 30x LMP coverage of a library with an insert size corresponding to the N80 of the current scaffolding results.

## 2.2 Assembly quality and contiguity

When all algorithmic heuristics perform as intended, the genome assembly results should become more accurate, and this accuracy can drive contiguity. Table 1 and Figure 1 show a clear example of an *A. thaliana* assembly becoming more accurate and contiguous, by changing parameters of the same heuristics to achieve better processing. The scaffolding results from both the *A. thaliana* contigs and scaffolds outperform both ABYSS and SOAP for our test runs on contiguity, and outperform DISCOVAR *denovo* contigs for both accuracy and contiguity. Details of the assembly parameters for each of these runs are given on Supplementary Material Section 5.

We also used the human HG004 Illumina 2x250bp PE and LMP datasets from the Genome in a Bottle Consortium (Zook et al., 2016), and we cite for comparison results from the ABYSS 2.0 publication (Jackman et al., 2016). It is important to highlight the main focus of w2rap is producing assemblies for comparative studies. Therefore, it is crucial to produce assemblies that are consistently comparable and not one-off optimal settings for a particular dataset. The default w2rap-contigger parameters achieve similar results to those of DISCOVAR *denovo*, as they run essentially the same heuristics with near-identical parameter choice, but achieve slightly less contiguity as seen for the HG004 dataset in Table 2. This comes from slightly more conservative choices of parameters. A ‘dv\_like’ mode can be activated when running the w2rap-contigger that produces results more similar to those of DISCOVAR *denovo*. The general results from these assemblies are comparable to those mentioned on the ABYSS 2.0 publication.

## 2.3 Computational performance of w2rap-contigger

Among the disadvantages of using DISCOVAR *denovo* to assemble large and complex genomes are its high memory consumption and long runtime and for projects aiming to generate multiple assemblies computational resources can become a major bottleneck. We reviewed the whole codebase and re-implemented many algorithms with openMP-based parallel approaches, testing and improving the performance on NUMA systems, such as those needed to handle the memory and processing requirements of larger genome assemblies. This led to

reduced processing times, both in NUMA systems and normal servers. Further reduction on the processing times can be achieved by correct parametrisation of the first steps of the contig assembly process, as this leads to simpler problems on the later stages (See Supplementary Material for more details).

Each step during contig assembly uses significantly different algorithmic approaches and data. We segmented the w2rap-contigger processing into eight steps which can be run independently thus enabling us to make more efficient usage of resources when running multiple assemblies or sharing computational resources with other projects. This change produced two desired outcomes: (i) each step runs with the resources required for that step only, thus avoiding a waste of computing resources on large-memory multi-processor machines and, (ii) the granularity of running shorter steps rather than all steps combined allows for better control over the assembly, and provides the opportunity for a detailed check of results from intermediate steps. These modifications are important when assembling large and complex genomes, where the contigging steps can take over 10 days.

Assembler	Contig NG50	Contig NGA50	Scaffold NG50	Scaffold NGA50
ABYSS v1.9	57	56.57	412.58	370.42
SOAPdenovo2	21.5	21.49	937	365.86
w2rap	361.9	355.89	1318.81	1136.96

**Table 1:** Comparison between w2rap and other assembly tools for the *A. thaliana* dataset. All values in Kbp.

Assembler	Contig NG50 (Kbp)	Scaffold NG50 (Mbp)
ABYSS v1.9	30.0	4.36
SOAPdenovo2	3.8	0.17
w2rap	90.6	4.78

**Table 2:** Comparison between w2rap with other assembly tools for the HG004 *H. sapiens* dataset from Genome in a Bottle.

\* Data for ABYSS and SOAPdenovo2 results from ABYSS 2.0 preprint.

### 3 Methods

#### 3.1 Data generation

PCR-free Illumina paired end libraries are both cost-effective and less prone to representation bias (Huptas et al., 2016). Standard Illumina NGS library construction protocols target fragments with a 500bp insert size but these typically span 300bp to 700bp, and smaller molecules within a library are more likely to be sequenced over larger molecules. These libraries reduce effective cover-

age due to overlapping sequences and fail to provide the spatial information due to the predominance short insert fragments. For *de novo* genome assembly projects, greater accuracy and contiguity can be achieved if the unique sequence flanking repeats can be resolved within multiple single paired reads. We aim for fragments of 700bp and longer which can be achieved by using a more stringent AMPure XP bead based clean up (for reproducibility it is recommended to use a positive displacement pipette to accurately dispense the beads) or a size exclusion technology such as the Blue Pippin from Sage Science.

We recommend following our modifications to the Nextera LMP protocol to produce libraries with good size distributions and representation (Heavens et al., 2015). Processing of these is explained in a later section, but emphasis must be on QC. When sequencing large genomes, we recommend constructing multiple libraries, sequencing a combined pool in a single low-coverage (and low cost) run, then choose the libraries with the best characteristics for further sequencing.

### 3.2 Generating contigs with the w2rap-contigger

The w2rap-contigger is a extensively modified version of DISCOVAR *denovo*. The original DISCOVAR has some limitations in terms of large repeat-rich datasets, which made it impossible to run it on genomes such as that of hexaploid wheat. We fixed bugs and limitations on the repeat-resolution heuristics and implemented an extra repeat-resolution heuristic, the PathFinder, that is described in the Supplementary Material. We divided the original assembly heuristics into discrete steps, both to optimise the usage of computational resources and to make the heuristic processing easier to track. Supplementary Material Section 1 contains more detail about each step.

The w2rap-contigger provides a more extensive set of parameters than that of DISCOVAR *denovo*, although most of these were originally present in the code but fixed to values that were reasonable for mammalian genomes as sequenced by The Broad Institute. In general, when sequencing multiple related genomes for comparative studies, a sequencing recipe should be devised and then a set of parameters chosen for the whole study, to guarantee comparable results.

While trying to adjust parameters for contiguity is a widespread practice in genome assembly, we have shown in the results section and Figure 1 that increasing accuracy can lead to higher contiguity. This means the assembly process must be guided by the careful execution of each heuristic to achieve accuracy. There needs to be an understanding of each heuristic and a method to measure whether each heuristic is achieving the desired results.

#### 3.2.1 Understanding the w2rap-contigger metrics

At the beginning and end of each step, and at every relevant point during execution, the w2rap-contigger prints a set of assembly status metrics: kmers in the graph, graph contiguity, reads pathing (i.e. a single-end read placement) and pair status. (See Supplementary Material for details).

The w2rap-contigger represents the assembly graph internally as a list of edges (sequences or gaps) with a list of vertices representing K-1 overlaps. The graph is directed, with independent reverse complements, which means every sequence will be represented both in forward and reverse unless it is palindromic. This means the number of kmers in the graph will be roughly doubled. Alongside the number of kmers and edges, a set of N<sub>k20</sub>, N<sub>k50</sub> and N<sub>k80</sub> values show the length in kmers of the edges such that edges of that length or more cover 20%, 50% and 80% of the total kmer size of the graph. We use this value rather than the traditional NXX values because it makes sense to evaluate edge lengths using a kmer-based method.

In terms of read paths and pairs status, as the assembly progresses we should generally see an improvement in the number of ends that map to a unique location and the number of pairs with both ends mapped and satisfied. These are the more indicative metrics, and we recommend following them throughout all the steps described below. A more detailed and up-to-date explanation on how to use the metrics to guide the assembly will always be kept in the w2rap tutorial.

### 3.2.2 60-mer graph construction

The first three steps on the w2rap-contigger transform the reads into its binary format (step 1), produce a 60-mer count with neighbouring information (step 2) and construct a 60-mer graph (step 3). The main parameter to adjust in these steps is the minimum frequency of 60-mers. This can be adjusted on step 2, and then can be re-adjusted to higher values on step 3 if needed on high coverage datasets where errors are over-abundant. The 60-mer spectrum is written to the small\_K.freqs file, and we recommend choosing values of minimum frequency smaller than the first valley, which separates the bulk of the error distribution from the bulk of the genome's true 60-mers. At this step it is also advisable to check the fragment size distribution, saved in small\_K.frag.dist and control that small improvements on N<sub>kXX</sub> are not achieved by loosing placement for too many reads.

### 3.2.3 Optimising large\_K value

At this point, a collection of all possible paths through the graph are generated, effectively transforming the 60-mer graph into an exploded large-K graph that contains all possible large kmer paths. This graph is then evaluated for support, pruning paths that do not have support from the original reads. The main parameter for steps 4 and 5 is the kmer size for the second (and final) graph. This value again represents a trade-off between connectivity and noise. Increasing this parameter will disentangle repeats in the kmer spectrum, but will also generate more erroneous paths, thus decreasing the amount of reads mapped to the true paths and eventually making them discontiguous due to lack of support. A secondary parameter, representing the number of supporting reads required for a path to be considered valid is also available.

### 3.2.4 Optimising local assembly

At this point, a key requirement is to maximise the graph connectivity so the desired assembly can be found as a path through the graph. As previous steps may have incorrectly pruned true paths, a local heuristic to reconstruct them is now used. First, clusters of unsatisfied read pairs are generated: these are reads that connect a set of "left" and a set of "right" edges in the graph, which are not currently connected through existing paths. All left edges need to be connected between them and all right edges need to be connected between them. The set of these "bridging" reads is assembled using the local assembly methods described in the DISCOVAR *denovo* publication. A selection of possible paths through the previously unrepresented region is created. A new sequence graph is computed by including all the edges from the previous graph and the edges from the local assembly heuristics. This graph is the basis of the final assembly.

The number of reads in each cluster is sub-sampled to a fixed value. This is reasonable both because it places an upper-bound on computational resources, and because the reads are only spanning a region between two ends of a pair, which means a region of less than 1Kbp. The number of read pairs is defaulted to 200, but this can be adjusted with the pair.sample parameter. While increasing this parameter may in some cases improve the assembly, it is computationally expensive.

### 3.2.5 Optimising repeat resolution

This step cleans all artificial paths and edges in the graph and attempts to resolve repeats by using read-mapping information through the edges. There are 2 different heuristics for repeat resolution which can be used in different combinations and with different parameters;

- PullAparter: inherited from the original DISCOVAR *denovo* heuristics, expands edges with 2 neighbours on each side, where there is read evidence to separate the two instances of the repeat. This method has been optimised to run faster on complex genomes but the simple heuristics remain the same.
- PathFinder: expands loops using a combination of read support and coverage-based heuristics, then looks for single-flow repeat regions. These are regions where a number of small, complex connected paths, flow in a single direction from a set of N inputs into a set of N outputs, and where all these inputs and outputs are assumed to be unique sequence on the assembly. Reads mapping from the input edges to the output edges is then evaluated to score in-out combinations. If an in-out 1-to-1 combination with correct support is found, and it is possible to reconstruct the path through the single-flow region based on read mappings, then the whole region is expanded and solved for each 1-to-1 pairing (see Supplementary figure).

The dv-like mode of the contigger runs just the PullAparter, like the original DISCOVAR. The default mode runs the PullAparter (to solve the easier cases) and then the PathFinder.

### 3.2.6 Parameterisation

It is important to note that modifying parameters at each stage of the w2rap-contigger may significantly affect runtime. If the first 60-mer graph produces a clean, highly resolved assembly, many of the complex heuristics for cleanup and specially for local assembly and repeat resolution won't need to be used to analyse the bulk of the graph. This again highlights the value of being able to run the algorithms in a consistent manner with metrics that show when the parameters are being set correctly.

## 3.3 Scaffolding

### 3.3.1 Preparing Nextera LMP reads

The pipeline for processing Nextera LMP reads is based on Nextclip (Leggett et al., 2013) and is designed to recover correctly generated LMP reads from raw sequencing reads. The Nextera protocol generates circularised constructs containing the Nextera adapter at the junction between the two reads. However, the adapter doesn't necessarily occur at the end of each read. To account for this, reads are first combined into a single sequence where they overlap using FLASH (Magoc and Salzberg, 2011) to generate a single sequence which should contain the Nextera adapter. This sequence is then reverse complemented to generate the other read in the pair. Nextclip (Leggett et al., 2013) is used to classify reads according to whether the adapter is found and where in the read-pair it is located. This whole process is encapsulated within a single script provided as part of the pipeline.

The K-mer Analysis Toolkit can be used to compare the LMP reads to the PE reads to highlight sequence representation issues. A subset of the reads are then mapped to the previously assembled contigs to check the fragment size distribution.

### 3.3.2 Optimising LMP mapping and scaffolding

A bundled version of SOAPdenovo is used for scaffolding. The main parameters at this stage are the kmer size used to map the reads and the read support to call a link, as in the original SOAP scaffolder. A kmer size of 71 is reasonable in most cases of complex genomes, but checking read placement stats and link status on the output files is recommended. Finally, s\_scaff is used to generate scaffolds. The output file generated at this stage gives useful information about how the scaffolding performed.

### 3.3.3 Recovering gaps and creating releases

Before scaffolding, SOAPdenovo converts gaps in contigs (Ns) to Cs and Gs so these are converted back to Ns by mapping the contigs back to scaffolds using the output files from SOAPdenovo. A script is provided for this as outlined in the tutorial.

When deciding on a length cut-off for scaffolds in a final release, the K-mer Analysis Toolkit can be used to make sure this doesn't result in a loss of content. The scaffolds should also be checked for contamination (such as phiX) and Illumina adapters.

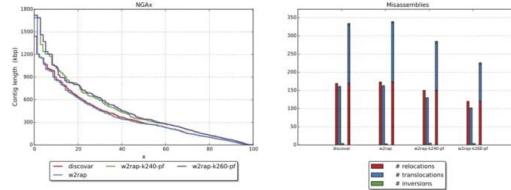


Figure 2: Correct parametrisation improves contiguity and accuracy. Contigs generated using K=260 and the PathFinder heuristic show increased aligned contiguity when evaluated against the TAIR10 reference. At the same time the correct parametrisation shows a decrease in misassemblies, which indicates improved performance from the algorithms instead of just contiguity gains by joining less-supported links.

## 4 Conclusion

Our assembly method to construct contigs and scaffolds from short read Illumina data produces high-quality assemblies in a cost-effective way, unlocking information in complex genomes. The focus on metrics enables complete tracking of how the datasets and algorithms are performing, which becomes particularly important for comparative studies where multiple similar genomes are assembled from equivalent datasets. By using w2rap, performance of the whole assembly process can be tracked to ensure reproducibility.

We have shown the effectiveness of combining the w2rap-contigger's short-range accuracy, based on the DISCOVAR heuristics originally designed to preserve variation on the assembly datasets, with a quality focused long mate paired sequencing method and the simple but proven heuristics of SOAPdenovo2's scaffolding modules. While more expensive or specific approaches could produce particular one-off results outmatching w2rap's performance, these assemblies are a good starting point for many comparative genomics projects where robustness, accuracy and price are the most important factors to consider. As

shown by its performance on the *H. sapiens* dataset, w2rap also scales well with complexity, and is already in use for even more complex genomes including the highly complex, 17Gbp genome of hexaploid wheat.

## 5 Author contributions

BJC designed the assembly approach and the pipeline. BJC and GGA programmed the w2rap-contigger. JW, GGA and BJC programmed the scripts for the w2rap pipeline. DH tweaked sequencing protocols and produced test datasets. BJC, LY and GGA tuned and optimised software, compilation chains and architecture. BJC, GGA, JW and KB tested assembly results and evaluated the pipeline. FDP provided initial collaboration with the Broad Institute and feedback throughout the project. BJC and FDP wrote the manuscript, with contributions from all authors.

## 6 Acknowledgments

Thanks to David Jaffe and Neil Weisenfeld for their support on reusing the DISCOVAR codebase. Thanks to the EI PP team for continuous efforts on producing great data, and many helpful discussions and feedback. Thanks to all members of the BBSRC Wheat LOLA for continuous feedback and support.

This work was strategically funded by the BBSRC, Institute Strategic Programme Grant BB/J004669/1. Work on wheat assembly was funded by BBSRC strategic LOLA Award BB/J003743/1. This research was supported in part by the NBIP Computing infrastructure for Science (CiS) group.

## References

Gnerre,S. et al. (2010) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, 108, 1513–1518.

Weisenfeld,N.I. et al. (2014) Comprehensive variation discovery in single human genomes.. *Nat Genet*, 46, 1350–5.

Love,R.R. et al. (2016) Evaluation of DISCOVAR de novo using a mosquito sample for cost-effective short-read genome assembly. *BMC Genomics*, 17.

Luo,R. et al. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1.

Clavijo,B.J. et al. (2016) An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations..

Salzberg,S.L. et al. (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, 22, 557–567.

Genome in a bottle: A human DNA standard (2015) *Nature Biotechnology*, 33, 675–675.

Gurevich,A. et al. (2013) QUAST: quality assessment tool for genome assemblies.. *Bioinformatics*, 29, 1072–1075.

Gnerre,S. et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data.. *Proc Natl Acad Sci U S A*, 108, 1513–8.

Heavens,D. et al. (2015) A method to simultaneously construct up to 12 differently sized Illumina Nextera long mate pair libraries with reduced DNA input, time, and cost.. *Biotechniques*, 59, 42–5.

Zook,J.M. et al. (2016) Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, 3, 160025.

Jackman,S.D. et al. (2016) ABySS 2.0: Resource-Efficient Assembly of Large Genomes using a Bloom Filter. *bioRxiv*.

Huptas,C. et al. (2016) Optimized Illumina PCR-free library preparation for bacterial whole genome sequencing and analysis of factors influencing de novo assembly.. *BMC Res Notes*, 9, 269.

Leggett,R.M. et al. (2013) NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics*, 30, 566–568.

Magoc,T. and Salzberg,S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27, 2957–2963.

# An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations

Bernardo J. Clavijo,<sup>1,9</sup> Luca Venturini,<sup>1,9</sup> Christian Schudoma,<sup>1</sup> Gonzalo Garcia Accinelli,<sup>1</sup> Gemy Kaithakottil,<sup>1</sup> Jonathan Wright,<sup>1</sup> Philippa Borrill,<sup>2</sup> George Kettleborough,<sup>1</sup> Darren Heavens,<sup>1</sup> Helen Chapman,<sup>1</sup> James Lipscombe,<sup>1</sup> Tom Barker,<sup>1</sup> Fu-Hao Lu,<sup>2</sup> Neil McKenzie,<sup>2</sup> Dina Raats,<sup>1</sup> Ricardo H. Ramirez-Gonzalez,<sup>1,2</sup> Aurore Coince,<sup>1</sup> Ned Peel,<sup>1</sup> Lawrence Percival-Alwyn,<sup>1</sup> Owen Duncan,<sup>3</sup> Josua Trösch,<sup>3</sup> Guotai Yu,<sup>2</sup> Dan M. Bolser,<sup>4</sup> Guy Namaati,<sup>4</sup> Arnaud Kerhornou,<sup>4</sup> Manuel Spannagl,<sup>5</sup> Heidrun Gundlach,<sup>5</sup> Georg Haberer,<sup>5</sup> Robert P. Davey,<sup>1,6</sup> Christine Fosker,<sup>1</sup> Federica Di Palma,<sup>1,6</sup> Andrew L. Phillips,<sup>7</sup> A. Harvey Millar,<sup>3</sup> Paul J. Kersey,<sup>4</sup> Cristobal Uauy,<sup>2</sup> Ksenia V. Krasileva,<sup>1,6,8</sup> David Swarbreck,<sup>1,6</sup> Michael W. Bevan,<sup>2</sup> and Matthew D. Clark<sup>1,6</sup>

<sup>1</sup>Earlham Institute, Norwich, NR4 7UZ, United Kingdom; <sup>2</sup>John Innes Centre, Norwich, NR4 7UH, United Kingdom; <sup>3</sup>ARC Centre of Excellence in Plant Energy Biology, The University of Western Australia, Crawley Western Australia 6009, Australia; <sup>4</sup>EMBL European Bioinformatics Institute, Hinxton, CB10 1SD, United Kingdom; <sup>5</sup>Plant Genome and Systems Biology, Helmholtz Center Munich, 85764 Neuherberg, Germany; <sup>6</sup>University of East Anglia, Norwich, NR4 7TJ, United Kingdom; <sup>7</sup>Rothamsted Research, Harpenden, AL5 2JQ, United Kingdom; <sup>8</sup>The Sainsbury Laboratory, Norwich, NR4 7UH, United Kingdom

Advances in genome sequencing and assembly technologies are generating many high-quality genome sequences, but assemblies of large, repeat-rich polyploid genomes, such as that of bread wheat, remain fragmented and incomplete. We have generated a new wheat whole-genome shotgun sequence assembly using a combination of optimized data types and an assembly algorithm designed to deal with large and complex genomes. The new assembly represents >78% of the genome with a scaffold N50 of 88.8 kb that has a high fidelity to the input data. Our new annotation combines strand-specific Illumina RNA-seq and Pacific Biosciences (PacBio) full-length cDNAs to identify 104,091 high-confidence protein-coding genes and 10,156 noncoding RNA genes. We confirmed three known and identified one novel genome rearrangements. Our approach enables the rapid and scalable assembly of wheat genomes, the identification of structural variants, and the definition of complete gene models, all powerful resources for trait analysis and breeding of this key global crop.

[Supplemental material is available for this article.]

Improvements in sequencing read lengths and throughput have enabled the rapid and cost-effective assembly of many large and complex genomes (Gnerre et al. 2011; Lam et al. 2011). Comparisons between assembled genomes have revealed many classes of sequence variation of major functional significance that were not detected by direct alignment of sequence reads to a common reference (The 1000 Genomes Project Consortium 2010; Gan et al. 2011; Bishara et al. 2015). Therefore, accurate comparative genomics requires that genome sequences are assembled

prior to alignment, but in many eukaryotic genomes, assembly is complicated by the presence of large tracts of repetitive sequences (Treangen and Salzberg 2012; Chaisson et al. 2015) and the common occurrence of genome duplications, for example, in polyploids (Blazquez and Wolfe 2004; Berthelot et al. 2014).

Recent innovations in sequence library preparation, assembly algorithms, and long-range scaffolding have dramatically improved whole-genome shotgun assemblies from short-read sequences. These include PCR-free library preparation to reduce bias (Aird et al. 2011), longer sequence reads, and algorithms that preserve allelic diversity during assembly (Weisenfeld et al. 2014). Short-read assemblies have been linked into larger

<sup>9</sup>These authors contributed equally to this work.

Corresponding authors: matt.clark@earlham.ac.uk, David.Swarbreck@earlham.ac.uk, michael.bevan@jic.ac.uk

Article published online before print. Article, supplemental material, and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.217117.116>. Freely available online through the *Genome Research* Open Access option.

© 2017 Clavijo et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

chromosome-scale scaffolds by Hi-C in vivo (Lieberman-Aiden et al. 2009) and in vitro (Putnam et al. 2016) chromatin proximity ligation, as well as by linked-read sequencing technologies (Mostovoy et al. 2016; Weisenfeld et al. 2016). Although it is more expensive than short-read sequencing approaches, single-molecule real-time (SMRT) sequencing improved the contiguity and repeat representation of mammalian (Pendleton et al. 2015; Gordon et al. 2016; Bickhart et al. 2017) and diploid grass genomes (Zimin et al. 2017). SMRT technologies are also being used to generate the complete sequence of transcripts, increasing the accuracy of splicing isoform definition (Abdel-Ghany et al. 2016).

The assembly of the 17Gb allohexaploid genome of bread wheat (*Triticum aestivum*) has posed major difficulties, as it is composed of three large, repetitive, and closely related genomes (Moore et al. 1995). Despite progressive improvements, an accurate and near-complete wheat genome sequence assembly and corresponding high-quality gene annotation has not yet been generated. Initial whole-genome sequencing used orthologous Poaceae protein sequences to generate highly fragmented gene assemblies (Brenchley et al. 2012). A BAC-based assembly of Chromosome 3B provided major insights into wheat chromosome organization (Choulet et al. 2014). Illumina sequencing and assembly of flow-sorted chromosome arm DNA (Chromosome Survey Sequencing [CSS]) identified homoeologous relationships between genes in the three genomes, but the assemblies remained highly fragmented (The International Wheat Genome Sequencing Consortium 2014). Recently, a whole-genome shotgun sequence of hexaploid wheat was assembled and anchored, though not annotated, using an ultradense genetic map (Chapman et al. 2015). The assembly contained ~48.2% of the genome with contig and scaffold N50 lengths of 8.3 and 25 kb, respectively.

Here we report the most complete and accurate sequence assembly and annotation to date of the allohexaploid wheat reference accession, Chinese Spring (CS42). Our approach is open source, rapid, and scalable and enables a more in-depth analysis of sequence and structural variation in this key global crop.

## Results

### DNA library preparation and sequencing

We aimed to reduce bias and retain maximum sequence complexity by using unamplified libraries for contig generation (Kozarewa et al. 2009) and to improve scaffolding by using precisely sized mate-pair libraries (Heavens et al. 2015). Libraries were sequenced using Illumina paired-end (PE) 250-bp reads to distinguish closely related sequences. In total, 1.1 billion PE reads were generated to provide 33 $\times$  sequence coverage of the CS42 genome (Supplemental Information S1; Supplemental Table S4.1). For scaffolding, long mate-pair (LMP) libraries with insert sizes ranging from 2480–11,600 bp provided 53 $\times$  sequence coverage, and Tight, Amplification-free, Large Insert PE Libraries (TALL) with an insert size of 690 bp provided 15 $\times$  sequence coverage (Supplemental Information S1; Supplemental Table S4.2).

### Genome assembly

Nearly 3 million contigs (of length >500 bp) were generated using the w2rap-contigger (Clavijo et al. 2017) with an N50 of 16.7 kb (Supplemental Information S1; Supplemental Table S4.3). After scaffolding using SOAPdenovo (Luo et al. 2012), the assembly contained 1.3 million sequences with an N50 of 83.9 kbp. The TGACv1 scaffolds were classified to chromosome arms using

raw CSS reads (The International Wheat Genome Sequencing Consortium 2014) and subsequently screened with a two-tiered filter based first on their length and their  $k$ -mer content (see Supplemental Information S1, section S4.5). The approach removed short, redundant sequences from the assembly minimizing the loss of unique sequence content, leading to an increase in scaffold N50 to 88.8 kb. Contig accuracy was assessed by mapping links from the 11-kb LMP library, which was not used in the contig assembly. Breaks in the linkage at different mate-pair mapping coverages only affected a very small portion of the content and did not reduce N50 contiguity significantly (Supplemental Information S1; Supplemental Figs. S4.4, S4.5). Supplemental Tables S4.5 and S4.6 in S1 show that 91.1% of TGACv1 genes were correctly assigned to Chromosome 3B, with no discrepancies in gene order identified.

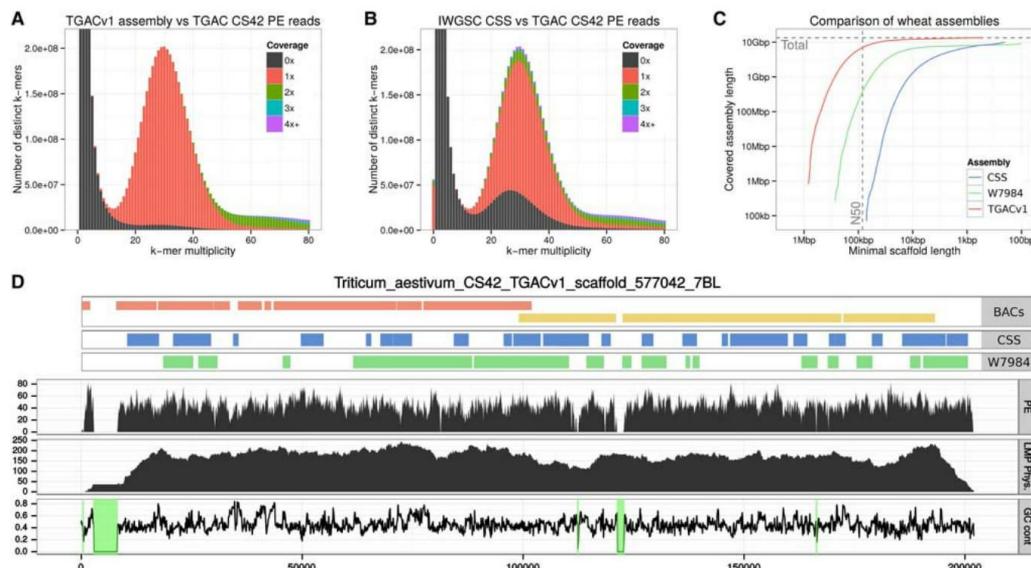
The genome of a synthetic wheat line W7984 was previously assembled with an improved version of meraculous (Chapman et al. 2011) using 150-bp PE libraries with varying insert sizes, for a combined genome coverage of 34.3 $\times$ , together with 1.5- and 4-kb LMP libraries for scaffolding (Chapman et al. 2015). This contig assembly, with an N50 of 8.3 kb, covered 8 Gb of the genome while the scaffold assembly covered 8.21 Gb with an N50 of 24.8 kb. In comparison, the TGACv1 assembly represents ~80% of the 17-Gb genome, a 60% improvement in genome coverage. The contiguity of the TGACv1 assembly, as measured by scaffold N50 values, is 3.7-fold greater than that of the W7984 assembly and 30 times that of the CSS assembly (Table 1; The International Wheat Genome Sequencing Consortium 2014).

A KAT  $k$ -mer spectra copy number plot provides information to analyze how much and what type of  $k$ -mer content from reads is present in an assembly (Mapleson et al. 2017). It decomposes the  $k$ -mer spectrum of a read data set by the frequency in which the  $k$ -mers are encountered in the assembly. The plot generated from TGACv1 (Fig. 1A) showed that  $k$ -mers found at low frequency (less than 12), representing sequencing errors, were not found in the assembly (shown by the black distribution at  $k$ -mer multiplicity less than 12). Most sequence content was represented in the assembly once (shown by the main red distribution), with  $k$ -mers originating from the repetitive and the homoeologous regions of the genome represented at higher frequencies (more than 50). The absence of  $k$ -mers in the assembly that are not present in the reads indicated that the assembled contigs accurately reflected the input data. A similar analysis of the CSS assembly (Fig. 1B) identified approximately 50 million  $k$ -mers that were not found as sequenced content in the PCR-free paired-end data, as shown by the red bar at  $k$ -mer multiplicity equal to zero. This is indicative of chimeric sequences or consensus inconsistencies in the CSS assembly. The black distribution between  $k$ -mer multiplicity 15 and 45 shows  $k$ -mers from the PCR-free reads that were not present in

**Table 1.** Comparison of TGACv1 scaffolds to the IWGSC and Chapman assemblies of hexaploid wheat

	Size (Gb)	Seq. count	N20 (kb)	N50 (kb)	N80 (kb)	% Ns	% of genome
TGACv1	13.43	735,943	180.1	88.8	32.8	5.7	78.8
W7984	8.21	955,122	47.1	24.8	9.9	15.2	48.2
CSS	8.32	4,061,833	8.6	3.3	1.2	1.0	48.9

Numbers are calculated using sequences >500 bp and including gaps (Ns) for each assembly. (IWGSC) International Wheat Genome Sequencing Consortium.



**Figure 1.** Summary of the TGACv1 wheat genome sequence assembly. (A, B) KAT spectra-cn plots comparing the PE reads to the TGACv1 scaffolds (A) and CSS scaffolds (B). Plots are colored to show how many times fixed length words (k-mers) from the reads appear in the assembly; frequency of occurrence (multiplicity; x-axis) and number of distinct k-mers (y-axis). Black represents k-mers missing from the assembly; red, k-mers that appear once in the assembly; green, twice; etc. Plots were generated using  $k=31$ . The black distribution between k-mer multiplicity 15 and 45 in B represents k-mers that do not appear in the CSS assembly. (C) Comparison of scaffold lengths and total assembly sizes of the TGACv1, W7984, and CSS assemblies. (D) Scaffold 577042 of the TGACv1 assembly. Tracks from top to bottom: aligned BAC contigs, CSS contigs, W7984 contigs, coverage of PE reads, coverage of LMP fragments, and GC content with scaffolded gaps (N stretches) with 0% GC highlighted in green. There are two BACs (composed of seven and four contigs each), 22 CSS contigs, and 15 W7984 contigs across the single TGACv1 scaffold.

the CSS assembly, most probably coming from the one-third of the genome not represented by the CSS assembly. The PCR-free library is expected to capture unbiased coverage of the genome, which is reflected in the increased size of the TGACv1 assembly compared with the CSS assembly. Greater amounts of duplication were observed in the single copy regions of the CSS assembly, corresponding to the purple and green areas above the main red distribution.

The content and order of genes in TGACv1 scaffolds assigned to Chromosome 3B (Supplemental Information S1; Supplemental Table S4.4) was compared to that in the Chromosome 3B BAC-based assembly (Choulet et al. 2014); 91.2% of the genes previously identified on the 3B BAC-based assembly aligned to TGACv1 scaffolds (Table 2), with no discrepancies in gene order (Supplemental Information S1; Supplemental Table S4.5). This compared with 73.9% aligned to W7984 3B scaffolds and 68.0% aligned to CSS Chromosome 3B scaffolds, demonstrating the improved representation of the TGACv1 assembly.

Alignment of TGACv1 3B scaffolds to the 3B BAC-based pseudomolecule (Fig. 2A,C) showed that they were largely in agreement. Two examples of apparent disagreement are shown in Figure 2, B and D. Scaffold\_221671\_3B spanned a gap of 700 kb in the 3B BAC assembly, and reoriented and removed a duplication, by identifying both ends of a CACTA element (Fig. 2B). Scaffold\_220592\_3B spanned 582 kb and diverged in one location (Fig. 2D) and contained a Sabrina solo-LTR with a characteristic ATCAG target site duplication (TSD). In scaffold\_220592\_3B, the TSD was present on either side of the Sabrina\_3231 element, while

in the BAC-based scaffold Sabrina homology ended in Ns. In the BAC-based assembly, only one side of the disjunction showed alignment similarity to CACTA\_3026, which was found to be complete in scaffold\_220592\_3B and spanned the disjunction (Fig. 2D). These two examples illustrate how the TGACv1 assembly generated accurate scaffolds spanning typical complex and long tracts of repetitive DNA characterizing the wheat genome, which were misassembled in the BAC-based approach.

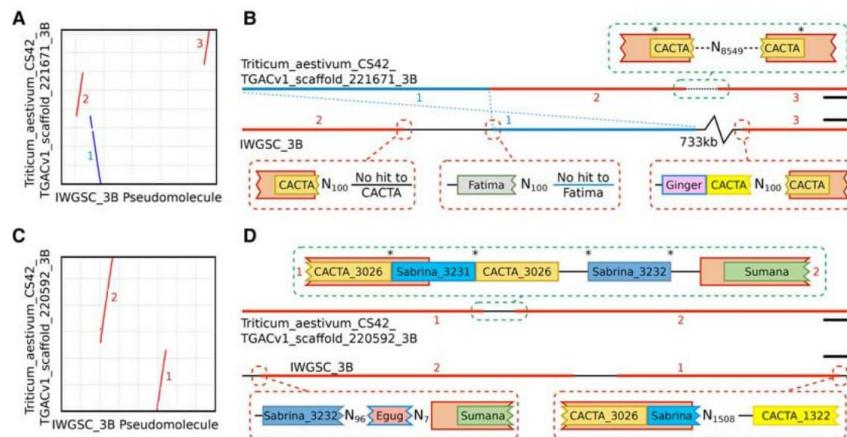
#### Repetitive DNA composition

More than 80% of the 13.4-Gb assembly was composed of approximately 9.7 million annotated transposable element entities, of which ~70% were retroelements (class I) and 13% DNA

**Table 2.** Comparison of TGACv1 Chromosome 3B scaffolds to BAC-based scaffolds (Choulet et al. 2014) and 3B scaffolds from the W7984 and CSS assemblies

	Scaffold count	N50 (kb)	Total seq. (Mb)	Gene count	% genes
3B ref.	2808	892.4	832.8	7703	100.0
TGACv1	29,090	116.5	790.0	6983	91.2
W7984	26,206	30.6	479.4	5671	73.9
CSS	272,072	3.4	557.2	5233	68.0

Numbers are calculated using sequences >500 bp and including gaps (Ns) for each assembly.



**Figure 2.** Comparative alignment of TGACv1 scaffolds with the 3B BAC-based pseudomolecule. (A,C) Dot plots between TGACv1 scaffolds and 3B show disruptions in sequence alignment, including rearrangements (red) and inversions (blue). (B,D) Graphical representation of sequence annotations in disrupted regions. Junctions in the TGACv1 scaffolds are consistent with a complete retroelement spanning the junction that includes identical TSD on either side of the retroelement (asterisks). Corresponding regions in the 3B BAC-based pseudomolecule are characterized by Ns that produce inconsistent alignment of retroelements across putative junctions. Retroelements of the same family (CACTA, Sabrina) but matching distinct members in the TREP database are indicated by different colors. Numbers adjacent to sequences correspond to regions shown in panel A and C, respectively. (B) Scale bars, 10 kbp; (D) scale bars, 30 kbp.

transposons (class II) (Supplemental Information S1; Supplemental Table S7.1). Among the class I elements, Gypsy and Copia LTR retroelements comprised the major component of the repeats, while CACTA DNA elements were highly predominant among class II DNA repeat types. No major differences in the repeat composition of the three genomes were apparent. Compared with *Brachypodium distachyon*, which has a related but much smaller genome (Vogel et al. 2010), there has been a greater than 100× increase in repeat content, driven by both class I and class II expansion. The preponderance of CACTA DNA elements in the wheat genome emerged during this massive expansion.

#### Gene prediction and annotation

A total of 217,907 loci and 273,739 transcripts were identified from a combination of cross-species protein alignments, 1.5 million high-quality long Pacific Biosciences (PacBio) cDNA reads,

and over 3.2 billion RNA-seq read pairs covering a range of tissues and developmental stages (Table 3; Supplemental Information S8).

Loci were identified as coding, long ncRNA, or repeat associated and were classified as high (HC) or low (LC) confidence based on similarity to known plant protein sequences and supporting evidence from wheat transcripts (Supplemental Information S8.5.5). We assigned 104,091 coding genes (154,798 transcripts) as HC, of which 95,827 spanned at least 80% of the length of the best identified homolog (termed protein rank 1, P1, in the annotation) (Supplemental Fig. S8.1; Supplemental Information S8.5.1). The HC protein-coding set contained 51,851 genes confirmed by a PacBio transcript (Transcript rank 1, T1) and an additional 29,996 genes fully supported by assembled RNA-seq data (T2), providing full transcriptome support for 81,847 (78.63%) HC genes. Gene predictions were assessed by identifying 2707 single copy genes common to *B. distachyon*, *Oryza sativa*, *Sorghum bicolor*, *Setaria italica*, and *Zea mays*. A single orthologous wheat gene was identified for 2686 (99.22%) of these, with 2665 (98.45%)

**Table 3.** Characteristics of predicted high (HC) and low (LC) confidence wheat genes including coding (mRNA) and long noncoding (ncRNA) RNA

	All TGAC Models	mRNA HC	mRNA LC	ncRNA HC	ncRNA LC	Repeat-associated
Genes	217,907	104,091	83,217	10,156	9933	10,510
Transcripts	273,739	154,798	85,778	11,591	10,438	11,134
Transcripts per gene	1.26	1.49	1.03	1.14	1.05	1.06
Transcript mean cDNA size (bp)	1766.12	2119.52	1304.53	1368.24	1083.98	1462.71
Exons per transcript	4.48	5.83	2.8	2.58	2.76	2.27
Exons mean size (bp)	394.15	363.73	465.27	530.25	392.24	644.09
Transcript mean CDS size (bp)	1,165.52	1,361.82	839.97	—	—	891.05
Mono-exonic transcripts	60,322	19,034	30,479	3061	3044	4704
22.04%	12.30%	35.53%	26.41%	29.16%	42.25%	
Genes with alternative splicing	32,616	28,608	2033	1037	460	478
14.97%	27.48%	2.44%	10.21%	4.63%	4.55%	

classified as HC and 21 (0.78%) in the LC set. A high coherence in gene length ( $r=0.969$ ) was found between wheat and *B. distachyon* proteins (Supplemental Fig. S8.2). These findings show that the HC gene set is robust and establishes a lower bound estimate for the total number of protein-coding genes in wheat. An additional 103,660 loci were defined as LC (i.e., gene models with all their transcripts either having <60% protein coverage or lacking wheat transcript support). These include bona fide genes that were fragmented due to breaks in the current assembly, wheat-specific genes, and genes without transcriptome support (Supplemental Table S8.8).

We also identified 10,156 HC noncoding genes with little similarity in protein databases and low protein-coding potential. The majority of these genes are located in intergenic regions (8854, or 87.18%), while most of the remaining 1302 are anti-sense to coding genes (1082, or 10.65%) (see Supplemental Information, section 8.5.8); 5413 of wheat noncoding genes (53.30%) were detected in at least one of the two sequenced wheat diploid progenitor species *Triticum urartu* and *Aegilops tauschii* (at least 90% coverage and 90% identity) (see Supplemental Information S8.5.8).

To obtain additional support for gene predictions, a proteome map was constructed from 27 wheat tissues (Supplemental Information S9). This identified 2,106,323 significant peptide spectrum matches corresponding to 102,379 distinct peptides. Of these, 96.20% matched HC genes, while 13.29% were assigned to LC genes. For 56,391 genes (43,431 HC, 12,960 LC), we were able to identify at least one peptide confirming the predicted coding sequence. Due to the hexaploid nature of wheat, only 22.1% of the peptides could be assigned to a single gene. Applying progressively stricter filters, by requiring at least two or five peptides, confirmed the protein sequence of 30,607 and 17,316 HC genes, respectively; 10,819 genes met the criteria of having support from multiple peptides with at least one uniquely identifying peptide and were considered as unambiguously corroborated by proteomic data. Among the LC genes, only 368 were identified by two or more peptides that did not match any HC gene, further supporting confidence assignments. Among these, 343 were classified as LC due to having <60% the length of the identified homolog, while the remaining 25 genes were classified as LC due to either repeat association or lack of wheat transcript support.

We compared the TGACv1, CSS (The International Wheat Genome Sequencing Consortium 2014), and Chromosome 3B (Choulet et al. 2014) gene models. Of the 100,344 HC genes in the International Wheat Genome Sequencing Consortium (IWGSC) annotation (PGSB/MIPS version 2.2 and INRA version 1.0 from Ensembl release 29), we were able to transfer 97,072 (97%) to the TGACv1 assembly with stringent alignment parameters (at least 90% coverage and 95% identity). Fewer (72%) of the IWGSC (The International Wheat Genome Sequencing Consortium 2014) LC, unsupported, repeat associated, and noncoding loci could be aligned (at least 90% coverage and 95% identity), likely reflecting differences between the assemblies of repeat rich and difficult to assemble regions. Of the TGACv1 HC genes, 61% overlapped with an aligned IWGSC HC gene and 78% to the full IWGSC gene set (Supplemental Information S8.5.7). Less agreement was found between TGACv1 LC and ncRNA genes and the IWGSC annotation, with only 8% overlapping IWGSC HC loci and 40% overlapping the full IWGSC gene set (Fig. 3A). Of the 22,904 (22%) HC TGACv1 genes not overlapping a transferred IWGSC gene, 19,810 (86%) had cross-species protein similarity support with 6665 (29%) fully supported by a PacBio transcript (Fig. 3B). We identified 13,609 TGACv1 genes that were over-

lapped by transcripts originating from two or more IWGSC genes in our annotation, indicating that they were likely fragmented in the CSS assembly. In 8175 of these cases (60%), we were able to find a PacBio read fully supporting our gene model. These differences reflect improvements in contiguity, a more comprehensive representation of the wheat gene space in our assembly, and improved transcriptome support for annotation.

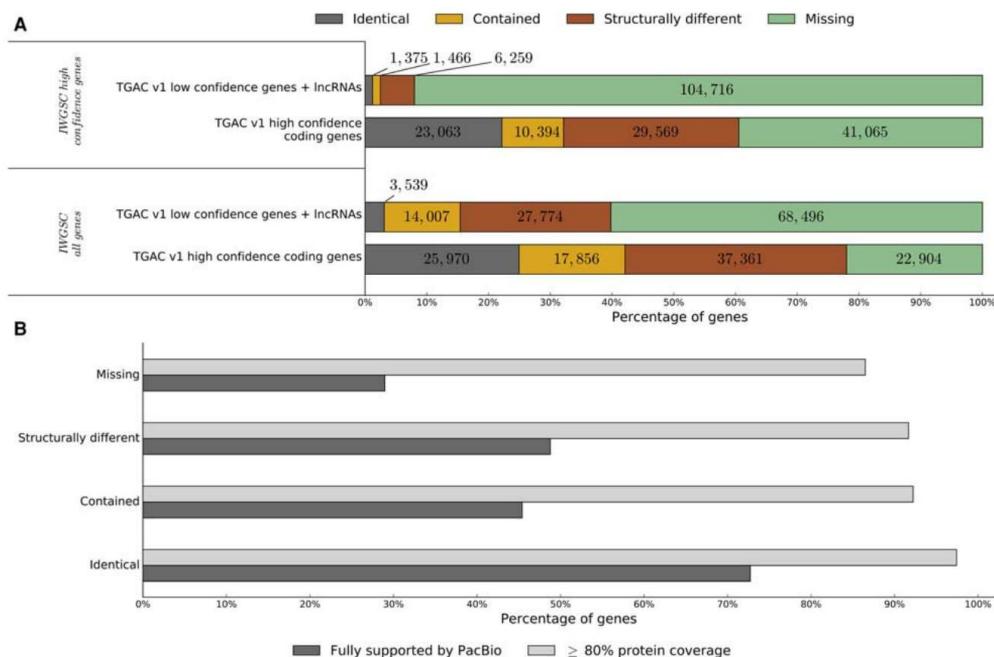
### Alternative splicing

Alternative splicing is an important mRNA processing step that increases transcriptome plasticity and proteome diversity (Staiger and Brown 2013). The TGACv1 annotation includes high-quality alternative splicing variants identified from PacBio transcriptome reads. To provide a more comprehensive representation of alternative splicing, we subsequently integrated transcript assemblies generated from six strand-specific Illumina libraries (Supplemental Information S8.6; Supplemental Table S8.1). This added a further 121,997 transcripts, increasing the number of genes with splice variants from 15% in the TGACv1 annotation to 31% in the supplemented set of transcripts (i.e., incorporating Illumina RNA-seq assemblies), as well as increasing the average number of transcripts per gene from 1.26 to 1.88. When considering only HC genes, the number of alternatively spliced genes was increased from 27.48% to 48.80% (2.36 transcripts per gene), similar to that observed in a wide range of plant species (Zhang et al. 2015).

Intron retention (IR) was the prevalent alternative splicing event in wheat (34%) followed by alternative 3' splice sites (A3SS; 27%), exon skipping (ES; 20%), alternative 5' splice sites (A5SS; 19%), and mutually exclusive exons (MXE; 0.04%). This was similar to previous analyses of Chromosome 3B (Pingault et al. 2015), and IR is also predominant in barley (Panahi et al. 2015). Alternative splicing coupled to nonsense mediated decay (NMD) regulates gene expression (Lykke-Andersen and Jensen 2015). We found 22% of all transcripts (17% of all genes) and 29% of multiexonic HC protein-coding transcripts (33% genes) may be potential targets for NMD. IR was the most common splicing event leading to NMD sensitivity, with 40% of IR transcripts identified as potential NMD targets (34% ES, 38% A5SS, 34% A3SS, 26% MXE). This suggests a potentially substantial role for alternative splicing/NMD in regulating gene expression in wheat.

### Gene families

HC and LC gene families were analyzed separately using OrthoMCL version 2.0 (Li et al. 2003; Supplemental Figs. S10.1, S10.2). Splice variants were removed from the HC gene data set, keeping the representative transcript for each gene model (see Supplemental Information S8.5.6, S10.1), and data sets were filtered for premature termination codons and incompatible reading frames. For the HC gene set, a total of 87,519 coding sequences were clustered into 25,132 gene families. The vast majority of HC gene families contained members from the A, B, and D genomes, consistent with the relatively recent common ancestry of the A and B genomes and the proposed hybrid origin of the D genome from ancestral A and B genomes (Marcussen et al. 2014). Subsets of gene families and singleton genes (those not clustered into any family) were classified to identify (1) genes and families that are A, B, or D genome specific; (2) gene families with expanded numbers in one genome; and (3) wheat gene families that are expanded relative to other species. These gene sets were analyzed for overrepresented Gene Ontology (GO) terms, shown in Supplemental File S2. Gene families that were significantly expanded in wheat



**Figure 3.** Comparison between IWGSC annotation and TGACv1 high (HC) and low confidence (LC) genes. IWGSC genes were aligned to the TGACv1 assembly (gmap,  $\geq 90\%$  coverage,  $\geq 95\%$  identity) and classified based on overlap with TGACv1 genes. (A) Identical indicates shared exon–intron structure; contained, exactly contained within the TGACv1 gene; structurally different, alternative exon–intron structure; and missing, no overlap with IWGSC. (B) Bar plot showing proportion of HC TGACv1 protein-coding genes supported by protein similarity or PacBio data. Genes are classified based on overlap with the full set of IWGSC genes.

compared with *Arabidopsis*, rice, sorghum, and *Brachypodium* include those encoding proteins involved in chromosome maintenance and reproductive processes, as well as protein and macromolecule modification and protein metabolism processes. The D genome has expanded gene families encoding phosphorylation, phosphate metabolism, and macromolecule modification activities, while the B genome has expanded gene families encoding components of chromosome organization, DNA integration and conformation/unwinding, and telomere maintenance. The B genome is derived from the *Sitopsis* section of the Triticeae, which has contributed genomes to many polyploid Triticeae species (Riley et al. 1961), suggesting B genomes may have contributed gene functions for establishing and maintaining polyploidy in the Triticeae. This is supported by the location of the major chromosome pairing *Ph1* locus on Chromosome 5B (Griffiths et al. 2006).

#### Genome organization

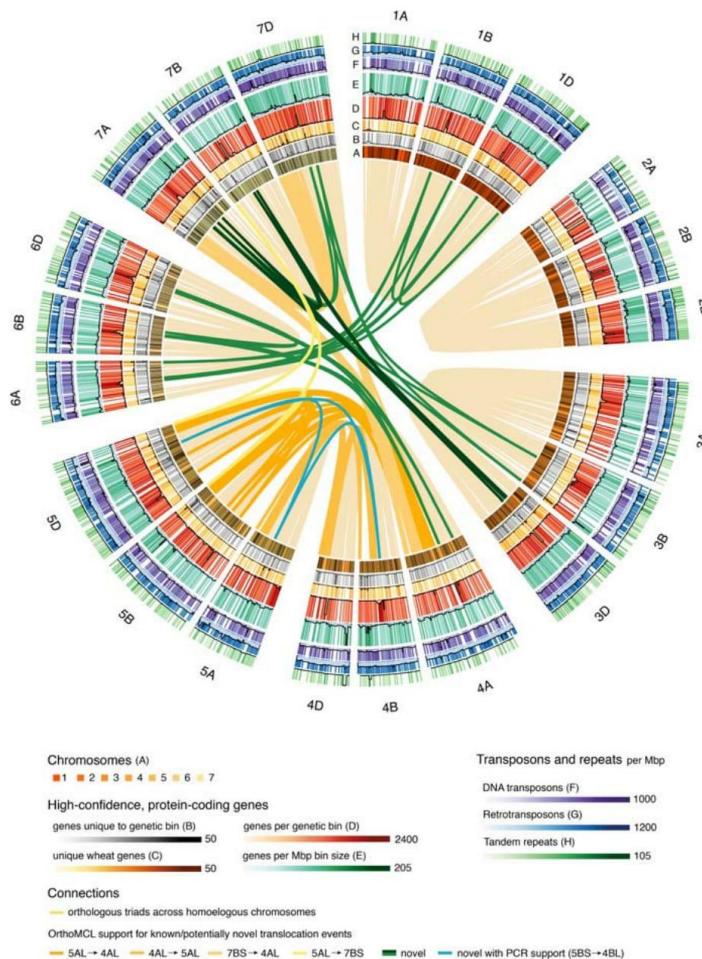
A corrected version of the POPSEQ genetic map (Chapman et al. 2015) was used to order TGACv1 scaffolds along chromosomes (Supplemental Information S5). This uniquely assigned 128,906 (17.5 %) of the 735,943 TGACv1 scaffolds to 1051 of 1187 genetic bins (class 1) (Supplemental Information S5) to form the final TGACv1 map. The total length of these scaffolds is 8,551,191,083 bp, representing 63.68% of the TGACv1 assembly

and 50.52% of the 17-Gbp wheat genome. A further 13,019 (1.77%) scaffolds were ambiguously assigned to different cM positions on the same chromosome (class 2), 489 (0.07%) scaffolds were assigned to homologous chromosomes (class 3), and 3320 (0.45%) scaffolds had matching markers with conflicting bin assignment (class 4).

The TGACv1 map also assigns unique chromosomal positions to 3927 (3.05%) scaffolds that were not previously assigned to a chromosome arm (class 5). The CSS-based chromosome arm assignments of 380 (0.295%) class 1 scaffolds and 11 (0.08%) class 2 scaffolds disagree with the map-based chromosome assignments (classes 6, 7). A list of scaffold classifications can be found in Supplemental Information S6.

The TGACv1 map encompasses 38,958 of the 53,792 scaffolds containing at least one annotated HC protein-coding gene (72.42%), comprising gene sequences of 307,085,968 bp (73.28% of total predicted gene sequence space). In total, we were able to assign genetic bins to 75,623 (72.65%) of the HC genes.

Chromosomal locations of related genes were identified by anchoring to the TGACv1 map and are displayed in Figure 4. Analysis of OrthoMCL outlier triads (Supplemental Information S1, sections S6, S10) provided genomic support for known ancestral reciprocal translocations between chromosome arms 4AL and 5AL, a combination of pericentromeric inversions between chromosome arms 4AL and 5AL, and a reciprocal exchange



**Figure 4.** Circular representation of the TGACv1 CS42 assembly. Chromosomes, genetic bins, and genomic features are visualized on the outer rings (A–H) and interchromosomal links identified known and potentially novel translocation events. (A) The seven chromosome groups of the A, B, and D genomes, scaled by number of genetic bins (black bands). (B–H) Combined heatmap/histogram representations of genomic features per genetic bin. With the exception of D, all counts are normalized by the size of the genetic bin in Mbp, calculated as the total size of all scaffolds assigned to the bin. (B) Distribution of unique genes, i.e., genes that did not have orthologs in a genome-wide OrthoMCL screen. (C) Distribution of wheat-specific genes. (D,E) Number of HC protein-coding genes. (F) Distribution of DTC, DTM, and DTH DNA transposons (Supplemental Information S1; Supplemental Table S7.1). (G) Distribution of RLX, RLC, RLG, RXX, and RIX retrotransposons. (H) Distribution of tandem duplications. Light yellow links connect homoeologous OrthoMCL triads. Dark yellow-colored links connect genetic bins harboring OrthoMCL outlier triads (Supplemental Information S1, section S6) that identify known translocation events. Dark green links connect genetic bins harboring at least three OrthoMCL outlier triads that may support novel translocation events. The cyan link shows a novel PCR-validated translocation event between Chromosomes 5BS–4BL.

between chromosome arms 4AL and 7BS (Devos et al. 1995). Several putative novel chromosomal translocations were also identified (Fig. 4; Supplemental File S3). As these may have originated in the parental lines used in the POPSEQ map rather than in CS42,

nine genes in the predicted translocations (six previously known and three novel) were tested using PCR assays on Chinese Spring chromosomal deletion stocks (Sears et al. 1966). Three known translocation events—4AL–5AL and 7BS–4AL (Devos et al. 1995) and 5AL–7BS (Ma et al. 2013)—and one previously unidentified translocation, 5BS–4BL, were validated by PCR assays.

#### Gene expression

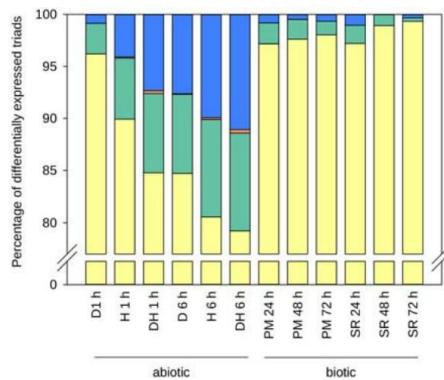
To explore global gene expression patterns, we mapped multiple wheat RNA-seq data sets to the TGACv1 transcriptome (Supplemental Information S1; Supplemental Table S11.1). Seventy-five percent of RNA-seq reads mapped to the TGACv1 transcriptome (Supplemental Information S1; Supplemental Table S11.1), and 78% of the HC protein-coding transcripts were expressed above the background level of 2 tpm (Wagner et al. 2013). Interestingly, 23% of the LC genes were also expressed above 2 tpm. Expression levels of genes across chromosomes were similar, with the exception of 19 genetic bins that had increased expression (defined as “hotspots” with a median expression level >20 tpm, containing on average 5 genes) across the six tissues examined (Supplemental Information; Supplemental Fig. S11.1). Hotspots tended to be enriched for genes encoding components of the cytoskeleton, ribosome biogenesis, and nucleosome assembly that were expressed at high levels in all tissues. Other notable hotspots were enriched in genes of photosystem I formation in leaf tissues, and nutrient reservoir activity in seed tissues.

The more complete and accurate annotation provided an opportunity to analyze patterns of transcript levels in homoeologous triads. Transcript levels of 964 triads were analyzed in response to biotic and abiotic stress using publicly available RNA-seq data sets, selected as they all used 7-d-old seedlings, were replicated, and assessed dynamic transcriptional responses to standardized treatments (Supplemental Information S1; Supplemental Table S11.2). Across treatments, 26% (2424 of 9159) of expressed triads showed higher expression in one or two genomes in at least one

stress condition (rather than balanced expression of three genomes) (see Supplemental Information S11.5). Abiotic stress led to more differentially regulated transcripts, compared with biotic stress responses, across all three genomes. To assess the

conservation of this stress response between homoeologs, we classified each homoeolog as either up-regulated (greater than twofold change, UP), down-regulated (less than 0.5-fold change, DOWN), or flat (between 0.5-fold to twofold change). We then assessed whether the individual homoeolog response to stress compared with control conditions was consistent (Supplemental Information S1; Supplemental Table S11.3). Eighty percent ( $\pm 5.1\%$  SE) of triads were not differentially expressed in response to the stress treatments and were excluded from further analysis. The most frequent pattern of differential triad expression was a single homoeolog UP or DOWN, with the other two remaining flat (79%–99% across conditions) (Fig. 5). Triads in which either all homoeologs were expressed in the same pattern ("3 UP" or "3 DOWN") were rare, as were triads in which homoeologs were expressed in opposite directions. This is consistent with Liu et al. (2015), who identified between 13% and 41% of homoeolog triads in which homoeologs did not respond to the same degree in response to stress conditions.

The genomic context of differences in homoeolog expression was explored in genomic regions containing at least five HC genes in synteny order on all three genomes, of which at least one homoeolog was expressed over background levels in root, shoot, and endosperm tissue at 10 and 20 d post anthesis (DPA; DRP000768 and ERP004505) (Supplemental Information S1; Supplemental Table S11.1; Pfeifer et al. 2014). Of the four blocks meeting these criteria, one showed equal expression of all 15 homoeologs in at least one of the tissues, while the other three blocks showed unbalanced expression of at least one homoeolog (Supplemental Information S1; Supplemental Fig. S11.2). All blocks exhibited major structural and promoter sequence differences, as well as variant transcription start sites (Supplemental Information S1; Supplemental Fig. S11.3). These multiple types of genomic differences all have the potential to contribute to unbalanced expression. To facilitate further expression studies the expression atlas at <http://www.wheat-expression.com> has been updated with the TGACv1 annotation and expression data from 424 RNA samples (Borrill et al. 2015).



**Figure 5.** Response of differentially expressed (DE) triads to stress treatments according to the number and pattern of DE homoeologs. Triads were classified as having one homoeolog DE (yellow), two homoeologs DE with same direction of change (green), three homoeologs DE with same direction of change (orange), or opposite direction of change between DE homoeologs (blue). The stresses applied were drought (D), heat (H), drought and heat combined (DH), powdery mildew (PM), and stripe rust (SR), with the duration of stress application indicated in hours (h).

## Gene families of agronomic interest

### Wheat disease-resistance genes

Plant disease-resistance (*R*-) genes termed nucleotide binding site-leucine rich receptors (NBS-LRRs) (Dodds and Rathjen 2010) are challenging to assemble as they are often organized in multigenic clusters with many tandem duplications and rapid pseudogenization. The TGACv1 assembly contains 2595 NBS-containing genes (Table 4) of which 1185 are NBS-LRR genes. Among these, 98% have complete transcripts compared with only 2% in the CSS assembly. We also used NLR-parser (Steuernagel et al. 2015) to predict the coiled-coil (CC-) NBS-LRR subclass of *R*-genes. We identified 859 complete CC-NBS-LRR genes supported by specific MEME motifs (Jupe et al. 2012) compared with 225 in the CSS assembly (Table 4). The total of 1185 wheat NBS-NLRs was consistent with that found in diploid wheat progenitors (402 NLRs in *T. urartu*) and diploid relatives (438 in *O. sativa*) (Sarris et al. 2016). Nearly 90% of CS42 *R*-genes were unambiguously assigned to chromosome arms, and 57% (674/1185) were anchored to the TGACv1 map. The number of *R*-genes per scaffold ranged from one to 31, compared with only two to three *R*-gene per scaffold in the CSS wheat assembly (The International Wheat Genome Sequencing Consortium 2014). This finding is corroborated by BAC sequence assemblies (Supplemental Information S1; Supplemental Fig. S12.1).

### Gluten genes

Glutens form the major group of grain storage proteins, accounting for 10%–15% of grain dry weight, and confer viscoelastic properties essential for bread-making (Shewry et al. 1995). Gluten genes encode proteins rich in glutamines and prolines that form low-complexity sequences composed of PxQ motifs, and occur in tandem repeats in highly complex loci that have posed significant challenges for their assembly and annotation. We characterized the gluten genes in the TGACv1 assembly and showed that most of the known genes were fully assembled. Gluten loci, while still fragmented, exhibit much greater contiguity than in the CSS assembly (The International Wheat Genome Sequencing Consortium 2014) with up to six genes per scaffold (Supplemental Information S1; Supplemental Fig. S12.2). We identified all assembly regions with nucleotide similarity to publicly available gluten sequences, adding an additional 33 gluten genes to the annotation and manually correcting 21 gene models. In total, we identified 105 full-length or partial gluten genes and 13 pseudogenes in the TGACv1 assembly (Table 4; Supplemental Information S1, section S12.2).

### The gibberellin biosynthetic and signaling pathway

Mutations in the gibberellin (GA) biosynthetic and signal transduction pathways have been exploited in wheat, where gain-of-function mutations in the GA signaling protein *Rht-1* confer GA insensitivity and a range of dwarfing effects. Most modern wheat cultivars carry semi-dominant *Rht-1* alleles (Phillips 2016), but these alleles also confer negative pleiotropic effects, including reduced male fertility and grain size. Hence, there is considerable interest in developing alternative dwarfing alleles based on GA-biosynthetic genes such as *GA20ox2*. A prerequisite for this is access to a complete set of genes encoding the biosynthetic pathway. Figure 6 shows that the TGACv1 assembly contains full-length sequences for 67 of the expected 72 GA pathway genes, in contrast to only 23 genes in the CSS assembly (The International Wheat

**Table 4.** Disease-resistance and gluten gene repertoires in the TGACv1 assembly

R-genes			Gluten genes	
	CSS	TGACv1	CDS	Pseudogenes
<i>NBS</i> -containing (Pfam)	1224	2595	<i>Gliadins</i>	
Fragmented	1188	65	Alpha	29
Complete transcript	36	2530	Gamma	18
No. of scaffolds	1195	1853	Unknown	14
Maximum genes per scaffold	3	31	Omega	10
<i>NBS-LRR</i> (Pfam)	627	1185	<i>Glutenins</i>	
Partial genes	611	11	HMW	6
Full-length genes	16	1174	LMW	16
No. of scaffolds	613	979	<i>Prolamins</i>	
Maximum genes per scaffold	2	13	Avenin	4
<i>CC-NBS-LRR</i> (NLR-parser)	225	859	Farinin	4
			Globulin	2
			Hordein	1
			Unknown	1
			Total	105
				13

Resistance genes were identified by their characteristic domain architecture (Sarris et al. 2016). Gluten genes were identified by sequence similarity to either a gliadin, glutenin, or generic prolamin class, representing prolamin-like glutens discovered in oat (avenin), wheat (farinin), or barley (hordein). See Supplemental Information, section 12.

Genome Sequencing Consortium 2014). Two paralogs of *GA20ox3* on Chromosome 3D are separated by 460 kb, and *GA1ox-B1* and *GA3ox-B3* are separated by 3.2 kb, suggesting common ancestry of these two enzymes with different catalytic activities (Pearce et al. 2015).

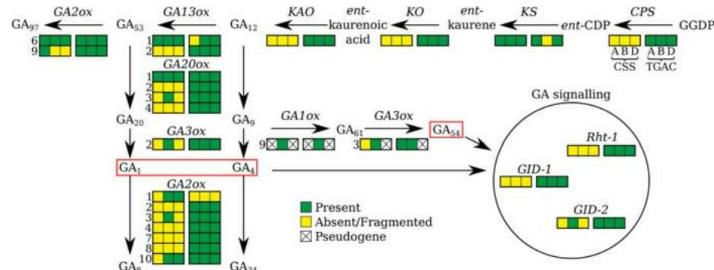
## Discussion

Access to a complete and robust wheat genome assembly is essential for the continued improvement of wheat, a staple crop of global significance with 728 M tonnes produced in 2014 (<http://fenix.fao.org/faostat/beta/en/#home>). The capacity to assemble and annotate wheat genomes accurately, rapidly, and cost-effectively addresses key social, economic, and academic priorities by facilitating trait analyses, by exploiting diverse germplasm resources, and by accelerating plant breeding. However, polyploidy and the extensive repeat regions in wheat have limited the completeness of previous assembly efforts (Brenchley et al. 2012; The International Wheat Genome Sequencing Consortium 2014; Chapman et al. 2015), reducing their utility.

Here we report a much more complete wheat genome assembly, representing ~80% of the 17-Gb genome in large scaffolds. We combined high-quality PCR-free libraries and precisely size-selected LMP libraries (Heavens et al. 2015) with the w2rap assembly software (Clavijo et al. 2017) to generate contiguous and complete assemblies from relatively low (about 33×) Illumina PE read coverage and LMP libraries. The contiguity of the TGACv1 assembly allowed us to create a greatly improved gene annotation supported by extensive transcriptome data. Over 78% of the 104,091 HC protein-coding genes are fully supported by RNA-seq data. These improvements identified 22,904 genes that were absent from previous wheat gene sets

(The International Wheat Genome Sequencing Consortium 2014; Choulet et al. 2014), almost all of which have a homolog in other species (Fig. 3B). The robustness of the annotation is further supported by the use of high-quality PacBio data and agreement with proteomic data, with 42% of the HC gene models supported by sequenced peptides. This new wheat gene set provides an improved foundation for wheat research. Finally, incorporation of strand-specific Illumina RNA-seq libraries into the annotation showed that nearly half of the HC genes were alternatively spliced, in line with observations in many other plants (Zhang et al. 2015).

A well-defined gene set in large sequence scaffolds is an essential foundation for trait analyses in wheat. We identified the complement of disease-resistance genes, gluten protein genes that confer nutritional and bread-making quality of wheat grains, and the set of GA biosynthetic and signal transduction genes that are important determinants of crop height and yield. An accurate gene set is also essential for understanding expression of gene families in complex allopolyploid genomes. We observed that 20% of homoeologous triads showed differential expression in seedling



**Figure 6.** Genes encoding the gibberellin (GA) biosynthetic and signaling pathway in bread wheat. The GA biosynthesis, inactivation, and signal transduction pathway, illustrating the representation of the gene sequences in CSS and TGACv1 assemblies. If more than one paralog is known for a gene, its number according to the classification by Pearce et al. (2015) is indicated on the left of the box. Bioactive GAs are boxed in red.

leaves subject to biotic and abiotic stress conditions. This is consistent with coexpression analyses in developing grains (Pfeifer et al. 2014), where most differentially expressed genes were single homoeologs that were up/down-regulated. Taken together, these results identify widespread subfunctionalization of homoeologous genes due to differential regulation. The new assembly and annotation will enable the identification of multiple sequence differences in promoters, transcription start sites, gene splicing, and other features among strict homoeologs, providing a foundation for systematic analyses of the causes of these differences.

Generating complete and accurate wheat genome assemblies is essential for capturing the full range of genetic variation in wheat genomes. By identifying this variation, genomics will directly facilitate trait analyses and accelerate plant breeding. Our rapid, accurate, and cost-effective assembly approach is suitable for assembling multiple wheat and other Triticeae genomes in robust and comparable ways, using relatively inexpensive sequencing technologies based on PCR-free libraries and open-source software. We anticipate that researchers with access to suitable computational infrastructure will use the approaches described here to sequence multiple wheat varieties, including elite varieties, unimproved landraces, and progenitor species. These assemblies will reveal a wide spectrum of genetic variation, including large-scale structural changes such as translocations and chromosome additions that are known to play a major role in the adaptation of the wheat crop to different growing environments. By adopting this pan-genomics approach, we will enrich our understanding of complex genome evolution and the plasticity of genome regulation and empower new approaches to wheat improvement.

## Methods

### DNA library preparation and sequencing

A full description of the DNA preparation and sequencing methods is in Supplemental Information. PCR-free PE libraries were sequenced using 2x 250-bp reads on HiSeq2500 platforms for contig generation. TALL libraries and Nextera LMP libraries (Heavens et al. 2015) were used for scaffolding. Insert size distributions (Supplemental Information S1; Supplemental Figs. S4.1–S4.3) were checked by mapping to the CS42 Chromosome 3B pseudomolecule (Choulet et al. 2014) using the DRAGEN coprocessor (<http://www.edicogenome.com/dragen/>).

### Genome assembly

Assembly was performed using the Wheat/Whole-Genome Robust Assembly Pipeline, w2rap (Clavijo et al. 2017). It combines the w2rap-contigger, based on DISCOVAR de novo (Weisenfeld et al. 2014), an LMP preparation approach based on FLASH (Magoc and Salzberg 2011) and Nextclip (Leggett et al. 2014), and scaffold with SOAPdenovo2 (Luo et al. 2012). The w2rap-contigger takes advantage of DISCOVAR (Weisenfeld et al. 2014; Love et al. 2016) algorithms to preserve sequence variation during assembly but has been further developed to enable processing of much larger data volumes and complex genomic repeats. The paired-end read data set was assembled into contigs on a SGI UV200 machine with 7TB of shared RAM. The contig assembly took 38 d using 64 cpus, with the default settings of the w2rap-contigger from [https://github.com/bioinfologics/w2rap-contigger/releases/tag/CS42\\_TGACv1](https://github.com/bioinfologics/w2rap-contigger/releases/tag/CS42_TGACv1). Newer versions of w2rap can achieve similar results in half the time or less, using close to half the memory. Scaffolding with the LMP data took a total of 10 d and was execut-

ed on the same hardware but used 128 cpus and <1 TB of RAM. Contigs were scaffolded using the PE, LMP, and TALL reads and the SOAPdenovo2 (Luo et al. 2012) prepare→map→scaffold pipeline, run at  $k=71$ . Contigs and scaffolds were quality controlled using KAT spectra-cn plots (Mapleson et al. 2017) to assess motif representation.

### Gene annotation

A high-quality gene set for wheat was generated using a custom pipeline integrating wheat-specific transcriptomic data, protein similarity, and evidence-guided gene predictions generated with AUGUSTUS (Stanke and Morgenstern 2005). Full methods are in Supplemental Information S8. RNA-seq reads (ERP004714, ERP004505, and 250-bp PE strand-specific reads from six different tissues) were assembled using four alternative assembly methods (Trapnell et al. 2010; Haas et al. 2013; Pertea et al. 2015; Song et al. 2016) and integrated with PacBio transcripts into a coherent and nonredundant set of models using Mikado (<https://github.com/lucuenturini/mikado>). PacBio reads were then classified according to protein similarity and a subset of high-quality (e.g., full length, canonical splicing, nonredundant) transcripts used to train an AUGUSTUS wheat-specific gene prediction model. AUGUSTUS was then used to generate a first draft of the genome annotation, using as input Mikado-filtered transcript models, reliable junctions identified with Portcullis (<https://github.com/maplesond/portcullis>), and peptide alignments of proteins from five close wheat relatives (*B. distachyon*, maize, rice, *S. bicolor*, and *S. italica*). This draft annotation was refined by correcting probable gene fusions, missing loci and alternative splice variants. The annotation was functionally annotated, and all loci were assigned a confidence rank based on their similarity to known proteins and their agreement with transcriptome data.

### Data access

All data generated in this study have been submitted to the European Nucleotide Archive (ENA; <http://www.ebi.ac.uk/ena>) under accession numbers PRJEB15378, PRJEB15378 (PE and LMP reads used for genome assembly and scaffolding), PRJEB11773 (genome assembly), and PRJEB15048 (Illumina and PacBio reads used for genome annotation). The assembly and annotation are available in Ensembl Plants (release 32; Ensembl Plants, [http://plants.ensembl.org/Triticum\\_aestivum/Info/Index](http://plants.ensembl.org/Triticum_aestivum/Info/Index)) and from the Earlham Institute Open Data site (EI; [http://opendata.earlham.ac.uk/Triticum\\_aestivum/TGAC/v1/](http://opendata.earlham.ac.uk/Triticum_aestivum/TGAC/v1/)). BLAST services for these data sets are available via Grassroots Genomics (Grassroots; <https://wheatis.tgac.ac.uk/grassroots-portal/blast>).

### Acknowledgments

We thank Jane Rogers and Mario Caccamo for initial discussions and planning and Burkhard Steuernagel for assistance with NLR-parser. This work was funded by a Biotechnology and Biological Sciences Research Council (BBSRC) strategic LOLA Award to M.W.B. and C.U. (BB/J003557/1), M.D.C. (BB/J003743/1), P.J.K. (BB/J00328X/1), and A.L.P. (BB/J003913/1); a BBSRC Anniversary Future Leader Fellowship (BB/M014045/1) to P.B.; BBSRC Institute Strategic Programme Grants GRO (BB/J004588/1) to M.W.B. and C.U.; “2020 Wheat” (BBS/E/C/00005202) to A.L.P.; and Bioinformatics BB/J004669/1 to F.D.P. The German Ministry of Education and Research (BMBF) grant 031A536 “de.NBI” supported M.S., and the Australian Research Council (LP120200102, CE140100008) and Agilent Technologies Australia supported A.H.M. Next-generation sequencing and library construction was

delivered via the BBSRC National Capability in Genomics (BB/J010375/1) at the Earlham Institute (EI; formerly The Genome Analysis Centre, Norwich), by members of the Platforms and Pipelines Group. Open data access and BLAST databases and service are provided by the EI Data Infrastructure group.

## References

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.

Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, Ben-Hur A, Reddy ASN. 2016. A survey of the sorghum transcriptome using single-molecule long reads. *Nat Commun* **7**: 11706.

Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12**: R18.

Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, Bento P, Da Silva C, Labadie K, Alberti A, et al. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun* **5**: 3657.

Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastic AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, et al. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat Genet* (in press). doi: 10.1038/ng.3802.

Bishara A, Liu Y, Weng Z, Kashef-Haghghi D, Newburger DE, West R, Sidow A, Batzoglou S. 2015. Read clouds uncover conservation in complex regions of the human genome. *Genome Res* **25**: 1570–1580.

Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Society* **16**: 1667–1678.

Borrelli P, Adamski N, Uauy C. 2015. Genomics as the key to unlocking the polyploid potential of wheat. *New Phytologist* **208**: 1008–1022.

Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, et al. 2012. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**: 705–710.

Chaisson MJ, Wilson RK, Eichler EE. 2015. Genetic variation and the de novo assembly of human genomes. *Nat Rev Genet* **16**: 627–640.

Chapman JA, Ho I, Sunkara S, Luo S, Schrotto GP, Rokhsar DS. 2011. Meraculous: de novo genome assembly with short paired-end reads. *PLoS One* **6**: e23501.

Chapman JA, Mascher M, Bulut A, Barry K, Georganas E, Session A, Strnadova V, Jenkins J, Sehgal S, Oliker L, et al. 2015. A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biol* **16**: 26.

Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, et al. 2014. Structural and functional partitioning of bread wheat chromosomes 3B. *Science* **345**: 1249721.

Clavijo B, Garcia Accinelli G, Wright J, Heavens D, Barr K, Yanes L, Di Palma F. 2017. W2RAP: a pipeline for high quality, robust assemblies of large complex genomes from short read data. *bioRxiv* **2017**: 110999.

DeVos KM, Dubcovsky J, Dvorak J, Chinoy CN, Gale MD. 1995. Structural evolution of wheat chromosomes 4A, 5A and 7B and its impact on recombination. *Theor Appl Genet* **91**: 282–288.

Dodds PN, Rathjen JP. 2010. Plant immunity: towards an integrated view of plant-pathogen interactions. *Nat Rev Genet* **11**: 539–548.

Gan X, Stegle O, Behr J, Steffens JG, Drewe P, Hildebrand KL, Lyngsøe R, Schultheiss SJ, Osborne BJ, Sreedharan VT, et al. 2011. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **477**: 419–423.

Gnerre S, MacCallum I, Przyblyski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci* **108**: 1513–1518.

Gordon D, Huddleston J, Chaisson MJ, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW, et al. 2016. Long-read sequence assembly of the gorilla genome. *Science* **352**: aae0344.

Griffiths S, Sharp R, Foote TN, Bertin I, Wanous M, Reader S, Colas I, Moore G. 2006. Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. *Nature* **439**: 749–752.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**: 1494–1512.

Heavens D, Accinelli GG, Clavijo B, Clark MD. 2015. A method to simultaneously construct up to 12 differently sized Illumina Nextera long mate pair libraries with reduced DNA input, time, and cost. *Biotechniques* **59**: 42–45.

The International Wheat Genome Sequencing Consortium. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* **345**: 1251788.

Jupe F, Pritchard L, Etherington GJ, MacKenzie K, Cock PJ, Wright F, Sharma SK, Bolser D, Bryan GJ, Jones JD, et al. 2012. Identification and localisation of the NB-LRR gene family within the potato genome. *BMC Genomics* **13**: 75.

Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* **6**: 291–295.

Lam HYK, Clark MJ, Chen R, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley EA, Gerstein MB, et al. 2011. Performance comparison of whole-genome sequencing platforms. *Nat Biotechnol* **30**: 78–82.

Leggett RM, Clavijo BJ, Clissold L, Clark MD, Caccamo M. 2014. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics* **30**: 566–568.

Li L. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**: 2178–2189.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York NY)* **326**: 289–293.

Liu Z, Xin M, Qin J, Peng H, Ni Z, Yao Y, Sun Q. 2015. Temporal transcriptome profiling reveals expression partitioning of homeologous genes contributing to heat and drought acclimation in wheat (*Triticum aestivum* L.). *BMC Plant Biol* **15**: 152.

Love RR, Weisenfeld NI, Jaffe DB, Besansky NJ, Neafsey DE. 2016. Evaluation of DISCOVOR de novo using a mosquito sample for cost-effective short-read genome assembly. *BMC Genomics* **17**: 187.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**: 18.

Lykke-Andersen S, Jensen TH. 2015. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat Rev Mol Cell Biol* **16**: 665–677.

Ma J, Stiller J, Berkman PJ, Wei Y, Rogers J, Feuillet C, Dolezel J, Mayer KF, Eversole K, Zheng Y-L, et al. 2013. Sequence-based analysis of translocations and inversions in bread wheat (*Triticum aestivum* L.). *PLoS One* **8**: e79329.

Magoc T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**: 2957–2963.

Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. 2017. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**: 574–576.

Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, Jakobsen KS, Wulff BBH, Steuernagel B, Mayer KFX, Olsen O-A, et al. 2014. Ancient hybridizations among the ancestral genomes of bread wheat. *Science* **345**: 1250092.

Moore G, Devos K, Wang Z, Gale M. 1995. Cereal genome evolution: grasses, line up and form a circle. *Curr Biol* **5**: 737–739.

Mostovoy Y, Levy-Sakin M, Lam J, Lam ET, Hastic AR, Marks P, Lee J, Chu C, Lin C, Dzakula Z, et al. 2016. A hybrid approach for de novo human genome sequence assembly and phasing. *Nat Methods* **13**: 587–590.

Panahi B, Mohammadi SA, Khaksefid RE, Fallah Mehrabadi J, Ebrahimi E. 2015. Genome-wide analysis of alternative splicing events in *Hordeum vulgare*: highlighting retention of intron-based splicing and its possible function through network analysis. *FEBS Lett* **589**: 3564–3575.

Pearce S, Huttly AK, Prosser IM, Li Y-d, Vaughan SP, Gallova B, Patil A, Coghill JA, Dubcovsky J, Hedden P, et al. 2015. Heterologous expression and transcript analysis of gibberellin biosynthetic genes of grasses reveals novel functionality in the GA3ox family. *BMC Plant Biol* **15**: 130.

Pendleton M, Sebra R, Pang AWC, Ummat A, Franzen O, Rausch T, Stütz AM, Stedman W, Anantharaman T, Hastic A, et al. 2015. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods* **12**: 780–786.

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTIE enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**: 290–295.

Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvistendahl TR, International Wheat Genome Sequencing Consortium, Mayer KFX, Olsen O-A. 2014. Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science* **345**: 1250091.

Phillips AL. 2016. Genetic control of gibberellin metabolism and signalling in crop improvement. In *Annual plant reviews*, Vol. 49, pp. 405–430. John Wiley & Sons, Chichester, UK.

Pingault L, Choulet F, Alberti A, Glover N, Wincker P, Feuillet C, Paux E. 2015. Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome. *Genome Biol* **16**: 29.

Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, et al. 2016. Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *Genome Res* **26**: 342–350.

Riley R, Kimber G, Chapman V. 1961. Origin of genetic control of diploid-like behavior of polyploid wheat. *J Heredity* **52**: 22–25.

Sarris PF, Cevik V, Dagdas G, Jones JDG, Krasileva KV. 2016. Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens. *BMC Biol* **14**: 8.

Sears ER. 1966. Nullisomic-tetrasomic combinations in hexaploid wheat. In *Chromosome manipulations and plant genetics*, pp. 29–45. Springer, Boston, MA.

Shewry PR, Tatham AS, Barro F, Barcelo P, Lazzari P. 1995. Biotechnology of breadmaking: unravelling and manipulating the multi-protein gluten complex. *Biotechnology* **13**: 1185–1190.

Song L, Sabuncyan S, Florea L. 2016. CLASS2: accurate and efficient splice variant annotation from RNA-seq reads. *Nucleic Acids Res* **44**: e98.

Staiger D, Brown JWS. 2013. Alternative splicing at the intersection of biological timing, development, and stress responses. *Plant Cell* **25**: 3640–3656.

Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* **33**: W465–W467.

Steuernagel B, Jupe F, Witek K, Jones JDG, Wulff BBH. 2015. NLR-parser: rapid annotation of plant NLR complements. *Bioinformatics* **31**: 1665–1667.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511–515.

Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**: 36–46.

Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, Barry K, Lucas S, Harmon-Smith M, Lail K, et al. 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**: 763–768.

Wagner GP, Kin K, Lynch VJ. 2013. A model based criterion for gene expression calls using RNA-seq data. *Theory Biosci* **132**: 159–164.

Weisenfeld NI, Yin S, Sharpe T, Lau B, Hegarty R, Holmes L, Sogoloff B, Tabbaa D, Williams L, Russ C, et al. 2014. Comprehensive variation discovery in single human genomes. *Nat Genet* **46**: 1350–1355.

Weisenfeld NI, Kumar V, Shah P, Church D, Jaffe DB. 2016. Direct determination of diploid genome sequences. *Genome Res* (this issue). doi: 10.1101/gr.214874.116.

Zhang C, Yang H, Yang H. 2015. Evolutionary character of alternative splicing in plants. *Bioinformatics Biol Insights* **9**: 47–52.

Zimin AV, Puu D, Luo M-C, Zhu T, Koren S, Marcais G, Yorke JA, Dvorak J, Salzberg SL. 2017. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the mega-reads algorithm. *Genome Res* (this issue). doi: 10.1101/gr.213405.116.

Received October 13, 2016; accepted in revised form March 14, 2017.

**Supplemental material for *An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations.***

March 14, 2017

**Contents**

<b>1 Germplasm and DNA isolation</b>	<b>3</b>
1.1 Germplasm . . . . .	3
1.2 DNA isolation . . . . .	3
<b>2 Sequencing library preparation</b>	<b>4</b>
2.1 Amplification-free paired-end library construction protocol . . . . .	4
2.2 Tight, Amplification-free, Large insert Libraries (TALL) paired-end library construction protocol . . . . .	4
2.3 Long mate-pair library construction protocol . . . . .	4
<b>3 RNA sequencing</b>	<b>6</b>
3.1 Stranded RNA sequencing on Illumina HiSeq2500 . . . . .	6
3.2 Isoform Sequencing (Iso-Seq™) . . . . .	6
<b>4 Genomic assembly and scaffolding</b>	<b>7</b>
4.1 Contig assembly . . . . .	7
4.2 Scaffolding . . . . .	7
4.3 Contamination screening and filtering . . . . .	9
4.4 Chromosome arm binning . . . . .	9
4.5 Sequence length and content filter . . . . .	9
4.6 Assignment of scaffold identifiers . . . . .	9
4.7 Comparison of TGACv1 scaffolds to Chapman and CSS assemblies . . . . .	9
4.8 Assembly validation . . . . .	11
4.8.1 Contig validation by mate-pair link support . . . . .	11
4.8.2 Scaffold validation by gene order between 3B and TGACv1 3B sequences . . . . .	11
4.9 Assessment of chromosome arm assignment accuracy . . . . .	12
<b>5 Integration with genetic maps and chromosomal alignments</b>	<b>13</b>
<b>6 Detection and confirmation of chromosomal translocations</b>	<b>14</b>
6.1 Detection of translocations from OrthoMCL output . . . . .	14
6.2 PCR assays of suspected translocations . . . . .	14
6.3 Cross-validation of translocations by using the genetic map . . . . .	15
<b>7 Repeat analysis</b>	<b>17</b>
<b>8 Construction of the wheat gene set</b>	<b>18</b>
8.1 Reference guided transcriptome reconstruction . . . . .	18
8.1.1 Alignment of Illumina RNA-seq data . . . . .	18
8.1.2 Alignment of PacBio RNA-seq data . . . . .	19
8.1.3 Transcript assembly . . . . .	19
8.2 Gene predictor training . . . . .	20
8.3 Gene prediction using evidence guided AUGUSTUS . . . . .	20
8.3.1 Generation of external hints for gene prediction . . . . .	20
8.3.2 Gene prediction . . . . .	21
8.4 Gene model refinement . . . . .	22

8.5	Assignment of gene biotypes and confidence classification . . . . .	22
8.5.1	Cross species protein similarity ranking . . . . .	23
8.5.2	Wheat transcript support ranking . . . . .	23
8.5.3	Assignment of a locus biotype . . . . .	23
8.5.4	Removal of spurious genes . . . . .	24
8.5.5	Assignment of high and low confidence tags . . . . .	24
8.5.6	Assignment of a representative gene model . . . . .	26
8.5.7	Assessment of the TGACv1 annotation . . . . .	26
8.5.8	Evaluation of non-coding RNAs . . . . .	29
8.6	Alternative splicing analysis . . . . .	30
8.7	Functional annotation of protein coding transcripts . . . . .	30
8.8	Data Access . . . . .	30
<b>9</b>	<b>Proteomics</b>	<b>31</b>
<b>10</b>	<b>Orthologous gene family analyses</b>	<b>32</b>
10.1	OrthoMCL gene family clustering of wheat subgenome genes . . . . .	32
10.2	OrthoMCL gene family clustering of the bread wheat genome and related species . . . . .	32
10.3	GO over-/under-representation for specific groups/singletons . . . . .	33
10.4	Expanded gene families in OrthoMCL and GO over-representation within . . . . .	33
<b>11</b>	<b>Gene expression analyses</b>	<b>35</b>
11.1	Expression quantification and analysis . . . . .	35
11.1.1	Gene expression quantification . . . . .	35
11.1.2	Differential gene expression analysis . . . . .	35
11.1.3	Visualisation of gene expression . . . . .	36
11.2	Gene expression across 17 diverse RNA-seq studies . . . . .	36
11.3	Gene expression patterns across chromosome regions . . . . .	36
11.4	Analysis of homoeolog gene expression in stress conditions . . . . .	36
11.5	Homoeologous gene expression analysis . . . . .	37
11.6	Gene expression in syntenic loci . . . . .	38
<b>12</b>	<b>Gene families of agronomic importance</b>	<b>41</b>
12.1	Disease resistance genes . . . . .	41
12.2	Gluten genes . . . . .	41
12.3	Gibberellin genes . . . . .	42
12.4	BAC analysis . . . . .	42
<b>13</b>	<b>Authors' contributions</b>	<b>43</b>
<b>14</b>	<b>File list</b>	<b>43</b>
<b>15</b>	<b>References</b>	<b>44</b>

## 1 Germplasm and DNA isolation

### 1.1 Germplasm

A single seed descent line of *Triticum aestivum* Chinese Spring (called CS42) was used for DNA extraction. The provenance of the line has been traced to original Sears material.

### 1.2 DNA isolation

High molecular weight wheat DNA was isolated from leaf material of 2–3 week old CS42 plants that had been kept in the dark for 48 hours to reduce starch levels. Leaf material (60–80g) was frozen in liquid N<sub>2</sub> and ground to a fine powder in a mortar and pestle. Ground leaf tissue was transferred into ice-cold SEB buffer + mercaptoethanol (ME), using a ratio of 15ml SEB+ME per gram of leaf material. The leaf tissue and buffer was gently mixed for 20 seconds every 2 minutes for 10–15 minutes on ice, and then filtered twice through two layers of Miracloth with gentle squeezing. 1/20 volume of SEB+ME+ 10% v/v Triton X100 was added and mixed for 20 seconds every 2 minutes for a total of 10 minutes. The mixture was centrifuged at 600× g for 20 minutes at 4°C in 250ml polypropylene centrifuge bottles. The supernatant was removed gently with a pipette, and 1ml SEB+ME added to each pellet to gently resuspend it. SEB+ME was added to a total of 20ml and the crude nuclei were centrifuged again. This step was repeated twice, and washed nuclei were resuspended in a total of 7.5ml SEB+ME. 20% w/v SDS was added to final concentration of 2% w/v, and the mixture inverted gently to lyse the nuclei. The lysed nuclei were heated at 60°C for 10 minutes in a waterbath, cooled to room temperature, and 5M sodium perchlorate added to a final concentration of 1M to further disrupt protein-nucleic acid interactions. The lysate was centrifuged at 500× g for 20 minutes at 10°C to pellet starch grains, and the supernatant transferred to a new 15ml tube using a cut-off 1ml pipette tip to minimise shearing DNA. The nucleic acid solution was extracted with an equal volume of phenol/chloroform/isoamyl alcohol (25:24:1) and gently rocked (18 cycles/ minutes for 15 minutes. The mixture was centrifuged at 3000× g for 10 minutes in a swinging bucket rotor, the supernatant transferred to a new tube, and re-extracted. The final aqueous phase was dialysed in TE pH 7.0 at 4°C overnight. RNase T1 and RNase A were added to the dialysate to 50U/ml and 50µg/ml respectively, gently mixed by inversion, and incubated at 37°C for 45–60 minutes. Proteinase K was added to a final concentration of 150µg/ml and incubated for a further 45–60 minutes. The DNA was then extracted twice with phenol/chloroform/iso-amyl alcohol, and DNA was precipitated from the final aqueous phase by the addition of 1/10 volume of 3M sodium acetate (pH 5.2) and 2.5 volumes of ethanol. DNA was precipitated by centrifugation at 5000× g for 30 minutes at 4°C. The pellet was rinsed in 1ml of 70% ethanol, air dried for 1 hour, and resuspended in TE buffer. Final yields were 50–100µg DNA per 100g of leaf material.

#### Buffers

TKE
Tris (0.1M) 6.055g
KCl (1M) 37.275g
EDTA (0.1M) 18.61g
MBG water to 500ml

Store at 4°C. Do not adjust pH.

#### SEB

Sucrose	171.2g
PEG 800	1.2g
Carbamic acid	1.3g
Spermine	0.35g (Place at 37°C if forms a solid block)
Spermidine	1g
TKE	100ml
MBG water	1000ml

Adjust to pH 9.5 with concentrated HCl if needed.

Add 2ml Mercaptoethanol (BME) to the SEB just before use.

## 2 Sequencing library preparation

### 2.1 Amplification-free paired-end library construction protocol

A total of 600ng of DNA was sheared in a 60 $\mu$ l volume on a Covaris S2 (Covaris, Massachusetts, USA) for 1 cycle of 40 seconds with a duty cycle of 5%, cycles per burst of 200 and intensity of 3. The fragmented molecules were then end repaired in 100 $\mu$ l volume using the NEB End Repair Module (NEB, Hitchin, UK) incubating the reaction at 22°C for 30 minutes. Post incubation 58 $\mu$ l beads of CleanPCR beads (GC Biotech, Alphen aan den Rijn, The Netherlands) were added using a positive displacement pipette to ensure accuracy and the DNA precipitated onto the beads. They were then washed twice with 70% ethanol and the end repaired molecules eluted in 25 $\mu$ l Nuclease free water (Qiagen, Manchester, UK). End repaired molecules were then A tailed in 30 $\mu$ l volume using in the NEB A tailing module (NEB) incubating the reaction at 37°C for 30 minutes. To the A tailed library molecules 1 $\mu$ l of an appropriate Illumina TruSeq Index adapter (Illumina, San Diego, USA) was added and mixed then 31 $\mu$ l of Blunt/ TA ligase (NEB) added and incubated at 22°C for 10 minutes. Post incubation 5 $\mu$ l of stop ligation was added and then the reaction incubated at room temperature for 5 minutes. Following this incubation 67 $\mu$ l beads of CleanPCR beads (GC Biotech, Alphen aan den Rijn, The Netherlands) were added and the DNA precipitated onto the beads. They were then washed twice with 70% ethanol and the library molecules eluted in 100 $\mu$ l nuclease free water. Two further CleanPCR bead based purifications were undertaken to remove any adapter dimer molecules that may have formed during the adapter ligation step. The first with 0.9 $\times$  volume beads, the second with 0.6 $\times$  and the final library eluted in 25 $\mu$ l Resuspension Buffer (Illumina).

Library QC was performed by running a 1 $\mu$ l aliquot on a High Sensitivity BioAnalyser chip (Agilent, Stockport, UK) and the DNA concentration measured using the High Sensitivity Qubit (Thermo Fisher, Cambridge, UK). To determine the number of viable library molecules the library was subjected to quantification by the Kappa qPCR Illumina quantification kit (Kapa Biosystems, London, UK) and a test lane run at 10pM on a MiSeq (Illumina) with 2 $\times$ 300bp reads to allow the library to be characterised prior to generation of the 60 $\times$  coverage required on the Hiseq2500s (Illumina) with a 2 $\times$ 250bp read metric.

### 2.2 Tight, Amplification-free, Large insert Libraries (TALL) paired-end library construction protocol

A total of 3 $\mu$ g of DNA was sheared in a 60 $\mu$ l volume on a Covaris S2 (Covaris, Massachusetts, USA) for 1 cycle of 40 seconds with a duty cycle of 5%, cycles per burst of 200 and intensity of 3. The fragmented DNA was then subjected to size selection on a Blue Pippin (Sage Science, Beverly, USA). The 40 $\mu$ l in each of collection wells was replaced with fresh buffer and the separation and elution current checked prior to loading the sample. To 30 $\mu$ l of the end repaired molecules 10 $\mu$ l of R2 marker solution was added and then loaded onto a 1.5% Cassette. The Blue Pippin was configured to collect fragments at 800bp using the tight settings. Post size selection, the 40 $\mu$ l from the collection well was recovered and the size isolated estimated on High Sensitivity BioAnalyser Chip and DNA concentration determined using a Qubit HS Assay.

The size selected molecules were then end repaired in 100 $\mu$ l volume using the NEB End Repair Module (NEB, Hitchin, UK) incubating the reaction at 22°C for 30 minutes. Post incubation 100 $\mu$ l beads of CleanPCR beads (GC Biotech, Alphen aan den Rijn, The Netherlands) were added and the DNA precipitated onto the beads. They were then washed twice with 70% ethanol and the end repaired molecules eluted in 25 $\mu$ l Nuclease free water (Qiagen, Manchester, UK).

End repaired molecules were then A tailed in 30 $\mu$ l volume using in the NEB A tailing module (NEB) incubating the reaction at 37°C for 30 minutes. To the A tailed library molecules 1 $\mu$ l of an appropriate Illumina TruSeq Index adapter (Illumina, San Diego, USA) was added and mixed then 31 $\mu$ l of Blunt/ TA ligase (NEB) added and incubated at 22°C for 10 minutes. Post incubation 5 $\mu$ l of stop ligation was added and then the reaction incubated at room temperature for 5 minutes. Following this incubation 67 $\mu$ l beads of CleanPCR beads (GC Biotech, Alphen aan den Rijn, The Netherlands) were added and the DNA precipitated onto the beads. They were then washed twice with 70% ethanol and the library molecules eluted in 100 $\mu$ l nuclease free water. Two further 1 $\times$  CleanPCR bead based purifications were undertaken to remove any adapter dimer molecules that may have formed during the adapter ligation step and the final library eluted in 25 $\mu$ l Resuspension Buffer (Illumina).

Library QC was performed by running a 1 $\mu$ l aliquot on a High Sensitivity BioAnalyser chip (Agilent, Stockport, UK) and the DNA concentration measured using the High Sensitivity Qubit (Thermo Fisher, Cambridge, UK). To determine the number of viable library molecules the library was subjected to quantification by the Kappa qPCR Illumina quantification kit (Kapa Biosystems, London, UK) and then sequenced on the Hiseq2500s (Illumina) with a 2 $\times$ 150bp read metric.

### 2.3 Long mate-pair library construction protocol

For the Tagmentation reactions 3 $\mu$ g and 6 $\mu$ g of Genomic DNA was prepared in 308 $\mu$ l and then 80 $\mu$ l 5 $\times$  Tagment Buffer Mate Pair (Illumina) added followed by 12 $\mu$ l Mate Pair Tagmentation Enzyme (Illumina) and the reaction gently vortexed to mix. This was then incubated for 30 minutes at 55°C, 100 $\mu$ l of Neutralize Tagment Buffer (Illumina) added and then incubated at room temperature for 5 minutes. A 1 $\times$  volume bead clean-up was performed with CleanPCR beads and the DNA eluted in 165 $\mu$ l of Nuclease free Water. A 1 $\mu$ l aliquot was run on a BioAnalyser 1200 chip and DNA concentration determined using a Qubit HS Assay.

Strand Displacement was performed by combining 162 $\mu$ l of tagmented DNA, 20 $\mu$ l 10 $\times$  Strand Displacement Buffer (Illumina), 8 $\mu$ l dNTPs (Illumina) and 10 $\mu$ l Strand Displacement Polymerase (Illumina). This was then incubated at room temperature for 30 minutes. A 0.75 $\times$  volume bead clean-up was performed with CleanPCR beads and the DNA eluted in 16 $\mu$ l of Nuclease free Water and the eluted DNA from the 3 $\mu$ g and 6 $\mu$ g reactions pooled. A 1 $\mu$ l aliquot was diluted 1:6 and run on a BioAnalyser 1200 chip and DNA concentration determined using a Qubit HS Assay.

Size selection was performed on a Sage Science ELF (Sage Science, Beverly, USA). The 30 $\mu$ l in each of collection wells was replaced with fresh buffer and the collection and elution current checked prior to loading the sample. To 30 $\mu$ l of the pooled Strand Displaced reaction 10 $\mu$ l of loading solution was added and then loaded onto a 0.75% Cassette which was configured to separate the sample for 3 hours 30 minutes and then eluting each fraction for 35 minutes. Post size selection, the 30 $\mu$ l from each of the 12 collection wells was recovered and the DNA concentration determined using a Qubit HS Assay.

Circularisation was performed by combining 30 $\mu$ l of size fractionated DNA, 12.5 $\mu$ l of 10 $\times$  circularisation buffer (Illumina), 3 $\mu$ l Circularisation Enzyme (Illumina) and 85 $\mu$ l nuclease free water. These were then incubated at 30°C overnight. Linear DNA was digested by adding 3.75 $\mu$ l Exonuclease (Illumina) and incubating at 37°C for 30 minutes followed by 70°C for 30 minutes to denature the enzyme and 5 $\mu$ l of stop ligation (Illumina) added. During exonuclease treatment 240 $\mu$ l of M280 Dynabeads (Thermo Fisher) were prepared by washing twice with 600 $\mu$ l Bead Bind Buffer (Illumina) before resuspending in 1560 $\mu$ l Bead Bind Buffer. Circularised DNA was then sheared in a 130 $\mu$ l volume on a Covaris S2 for 2 cycles of 37secs with a duty cycle of 10%, cycles per burst of 200 and intensity of 4.

To 130 $\mu$ l fragmented DNA 130 $\mu$ l of washed M280 beads was added, mixed and then placed on a lab rotator at room temperature for 20 minutes. Library molecules bound to M280 beads were then washed four times with 200 $\mu$ l Bead Washer Buffer (Illumina) and twice with 200 $\mu$ l Resuspension Buffer (Illumina).

A master mix containing 1105 $\mu$ l nuclease free water, 130 $\mu$ l 10 $\times$  End Repair Reaction Buffer (NEB, Hitchin, UK) and 65 $\mu$ l end repair enzyme mix (NEB) was prepared and 100 $\mu$ l added to each tube, mixed with the beads and incubated at room temperature for 30 minutes. End repaired library molecules bound to M280 beads were then washed four times with 200 $\mu$ l Bead Washer Buffer and twice with 200 $\mu$ l Resuspension Buffer.

A master mix containing 325 $\mu$ l nuclease free water, 39 $\mu$ l A Tailing 10 $\times$  Reaction Buffer (NEB) and 26 $\mu$ l A tailing enzyme mix (NEB) was prepared and 30 $\mu$ l added to each tube, mixed with the beads and incubated at 37°C for 30 minutes. To the A tailed library molecules 1 $\mu$ l of the appropriate Illumina Index adapter (Illumina) was added and mixed then 31 $\mu$ l of Blunt/ TA ligase (NEB) added and incubated at room temperature for 10 minutes. Post incubation 5 $\mu$ l of stop ligation added and then the adapter ligated library molecules bound to M280 beads were then washed four times with 200 $\mu$ l Bead Washer Buffer and twice with 200 $\mu$ l Resuspension Buffer.

A master mix containing 240 $\mu$ l nuclease free water, 300 $\mu$ l 2 $\times$  Kappa HiFi (Kappa Biosystems) and 60 $\mu$ l Illumina Primer Cocktail (Illumina) was prepared and 50 $\mu$ l added to each tube, mixed with the beads and the contents, including beads, transferred to a 200 $\mu$ l PCR tube. Each sample was then subjected to amplification on a Veriti Thermal Cycler (Thermo Fisher) with the following conditions: 98°C for 3 minutes, 8, 10 or 12 cycles of PCR depending upon copy number entering circularisation of 98°C for 10 seconds, 60°C for 30 seconds, 72°C for 30 seconds followed by 72°C for 5 minutes and Hold at 4°C.

Post amplification the PCR tubes were placed on a magnetic plate, the beads allowed to pellet and then 45 $\mu$ l of the PCR transferred to a 2ml Lobind Eppendorf Tube. To this 31.5 $\mu$ l beads of CleanPCR beads were added to precipitate the DNA, the beads washed twice with 70% ethanol and the final library eluted in 20 $\mu$ l resuspension buffer. Library QC was performed by running a 1 $\mu$ l aliquot on a High Sensitivity BioAnalyser chip (Agilent) and the DNA concentration measured using the High Sensitivity Qubit (Thermo Fisher). Each library was then equimolar pooled (except for the largest insert library which was considerably weaker than the others which was at 10% concentration) based on DNA concentration. The quantification of the pool was determined by the Kappa qPCR Illumina quantification kit (KAPPA) with the pool run at 10pM on a MiSeq with a 2 $\times$ 300bp reads read metric.

Reads generated were then processed through NextClip which takes LMP FASTA reads and looks to categorise them into four groups based on the presence of the Nextera adapter junction sequence. Category A pairs contain the adaptor in both reads, Category B pairs contain the adaptor in only read 2, Category C pairs contain the adaptor in only read 1, Category D pairs do not contain the adaptor in either read. NextClip also uses a k-mer-based approach to estimate the PCR duplication rate while reads are examined. Filtered reads in categories A, B and C were then mapped back to the Wheat Chromosome 3B reference using BWA mem with the default parameters. This uses the reference sequence and measures from the leftmost to the rightmost aligned bases within the reads to determine the insert size.

Once characterised the libraries with inserts centred at 9kbp (Fraction 4) and 11kbp (Fraction 3) were then sequenced to greater depth as 2 $\times$ 250bp reads on HiSeq2500s.

### 3 RNA sequencing

#### 3.1 Stranded RNA sequencing on Illumina HiSeq2500

Quality checked libraries were quantified to range from 2.2nM to 9.87nM. Each library was then diluted to 2nM with NaOH and 5 $\mu$ l transferred into 995 $\mu$ l HT1 (Illumina) to give a final concentration of 10pM. 135 $\mu$ l of the diluted library pool was then transferred into a 200 $\mu$ l strip tube, spiked with 1% PhiX Control v3 and placed on ice before loading onto the Illumina cBot. The library was hybridised to the flow cell using HiSeq Rapid Paired End Cluster Generation Kit v2, following the Illumina RR\_TemplateHyb\_FirstExt\_VR recipe. Following the hybridisation procedure, the flow cell was loaded onto the Illumina HiSeq2500 instrument following the manufacturer's instructions. The sequencing chemistry utilised was HiSeq Rapid SBS v2 using HiSeq Control Software 2.2.58 and RTA 1.18.64. Each library was run across a single lane for 250 cycles for each paired end read. Reads in bcl format were demultiplexed based on the 6bp Illumina index by CASAVA 1.8, allowing for a one base-pair mismatch per library, and converted to FASTQ format by bcl2fastq.

#### 3.2 Isoform Sequencing (Iso-Seq™)

The procedures used followed the Pacific Biosciences protocol. <http://www.pacb.com/wp-content/uploads/Procedure-Checklist-Isoform-Sequencing-Iso-Seq-Analysis-using-the-Clontech-SMARTer-PCR-cDNA-Synthesis-Kit-and-SageELF-Size-Selection-System.pdf>

## 4 Genomic assembly and scaffolding

### 4.1 Contig assembly

We generated 1.1 billion 250bp paired-end reads from two PCR-free CS42 libraries (see Table S4.1) which provided 32.78 $\times$  coverage of the CS42 genome (approx. 30 $\times$  31-mer coverage). Insert size distributions of each library were checked by mapping to the CS42 chromosome 3B pseudo-molecule (Choulet et al., 2014) using the DRAGEN co-processor (EdicoGenome, 2014).

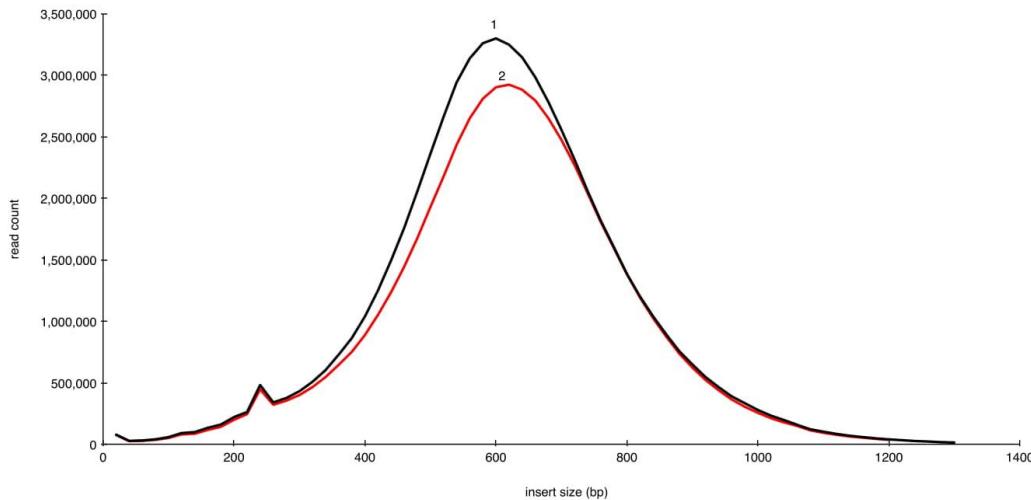


Figure S4.1: Insert size distributions of the two PE libraries.

A method based on DISCOVAR *de novo* (Weisenfeld et al., 2014) was chosen to assemble contigs as this approach utilises PCR-free libraries to reduce coverage bias, and uses long 250bp reads generated by the latest Illumina sequencing technology. Originally developed to assemble human genomes, the algorithm is designed to retain the majority of the variation present in the reads when generating the assembly, including variation between homologous chromosomes and repeat copies. This is important when assembling a repeat-rich hexaploid genome such as wheat to prevent collapsing of repeats and homologous/homoeologous regions during assembly. Contig assembly starts by correcting errors in the reads by creating “friend stacks” for each read in a read pair, then retaining only “true friends”, reads that perfectly match the original read pair (with an offset). A consensus sequence is called for each stack, in most cases including the gap between the read pairs consistent with the library fragmentation step. Overlaps between each stack are used to generate a ‘joint consensus’ for the original DNA fragment, typically yielding a single unambiguous joint consensus, a “closed pair”. The unipath graph is created from the consensus sequences, simplified to remove artifacts from the laboratory process, then the closed pair sequences are applied to the graph to join paths that overlap and pull apart regions containing collapsed repeats. The version of the contigger used is available in Github ([https://github.com/bioinfologics/w2rap-contigger/releases/tag/CS42\\_TGACv1](https://github.com/bioinfologics/w2rap-contigger/releases/tag/CS42_TGACv1)) and is fully described elsewhere (Clavijo et al., 2017). Contigs were QC’ed using KAT (Mapleson et al., 2017) spectra-cn plots to check motif representation. Importantly, our data generation was tailored to generate maximum complexity, precisely sized, low bias sampling.

### 4.2 Scaffolding

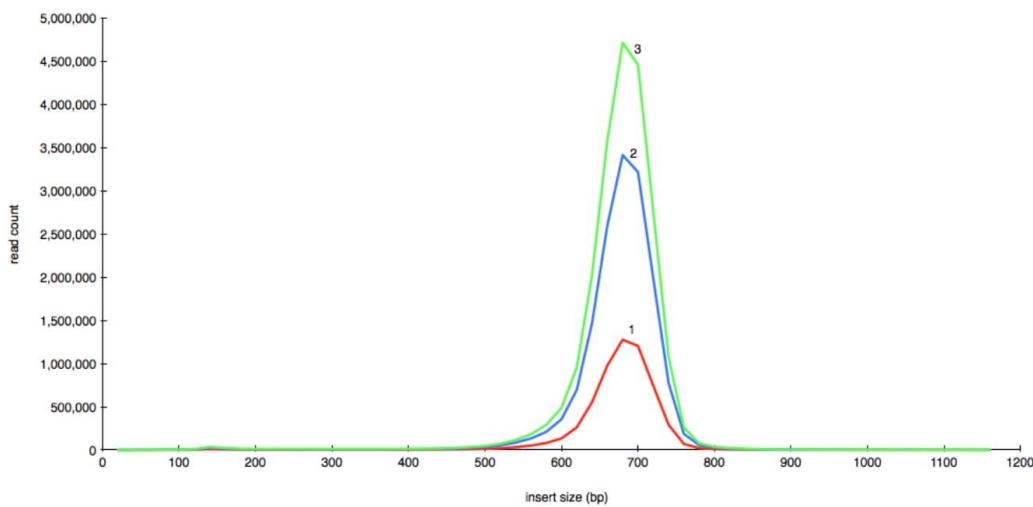
Multiple Nextera Long Mate Pair libraries were constructed as described above, QC’ed by alignment to the 3B pseudomolecule, and chosen for sequencing as described in our published LMP protocol (see Table S4.2; Heavens et al. (2015)). Raw reads were pre-processed using a pipeline based on NextClip (Leggett et al., 2014). Briefly, this pipeline merges overlapping read pairs with FLASH (Magoc and Salzberg, 2011), generates a read 2 by reverse complementing the read 1 sequence, then runs Nextclip to identify and trim reads containing the Nextera adaptor.

Table S4.1: Paired-end library details

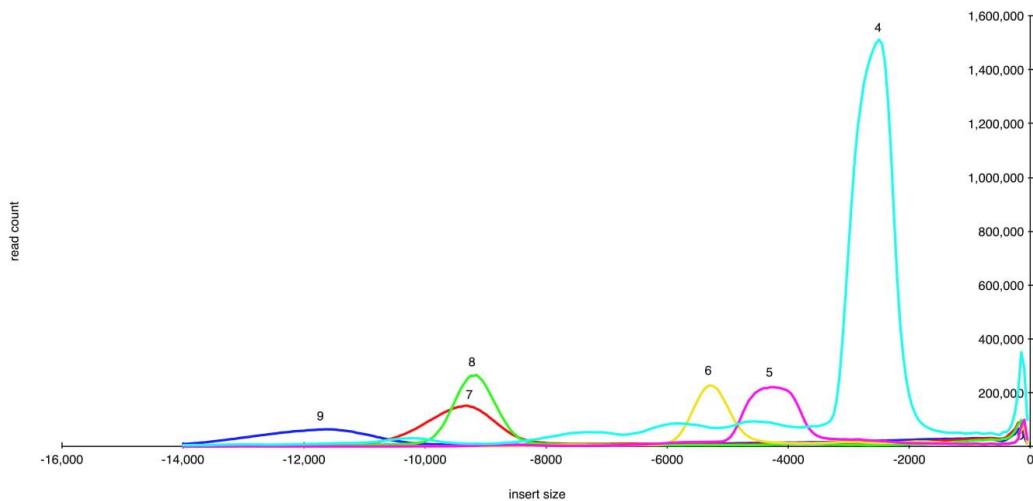
Library type	Read count	Read length (bps)	Insert size (bps)	Read coverage	
1	PCR-free	658,890,225	250	620	19.38
2	PCR-free	455,733,257	250	600	13.4

**Table S4.2:** Summary of library sequencing. \*Library 4 was sequenced twice, once generating 150bp reads and once generating 250bp reads.

Library	Type	Read count	Read length (bp)	Insert size (bp)	Read coverage	Fragment coverage
1	TALL	118,575,256	150	690	2.09	4.81
2	TALL	309,422,248	150	690	5.46	12.56
3	TALL	434,404,265	150	690	7.67	17.63
4*	MP	151,086,835	150	2,480	2.67	22.04
		508,236,686	250	2,480	14.95	74.14
5	MP	170,061,926	150	4,300	3.00	43.02
6	MP	142,304,055	150	5,250	2.51	43.95
7	MP	432,253,166	250	9,300	12.71	236.47
8	MP	173,921,104	250	9,180	5.12	93.92
9	MP	404,721,706	250	11,600	11.90	276.16



**Figure S4.2:** Insert size distribution for TALL libraries.



**Figure S4.3:** Insert size distribution for LMP libraries.

**Table S4.3:** Summary contig and N-remapped scaffolds for the TGACv1 assembly of Chinese Spring 42.

	Size (Gb)	Sequence count ( $\geq 500$ bp)	N20 (kb)	N50 (kb)	N80 (kb)	NG50 (kb)	L50	%N
Contigs	13.26	2,977,539	40.7	16.7	3.1	8.7	200,473	0.1
Scaffolds	14.07	1,333,497	175.6	83.9	25.4	63.1	47,111	5.6

In addition to the PE reads used to generate contigs, three Tight, Amplification-free, Large insert Library (TALL) and six mate-pair libraries were used for scaffolding. The TALL library protocol generates paired-end reads with a tight insert-size distribution without PCR-amplification and provided additional coverage for scaffolding. Contigs were scaffolded using SOAPdenovo2 (Luo et al., 2012). A  $k$ -mer length of 71 was used for the prepare and mapping stage. SOAPdenovo2 replaces N-stretches (gaps) in contigs with Cs and Gs during scaffolding, so to correct this contigs were mapped back to the scaffolds and the gaps converted back to Ns. Contig and scaffold contiguity statistics are shown in Table S4.3.

#### 4.3 Contamination screening and filtering

The scaffolds were checked for contamination against the NCBI nucleotide database using BLAST+ and the results joined to NCBI's taxonomy database. Filtering was applied to show hits of more than 98% identity over 90% of scaffold length. From this list, scaffolds identified with a taxonomy containing "BEP" (the grass BEP clade), "Poales" (the order encompassing grasses) or "eudicotyledons" (the dicot group of angiosperms) were kept and the remaining scaffolds were considered to be contamination. These were mainly short contigs containing PhiX.

#### 4.4 Chromosome arm binning

Scaffolds were classified into chromosome-arm bins using arm-specific Chromosome Survey Sequence (CSS) reads (The International Wheat Genome Sequencing Consortium, 2014). Scaffolds from 3B were not separated into short/long arm bins as individual arm datasets were not generated for this chromosome in the CSS project. The `sect` method of KAT was used to compute kmer coverage over each scaffold using each CSS read set. Each non-repetitive kmer in a scaffold was scored proportionally to coverage on each CSS arm and scaffolds were classified using the following set of rules:

1. Scaffolds with less than 10% of the kmers producing a vote were left as unclassified (marked as Chromosome arm "U"). These are mostly small and/or repetitive sequences.
2. Scaffolds with a top score towards a CSS set at least double the second top score were classified to the highest scoring chromosome arm.
3. Scaffolds with a top score towards a CSS set less than double the second top score were left as unclassified (marked as Chromosome arm "U", but with the two top scores and CSS sets included in the sequence name). This category contains scaffolds that are classified as combinations of the two arms from the same chromosome, probably due to imprecise identification during flow-sorting. It also contains scaffolds from regions of the genome with specific flow-sorting biases, and assembly chimeras.

#### 4.5 Sequence length and content filter

Rather than using a simple length cutoff to include scaffolds in the final assembly, a content filter was applied to the scaffolds classified into each chromosome-arm bin to ensure short scaffolds containing unique content were not excluded from the assembly. Scaffolds were sorted by length, longest first. Scaffolds longer than 5kb were automatically added to the assembly. Scaffolds between 5kb and 500bp were added from longest to smallest if 20% of the kmers in the scaffold were not already present in the assembly. Scaffolds shorter than 500bp were excluded.

#### 4.6 Assignment of scaffold identifiers

For assigned scaffolds, the arm assignment is included in the FASTA identifier. For unassigned scaffolds with more than 10% voting kmers, the highest and second highest vote is included in the FASTA identifier to indicate possible arms. Per chromosome statistics for the final classified scaffolds are given in Table S4.4.

#### 4.7 Comparison of TGACv1 scaffolds to Chapman and CSS assemblies

We compared our 3B scaffolds to 3B scaffolds from the Chapman (Chapman et al., 2015) and CSS (The International Wheat Genome Sequencing Consortium, 2014) assemblies. Although there are more scaffolds in the TGACv1 3B assembly than the Chapman 3B scaffolds, they are more contiguous and represent a much higher portion of the chromosome. To compare gene content between assemblies, the 7703 genes identified on 3B (Choulet et al., 2014) were aligned to the 3B scaffolds from each assembly using GMap (Wu and Watanabe, 2005). Genes were counted if they aligned with at least 95% identity over 80% of their length. We could align 91.2% of 3B genes to our 3B scaffolds compared to around 70% that aligned to Chapman and CSS 3B scaffolds indicating the increased completeness of our assembly.

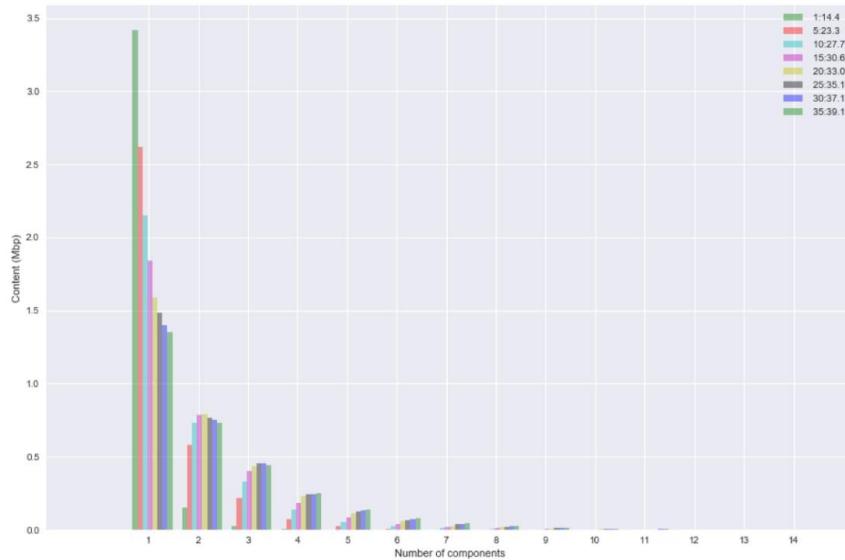
**Table S4.4:** Assembly statistics for classified scaffolds.

Arm	Total (bp)	N20	N50	N80	N%	Count
1AL	355,144,189	159,693	80,107	30,798	5.57	19,140
1AS	200,141,416	176,516	85,799	32,413	5.48	11,382
1BL	427,850,462	212,050	105,411	41,787	5.43	19,349
1BS	224,120,373	204,783	99,660	39,287	5.36	11,813
1DL	292,316,462	127,480	65,923	23,018	6.59	19,204
1DS	155,677,507	123,950	62,097	19,441	6.74	12,849
2AL	408,449,610	164,629	84,674	33,270	5.49	19,410
2AS	318,533,889	183,072	90,023	33,061	5.40	17,435
2BL	423,469,708	227,122	117,486	45,691	5.14	16,714
2BS	317,593,121	215,046	108,705	45,716	5.19	12,136
2DL	335,204,207	133,166	70,105	26,700	6.67	19,424
2DS	245,159,861	140,704	72,904	24,794	6.56	16,533
3AL	381,464,830	165,249	84,656	33,372	5.64	17,063
3AS	277,280,281	188,759	93,882	40,580	5.27	10,234
3B	789,970,040	223,860	116,546	47,041	5.13	29,090
3DL	340,636,885	136,140	68,689	24,264	6.53	22,646
3DS	228,916,862	145,224	72,644	23,143	6.42	16,817
4AL	363,230,010	179,374	89,157	33,873	5.46	18,295
4AS	276,247,067	181,019	91,272	35,335	4.98	14,167
4BL	272,849,020	240,935	127,687	58,815	4.99	7,632
4BS	310,515,948	224,543	110,746	45,899	4.90	14,697
4DL	306,806,261	171,404	80,284	28,140	6.31	18,791
4DS	171,621,745	137,248	68,499	21,787	6.30	13,021
5AL	413,139,451	161,674	81,944	33,128	5.90	18,826
5AS	231,190,161	180,634	89,316	35,125	5.14	11,705
5BL	466,173,773	207,503	107,733	43,825	5.21	19,325
5BS	182,789,732	209,845	107,461	40,181	5.16	9,793
5DL	345,449,775	130,074	65,820	23,183	7.02	23,851
5DS	173,821,965	133,804	64,345	18,898	6.58	14,481
6AL	302,563,130	168,100	85,773	33,526	5.53	14,457
6AS	264,274,034	160,498	81,455	30,863	5.68	14,315
6BL	362,924,849	203,268	110,331	45,402	5.22	13,913
6BS	299,250,616	185,879	100,360	38,835	5.51	13,349
6DL	236,649,310	143,791	71,511	24,364	6.34	16,246
6DS	178,741,401	146,601	65,202	21,073	6.62	13,586
7AL	334,861,391	184,024	92,381	37,818	5.49	13,158
7AS	259,954,140	187,229	99,434	47,521	5.56	7,777
7BL	406,571,657	203,402	107,841	45,705	5.17	15,233
7BS	287,930,109	222,106	119,366	48,224	4.95	10,813
7DL	273,279,341	135,861	69,599	23,246	6.84	18,964
7DS	303,641,845	133,599	68,218	24,284	6.63	19,510
U	680,947,588	192,507	78,842	6,368	6.58	88,799
Total	13,427,354,022	180,094	88,778	32,825	5.73	735,943

## 4.8 Assembly validation

### 4.8.1 Contig validation by mate-pair link support

To validate the contigs produced by the w2rap-contigger, we used the Nextera 11kbp mate pair library as an independent dataset, before it was incorporated into the assembly during scaffolding. We used this library to find unsupported regions in the contigs, by assessing the link support.



**Figure S4.4:** Content on contigs by number of components in the contig when splitting at breakpoints with support <percentile>:<link threshold>

The mate pair library was aligned using BWA to all contigs longer than 33kbp ( $3 \times$  the length of the library), the links were projected on each contig to obtain a measure of bridging link coverage. This reflected the amount of support at each position across the contig. Breakpoints were identified on any contig position with low link support; subsequently, the amount of sequence contained on contigs divided in different number of components and a corrected N50 for the set of broken contigs were computed. To choose link thresholds, a sample of 100 contigs was taken and the percentiles of the accumulated link distribution was computed (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 5, 10, 15, 20, 25, 30, 35th percentile); this procedure was repeated 500 times and the mean of those percentile distributions was used for the breakpoint calculations. These analyses are shown in Figures S4.4 and S4.5.

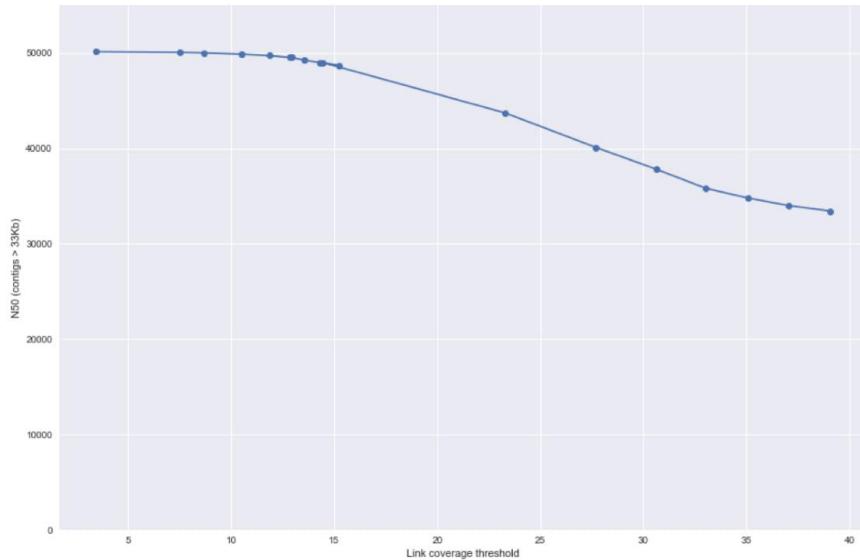
Both plots show a very small amount of affected content and a very small change of N50 in the lower percentiles of coverage, with values of linkage that are much higher than the usual accepted thresholds of 3 to 5 links. At the 1st percentile, with more than 14 links, most of the content is on single-component contigs, with only a small amount of content on 2-component contigs and negligible content on contigs broken into more components. Higher percentiles are included to show how the assembly breaks down as expected once the requirement for links is higher than the typical coverage, but we do not consider any of those thresholds to represent significant risk of misassemblies.

The code for this analysis is available on [https://github.com/bioinfologics/assembly\\_validation/tree/master/link\\_support](https://github.com/bioinfologics/assembly_validation/tree/master/link_support).

### 4.8.2 Scaffold validation by gene order between 3B and TGACv1 3B sequences

As a proxy to assess the accuracy of the scaffold linkage, we used the alignment of 3B genes to contigs and scaffolds to assess the coherence between our assembly and the 3B pseudo-molecule reference. We looked for blocks on our contigs and scaffolds where two or more genes aligned and compared the order of genes in these blocks to gene order in the 3B pseudo-molecule. In all cases, on both contigs and scaffolds we found gene order in full agreement with 3B (Table S4.5).

This provides extra evidence that at least on the genic level, our assemblies are consistent with the existing reference, with the scaffolds generating precise linkage over longer ranges.



**Figure S4.5:** N50 for the contig subset when splitting at breakpoints by link coverage threshold.

**Table S4.5:** Identification of gene blocks.

	Number of blocks	Genes in syntenic blocks	Sequence contained in blocks (Mb)
TGACv1 contigs	1,266	3,224	15.0
TGACv1 scaffolds	1,503	4,792	60.7

#### 4.9 Assessment of chromosome arm assignment accuracy

**Table S4.6:** Assessment of chromosome arm assignment accuracy.

	Genes aligned	Genes aligned to 3B classified scaffolds	Genes aligned to potential 3B scaffolds	Genes aligned to other arms
TGACv1 contigs	6,859	6,185	50	624
TGACv1 scaffolds	7,124	6,487	169	468

We aligned the genes identified on chromosome 3B (Choulet et al., 2014) to our assembled contigs and scaffolds in order to assess how accurately our algorithm assigned sequences to chromosome arms. We aligned these sequences with GMAP, using as a minimum threshold 95% alignment identity over 80% of the sequence length. We found that for contigs, 90.2% of genes aligned to 3B classified scaffolds with a further 0.7% aligning to potential 3B scaffolds (unclassified but with 3B as a suggested assignment). 9.1% aligned to other arms. For scaffolds, 91.1% aligned to 3B classified sequences, 2.4% to potential 3B scaffolds and 6.7% to other arms.

## 5 Integration with genetic maps and chromosomal alignments

To order the TGACv1 scaffolds along the wheat genome, we used the “Synthetic W7984”  $\times$  Opata M85 map, hereafter WGS map, described in Chapman et al. (2015). The WGS map was constructed using a whole genome shotgun approach on 78 double haploid (DH) lines derived from W7984/Opata F1 hybrids. In order to anchor the TGACv1 scaffolds to the WGS map, we used the 437,973 scaffolds of the W7984 assembly, which were assigned to the genetic bins of the WGS map, as markers. Given the relatively low functional population size and high number of markers, even small frequencies of scoring error will result in high rates of ordering ambiguities between markers within short genetic distances. We corrected the genetic distances between bins by iterating over the bins  $b$  and merging bins  $b_i$  and  $b'_{i+1}$  into  $b'_i$  if:

- $|b_i - b'_{i+1}| < 1.6$  recombinations (1 recombination represents 0.586cM on the WGS map)
- $b'_i$  did not span more than 2.5cM [Abraham Korol - pers. comm.]

The map position for each  $b'_i$  was calculated as the arithmetic mean of all bins merged into it. A mapping between the original WGS map bins and our corrected version can be found in (Supplementary File S4). Marker sequences were then aligned against all TGACv1 scaffolds with megablast (blast version 2.2.28, multithreaded). Only the best BLAST hit (`-max_target_seqs 1`) for each marker was taken into consideration. Markers that could be aligned equally well to more than one scaffold were discarded. BLAST hits were filtered by e-value (less than  $10 \times 10^{-10}$ ), percent identity (at least more than 98.5%), and alignment length (at least 1kbp of the marker sequence is aligned).

TGACv1 scaffolds were then anchored to the corrected WGS map by assigning them to the genetic bin of their matching marker sequences. In order to deal with ambiguous bin assignments due to multiple markers matching a scaffold, we classified the anchored scaffolds according to the following scheme:

1. *unique*: all matching markers are assigned to the same bin.
2. *ambiguous*: matching markers are assigned to the same chromosome but to different genetic bins.
3. *homoeolog*: matching markers are assigned to the same chromosome of different subgenomes.
4. *conflict*: matching markers are assigned to different, non-homeologous chromosomes.
5. *novel*: subset of class 1:unique comprising scaffolds that do not have a CSS-based chromosome arm assignment.
6. *cc\_unique*: subset of *unique*, comprising scaffolds with conflicting CSS-based and genetic map-based chromosome assignments (cc: Chapman/Clavijo conflict).
7. *cc\_ambiguous*: subset of *ambiguous*, comprising scaffolds with conflicting CSS-based and genetic map-based chromosome assignments.

The final TGACv1 map was constructed only from uniquely anchored scaffolds, i.e. scaffolds of classes 1, 5, and 6. The map is available in Supplementary File S5 and scaffold classifications for all anchored scaffolds in Supplementary File S6. Python scripts for generation of the TGACv1 map are available at <https://github.com/krasileva-group/tgac-map>.

## 6 Detection and confirmation of chromosomal translocations

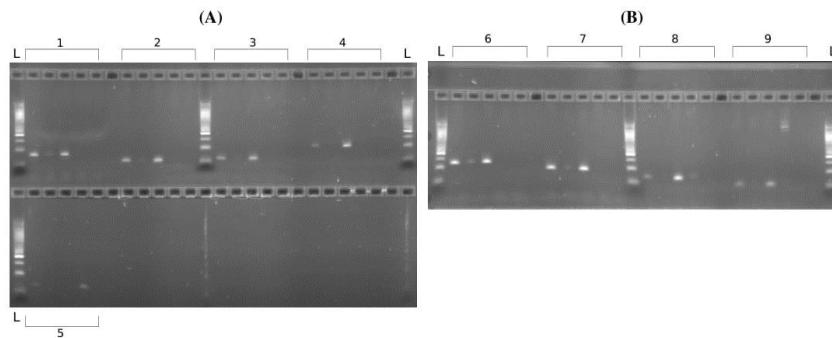
### 6.1 Detection of translocations from OrthoMCL output

Potential chromosome translocation events were identified as outlier triads of orthologous sequences (as identified by OrthoMCL, see Section 10). These triads are defined as three orthologous sequences that belong to the same OrthoMCL group, with two of the sequences being assigned to two different homoeologous chromosomes (e.g. 5B, 5D) and the third sequence, the “outlier”, to a different non-homoeologous chromosome (e.g. 4A). The translocated sequence is assumed to have moved from the missing chromosome (source) of the homoeologous triplet (5A in the example case) to the chromosome on which the outlier sequence is located (destination).

In the present analysis, we further included orthologs with multiple copies on either of the three involved chromosomes. As chromosomal translocations typically do not involve just a single gene but a whole chromosomal region and such copies could have occurred independently of a translocation, these occurrences would not prevent either of the copies (or all) to be translocated.

### 6.2 PCR assays of suspected translocations

Triads with sequences that are annotated as being transposon-associated were ignored. In order to validate these potential translocation events via PCR, primer pair candidates for the outlier sequence were designed using Polymarker (Ramirez-Gonzalez et al., 2015) without specifying marker SNPs. The candidate pairs were then checked for specificity via `blastn` (using Blast 2.2.28, multithreaded with `-task blastn-short`, `-evalue 20`, `-dust no`). Primer pairs were discarded if any off-target Blast hit with up to 3 mismatches/indels was found.



**Figure S6.1:** Gel images for gels 1–2 (panels A–B, respectively). Each gel contains five lanes per primer pair as described in the text. See Table S6.1 for details on each primer pair.

Primer pair	Gel	Lanes	Translocation type	Translocation group	Fragment length	Annealing temperature (°C)
1	1	1–5	4AL_5AL	group17899	78	55
2	1	7–11	4AL_5AL	group17588	57	55
3	1	13–17	7BS_4AL	group1175	58	55
4	1	19–23	5BL_4BS	group17187	101	55
5	1	25–29	7AL_3AL	group12953	67	55
6	2	1–5	4AL_5AL	group16803	118	60
7	2	7–11	5AL_4AL	group3295	89	60
8	2	13–17	7BS_4AL	group6850	69	60
9	2	19–23	5AL_7BS	group7391	50	60

**Table S6.1:** Primer pairs.

We tested a set of 9 genes corresponding to 3 known and 3 novel predicted translocations by PCR amplification of wild type and appropriate nullisomic lines (Sears, 1966). Five reactions were set up for each primer pair using the following template genomic DNA:

- 10ng Chinese Spring wheat
- 1ng Chinese Spring wheat
- 10ng nullisomic gDNA for the predicted source chromosome
- 10ng nullisomic gDNA for the destination chromosome predicted to receive the locus

- a negative control without DNA.

A reaction volume of 25 $\mu$ L was used with the following final concentrations:

- 1 $\times$  Flexi Buffer
- 2mM MgCl<sub>2</sub>
- 0.2mM dNTP each
- 0.5 $\mu$ M forward primer
- 0.5 $\mu$ M reverse primer
- 0.025U/ $\mu$ L of GoTaq Hot Start Polymerase

PCR was performed using a AB Verity with the following programme:

- 2min at 95°C
- 32 cycles of:
  - 30s at 95°C
  - 30s at 55°C
  - 10s at 72°C
- A final extension for 30s at 72°C

For primer pairs where non-specific bands were observed, the annealing temperature was increased from 55°C to 60°C in order to improve the stringency/specification. We ran 10 $\mu$ L of the amplicons on 4% agarose E-gels (Invitrogen) and scored PCRs that amplified the Chinese Spring control cleanly. As primers could produce some off-target amplifications (e.g. homoeologous copies) we scored departures and arrival nullisomics as negative if they produced a band at the same intensity as 1ng of Chinese Spring or lower, bands of the same intensity were scored as ambiguous. Details on all the primers and the experiments are reported in Supplementary File S7.

### 6.3 Cross-validation of translocations by using the genetic map

Overall 436 (35%) of all 1240 triads (supporting 152, or 40.75%, of 373 potential translocation events) could be anchored to the TGACv1 map (Supplementary File: S3). Of these, 416 (33.55%) triads (supporting 146 — 11.77% — potential translocation events) could be anchored without conflict between their CSS-based chromosome assignment and their genetic bin on the TGACv1 map. In 8 out of 20 conflicting triads the chromosome on the TGACv1 map is identical to the source chromosome of the potential translocation event (Table S6.2), rendering the event undetectable when relying solely on TGACv1 map information.

**Table S6.2:** Triads with conflicts between TGACv1 map and CSS chromosome arm assignment

Gene/Representative transcript	OrthoMCL group	Translocation		TGACv1 genetic bin	Note
		Source	Destination		
TRIAE_CS42_5DS_TGACv1_457137_AA1482860.1	group11550	2DL	5DS	2D:65.70	map chromosome is source chromosome
TRIAE_CS42_7BL_TGACv1_576879_AA1858370.1	group14472	3B	7BL	3B:45.71	map chromosome is source chromosome
TRIAE_CS42_7BL_TGACv1_576879_AA1858380.1	group14473	3B	7BL	3B:45.71	map chromosome is source chromosome
TRIAE_CS42_7BL_TGACv1_576879_AA1858390.1	group14474	3B	7BL	3B:45.71	map chromosome is source chromosome
TRIAE_CS42_5DL_TGACv1_436092_AA1457960.1	group15184	5AL	4AL	3A:49.69	
TRIAE_CS42_5DL_TGACv1_435790_AA1454970.1	group15512	5AL	4AL	5B:129.99	
TRIAE_CS42_5DL_TGACv1_436307_AA1459890.1	group15869	5AL	4AL	4A:106.06	
TRIAE_CS42_4DL_TGACv1_342399_AA1112520.1	group17975	4AL	5AL	6B:70.29	
TRIAE_CS42_4DL_TGACv1_342399_AA1112540.1	group17976	4AL	5AL	6B:70.29	
TRIAE_CS42_4DL_TGACv1_342399_AA1112570.1	group17977	4AL	5AL	6B:70.29	
TRIAE_CS42_6AS_TGACv1_485809_AA1552570.1	group20423	2AL	6AS	2A:82.29	map chromosome is source chromosome
TRIAE_CS42_5DL_TGACv1_433289_AA1408400.1	group22982	5AL	2AL	7B:51.03	
TRIAE_CS42_5DL_TGACv1_436092_AA1457970.1	group23416	5AL	4AL	3A:49.69	
TRIAE_CS42_5AS_TGACv1_392779_AA1264470.1	group23830	7AL	5AS	7A:63.00	map chromosome is source chromosome
TRIAE_CS42_2AS_TGACv1_112471_AA0338700.1	group3312	2BS	5BL	7A:68.56	
TRIAE_CS42_4DL_TGACv1_342436_AA1113710.2	group4786	4AL	5AL	7D:73.08	
TRIAE_CS42_2BS_TGACv1_146689_AA0470800.1	group5765	2DS	3DL	3B:48.27	
TRIAE_CS42_5AS_TGACv1_393155_AA1269210.1	group6137	3AL	5AS	3A:70.72	map chromosome is source chromosome
TRIAE_CS42_1BS_TGACv1_050024_AA0166120.1	group7241	3AL	1BS	3A:57.65	map chromosome is source chromosome
TRIAE_CS42_7DL_TGACv1_605399_AA2006440.1	group7349	7AL	4BL	4D:47.93	

## 7 Repeat analysis

Transposons were detected and classified by a homology search against the REdat\_9.7\_Triticeae section (13,229 elements, 100Mbp) from the PGSB transposon library (Spannagl et al., 2016). The program vmatch (<http://www.vmatch.de>) was used for that purpose as a fast and efficient matching tool suited for large and highly repetitive genomes with the following parameters: identity greater or equal to 70%, minimal hit length 75bp, seedlength 12bp; the exact commandline is:

```
-d -p -l 75 -identity 70 -seedlength 12 -exdrop 5
```

The vmatch output was filtered for redundant hits via a priority based approach, which assigns higher scoring matches first and either shortens (less than 90% coverage and at least 50bp rest length) or removes lower scoring overlaps to obtain an overlap free annotation.

Full-length LTR-retrotransposons elements were identified with LTRharvest (Ellinghaus et al., 2008), which reported 354,315 non overlapping candidate sequences under the following parameter settings:

```
overlaps best -seed 30 -minlenltr 100 -maxlenltr 2000 -mindistltr 3000
-maxdistltr 25000 -similar 85 -mintsd 4 -maxtsd 20 -motif tgca
-motifmis 1 -vic 60 -xdrop 5 -mat 2 -mis -2 -ins -3 -del -3
```

All candidates where annotated for PfamA domains with hmmer3 (Eddy, 2011) and stringently filtered for false positives by several criteria, the main ones being the presence of at least one typical retrotransposon domain (e.g. RT, RH, INT, GAG) and a tandem repeat content below 25%. The filtering steps led to a final set of 44,579 high confidence full-length LTR retrotransposons. The composition of repeats in the assembled genome can be observed in Table S7.1.

	% of Genome	% of TE	Number	Total (Mb)	Average length (bp)
<b>Mobile Element (TXX)</b>	81.10	100.00	9,673,829	10,268.4	1061
<b>Class I: Retroelement (RXX)</b>	67.70	83.50	7,249,022	8571.7	1182
LTR Retrotransposon (RLX)	67.30	83.00	7,171,177	8522.1	1188
Ty1/copia (RLC)	14.20	17.50	1,555,328	1792.5	1152
Ty3/gypsy (RLG)	30.80	37.90	2,971,111	3895.2	1311
Unclassified LTR (RLX)	20.80	25.60	2,621,553	2627.7	1002
non-LTR Retrotransposon (RXX)	0.40	0.50	77,845	49.6	638
LINE (RIX)	0.40	0.50	72,414	47.6	657
SINE (RSX)	0.00	0.00	5431	2.2	375
<b>Class II: DNA Transposon (DXX)</b>	12.90	15.90	2,233,197	1636.3	733
DNA Transposon Superfamily (DTX)	12.80	15.70	2,123,788	1616.6	761
CACTA superfamily (DTC)	12.40	15.30	1,951,401	1567.8	803
hAT superfamily (DTA)	0.01	0.01	1642	0.6	393
Mutator superfamily (DTM)	0.16	0.20	61,612	20.3	329
Tc1/Mariner superfamily (DTT)	0.04	0.05	37,550	5.0	134
PIF/Harbinger (DTH)	0.12	0.15	34,127	15.1	443
unclassified (DTX)	0.06	0.07	37,456	7.7	206
DNA Transposon Derivative (DXX)	0.13	0.16	102,275	16.5	162
HITE (DXX)	0.13	0.16	102,275	16.5	162
Helitron (DHH)	0.01	0.01	1965	1.5	765
unclassified DNA transposon (DXX)	0.01	0.02	5169	1.7	331
Unclassified Element (TXX)	0.48	0.59	191,610	60.3	315
<b>Retro-TE/DNA-TE ratio</b>	5.20				
<b>Gypsy/Copia ratio</b>	2.20				

Table S7.1: Repeat composition of the bread wheat genome.

## 8 Construction of the wheat gene set

The wheat gene set for wheat was generated using a custom pipeline integrating wheat-specific transcriptomic resources, including PacBio transcriptomic data, similarity to proteins of related species, and evidence-guided ab initio predictions generated with AUGUSTUS (Stanke et al., 2006).

The pipeline was divided in five different phases. In the first phase, RNA-Seq models were generated with 4 different assembly methods utilising data from multiple tissues and conditions, and integrated together with PacBio transcripts into a coherent and non-redundant set of models using Mikado (Venturini et al., 2016). In the second phase, PacBio reads were classified based on protein similarity and a subset of high quality (e.g. full length, canonical splicing, non-redundant) transcripts employed to train an AUGUSTUS wheat-specific gene prediction model. In the third phase, AUGUSTUS was used to generate a first draft of the genome annotation, using as input Mikado-filtered transcript models, reliable junctions identified with Portcullis (Mapleson et al., 2016), and peptide alignments of proteins from five different species closely related to wheat (*Brachypodium distachyon* 314 v. 3.1, *Zea mays* 284 v. 6a, *Oryza sativa* 204 v. 7.0, *Sorghum bicolor* 313 v. 3.1, and *Setaria italica* 312 v. 2.2, all downloaded from Phytozome (Goodstein et al., 2012)). In the fourth stage, this draft annotation was refined and polished by identifying and correcting probable gene fusions, missing loci and alternative splice variants. Finally, the polished annotation was functionally annotated and all loci were assigned a confidence rank based on their similarity to known proteins and their agreement with wheat transcriptomic data.

### 8.1 Reference guided transcriptome reconstruction

#### 8.1.1 Alignment of Illumina RNA-seq data

**Data preparation** RNA-Seq data from three different datasets was utilised for the annotation: ERP004714 (used for the annotation provided in The International Wheat Genome Sequencing Consortium (2014)), ERP004505 (used for the grain-development analyses in Pfeifer et al. (2014)) and an internally generated dataset of 250bp paired-end strand-specific reads from six different tissues (PRJEB15048; Table S8.1). In total, the three datasets comprised over 3.2 billion paired-end reads. For each dataset, read samples were collapsed by tissue and filtered using trim-galore v. 0.3.7 (BabrahamLab, 2014), with the command line options:

```
-q 20 --phred33 --stringency 5 --fastqc --length 60
```

Due to concerns of high concentration of ribosomal RNA in the internally produced samples, reads from that dataset were further filtered using SortMeRNA v. 2.0 (Kopylova et al., 2012), with the command line options:

```
--num_alignments 1 --fastx --paired_in
```

and using RFam (5S and 5.8S) and Silva (Archea 16S-23S, Bacteria 16S-23S, Eukariota 18S-28S) as databases.

**Alignment with STAR** Filtered reads were aligned to the wheat genome using a forked version of STAR-2.5.0-alpha (Dobin et al. (2013), commit f82c5a0028; see (<https://github.com/alexdobin/STAR/issues/85>)). The genome was indexed using the option

```
--genomeChrBinNbts 14
```

in accordance with STAR documentation, and the process had to be performed on a UV supercomputer due to the memory requirements (~2TB of RAM). Reads were aligned with stringent parameter in a two pass approach to ensure alignment accuracy, a first pass using the custom command-line options

```
--outFilterMismatchNmax 3 --alignEndsType EndToEnd
```

```
--alignIntronMin 20 --alignIntronMax 200000
```

```
--outSJfilterIntronMaxVsReadN 10000 10000 10000
```

to increase the accuracy of the alignments and

```
--outSAMattributes NH HI NM MD AS XS
```

to ensure the compatibility of the output with downstream tools such as Cufflinks (Trapnell et al., 2010). All 1,519,861 reliable junctions detected by STAR in at least one sample during this first pass were collapsed, and given as input for a second round of alignments, with the same command line parameters but also providing the merged junction file with the options:

```
--limitSjdbInsertNsJ 2000000 --sjdbOverhang 250
```

Finally, the alignments from all samples were filtered with portcullis v. 0.10.1 (Mapleson et al., 2016) to exclude spliced reads with non-canonical junctions that were on manual review identified as predominantly due to misalignment.

**Alignment with TopHat2** As the original IWGSC annotation had been created using the aligner TopHat2 (Kim et al., 2013), we also aligned reads from the ERP004714 dataset using this program. To retrieve splicing junctions related to the original annotation, IWGSC models were aligned against our reference using GMAP v. 2015-09-29 (Wu and Watanabe, 2005), with the command line options:

```
--min-identity=0.99 --min-trimmed-coverage=0.90 -n 1
```

and subsequently collapsed and filtered for models only with canonical junctions using gffread from Cufflinks v. 2.2.2beta (Trapnell et al., 2012; Roberts et al., 2011a,b). 281,562 unique splicing junctions from the aligned models were retrieved with a custom Python3 script from the surviving 85,242 models and provided to TopHat v.2.1.0 (patched to use Bowtie2.2.5 (Langmead and Salzberg, 2012) long indices; the patch was subsequently integrated into the later TopHat v.2.1.1). Reads from ERP004714 were then aligned in single pass using the CLI options

```
-a 13 -i 20 -I 400000 -g 20 --no-discordant -N 1 --read-edit-dist 1 --read-realign-edit-dist 1 --read-gap-length 1 --library-type fr-unstranded
```

and additionally providing the junction file from above.

**Table S8.1:** Sequencing reads used in this study. ERP004714: Grain, Leaf, Root, Spike and Stem, ERP004505: 10DPA, AL\_20DPA, AL.SE\_30DPA, REF\_20DPA, SE\_20DPA, SE\_30DPA and TC\_20DPA, PRJEB15048: seedling, root, leaf, stem, spike and seed.

	ERP004714	ERP004505	PRJEB15048
Number of samples	5	7	6
Number of reads	1,536,051,415	873,709,556	824,241,135
Number of filtered reads	1,412,029,174	873,550,049	731,931,657
Average no. filtered reads per sample	282,405,834.8	124,792,864.1	121,988,609.5
Aligned reads (STAR)	1,203,100,456	744,087,908	488,750,691
Aligned reads (STAR second pass)	1,267,816,403	759,278,032	579,642,183
Aligned reads (TopHat2)	1,299,830,440	NA	NA

**Table S8.2:** Number of PacBio reads, per sample and size-fraction.

Stage	Size Fraction	Leaf	Root	Seed	Seedling	Spike	Stem	Total
Reads of insert	0.7 - 2 kbps	345,566	482,417	410,969	227,253	353,196	210,462	2,029,863
	2-3 kbps	267,379	410,186	364,988	330,525	375,062	376,717	2,124,857
	3-5 kbps	367,571	356,396	301,030	110,628	311,537	370,739	1,817,901
	Total	980,516	1,248,999	1,076,987	668,406	1,039,795	957,918	5,972,621
IsoSeq + Quiver	0.7 - 2 kbps	69,817	116,164	86,031	77,211	98,848	79,909	527,980
	2-3 kbps	55,789	125,622	77,619	97,894	90,340	104,293	551,557
	3-5 kbps	73,513	73,351	56,315	34,818	88,516	103,272	429,785
	Total	199,119	315,137	219,965	209,923	277,704	287,474	1,509,322
Aligned		187,583	297,970	205,990	197,535	259,329	265,816	1,414,223
% aligned		94.21%	94.55%	93.65%	94.10%	93.38%	92.47%	93.70%

### 8.1.2 Alignment of PacBio RNA-seq data

**Data preparation** PacBio sequencing data from six tissues was analysed initially using the SMRTanalysis package (v2.3.0.140936), stopping at the quiver step. The “CircularConsensus” step of the ConsensusTools utility was called with the command-line options `--minFullPasses 0 --minPredictedAccuracy 75` while during the classification step the option `--min_seq_len 300` was invoked. The pipeline provided a total of over 1.5 million PacBio transcriptomic reads for downstream analyses (Table S8.2).

**Read alignment** PacBio reads were aligned using the gmapl utility from GMAP v. 2015-11-20 (Wu and Watanabe, 2005), with the command line options `-f 2 --no-chimera -n 1 --min-trimmed-coverage=0.90 --min-identity=0.95 --split-output`. We further discarded alignments deemed to be translocations by GMAP (those reported in the .transloc file).

### 8.1.3 Transcript assembly

The illumina RNA-Seq alignments (18 from STAR and 5 from TopHat2) were assembled by tissue/condition using three different tools: CLASS v. 2.12 (Song et al., 2016), Cufflinks v. 2.2.2 beta (commit 753c109e31; Trapnell et al. (2010); Roberts et al. (2011a,b)) and StringTie v.1.10 (Pertea et al., 2015). CLASS was called using the option `-F 0.05`; Cufflinks was invoked asking to limit the intron size to 200,000 and using both the fragment-bias correction and the multi-read rescue method:

`-I [200000] -b -u`

Samples from the internal dataset were assembled using also the option:

**Table S8.3:** Illumina and PacBio transcript assembly statistics. For each tool, assembled transcripts have been clustered into loci using `cuffcompare` (v.2.2.1, command line options “`-C -G`”; Trapnell et al. (2010))

Method	Loci	Transcripts	Average number of exons	Average cDNA size	Number of monoexonic transcripts
CLASS	181,259	3,188,679	5.48	1,304.55	326,210
Cufflinks	270,456	3,281,661	4.37	1,595.44	1,078,721
StringTie	285,728	3,826,431	4.47	1,554.83	1,117,717
Trinity	244,384	646,244	2.96	1,301.02	333,428
PacBio (4 samples)	81,752	1,020,650	6.80	2,109.06	131,357
PacBio (all 6 samples)	88,609	1,330,372	6.79	2,100.97	173,661

**Table S8.4:** Mikado transcript assembly statistics.

	Genes	Transcripts	Average number of exons	Average cDNA size	Number of monoexonic transcripts
Mikado (4 PacBio)	81,848	120,886	6.36	2,098.83	18,554
Mikado (6 PacBio)	83,144	128,030	6.29	2,182.37	19,175
Mikado (Illumina and PacBio)	273,243	373,861	4.07	1,377.70	93,564

--library-type fr-firststrand

StringTie was invoked asking for assemblies longer than 200bp ("–m 200"). In addition the alignments of reads from the internal dataset (6 tissues) were merged using the MergeSamFiles utility from picard (Wysokar et al., 2016). The merged BAM file was used as input for Trinity v.2.1.1 (Haas et al., 2013) in genome-guided mode, using the command line options:

--SS\_lib\_type RF --genome\_guided\_max\_intron 200000

The assembled transcripts were then aligned against the genome using gmap from GMAP v. 2015-11-20 (Wu and Watanabe, 2005), using the command line options:

-f 2 --min-trimmed-coverage=0.80 --min-identity=0.90

Uniquely and multiply mapping transcripts were further filtered using a custom python3 script to retain only those alignments in which the assembled transcript mapped against the same region from which its original read cluster originated from. The number and features of transcripts detected by each method is reported in Table S8.3.

We used Mikado (Venturini et al., 2016) to integrate the ~11 million Illumina assemblies generated by multiple assembly tools (CLASS, Cufflinks, StringTie, Trinity) and ~1.4 million aligned PacBio reads. Mikado leverages transcript assemblies generated by multiple methods to improve transcript reconstruction. Loci are first defined across all input assemblies with each assembled transcript scored based on metrics relating to ORF and cDNA size, relative position of the ORF within the transcript, UTR length and presence of multiple ORFs. The best scoring transcript assembly is then returned along with additional transcripts (splice variants) compatible with the representative transcript.

We generated three Mikado selected transcript sets for use in gene predictor training or annotation (Table S8.4):

1. Alignments from 4 PacBio samples (Root, Seedling, Spike, Stem) were analysed with Mikado 0.11.0, without BLAST data and disabling the "chimera\_split" algorithm. The transcript set was used in gene predictor training.
2. Mikado (v. 0.19.2) run on the full set of 6 PacBio samples, with BLAST data, and enabling the chimera\_split option in "PERMISSIVE" mode.
3. The 70 RNA-Seq assemblies (23 alignments \* 3 assemblers + Trinity) and PacBio alignments (Root, Seedling, Spike, Stem) were analysed using Mikado v. 0.18.0 with the "chimera\_split" option set to PERMISSIVE.

For Mikado runs incorporating BLAST data transcripts passing the "prepare" step were blasted against filtered and masked proteins of *B. distachyon*, *O. sativa*, *S. bicolor*, *S. italica* and *Z. mays* using BLAST+ v. 2.2.30 and limiting each result to the best 15 matches.

## 8.2 Gene predictor training

The primary PacBio alignments from 4 samples (Root, Seedling, Spike, Stem) analysed with Mikado 0.11.0 were filtered for full-length complete and coding transcripts using Full-lengtherNEXT (v0.0.8; Fernandez and Guerrero (2012)) with open reading frames (ORFs) predicted using TransDecoder v2.0.1 (Grabherr et al., 2011). A reliable set of transcripts were selected for training AUGUSTUS having single full length ORF, with 5' and 3' UTR present, consistent Full-lengtherNEXT and TransDecoder CDS coordinates, a minimum CDS to transcript ratio of 50% and a single transcript per gene. We excluded genes with a genomic overlap within 1000bp of a second gene and gene models that are homologous to each other with a coverage and identify of 80%. The filtered PacBio set contained 9952 transcripts selected for training AUGUSTUS. The trained AUGUSTUS model resulted in 0.941 sn, 0.844 sp nucleotide level, 0.798 sn, 0.756 sp exon level and 0.455 sn, 0.367 sp at the gene level.

## 8.3 Gene prediction using evidence guided AUGUSTUS

Protein coding genes were predicted using AUGUSTUS (Stanke et al., 2006) by means of a Generalized Hidden Markov Model (GHMM) that takes both intrinsic and extrinsic information into account.

### 8.3.1 Generation of external hints for gene prediction

**Junctions** RNA-Seq junctions (defining introns) were derived from RNA-Seq alignments (From TGAC: Leaf, Stem, Spike, Seed, Seedling and Root samples; From accession ERP004505: 10DPA, AL\_20DPA, AL\_SE\_30DPA, REF\_20DPA, SE\_20DPA, SE\_30DPA and TC\_20DPA samples; From accession ERP004714: Grain, Leaf, Root, Spike and Stem samples), using portcullis v.0.12.0 (Mapleson et al., 2016) and the default set of filtering parameters. Junctions that pass and fail the portcullis filter were classified as Gold and Silver respectively.

**Table S8.5:** Description of reference protein datasets used with AUGUSTUS (Stanke et al., 2006). Proteins were filtered at 50% identity and 80% coverage and junctions checked against the Illumina junctions as an additional filtering criterion. Any intron over 50kb resulted in the protein alignment being removed.

	<i>B. distachyon</i>	<i>O. sativa</i>	<i>S. bicolor</i>	<i>S. italica</i>	<i>Z. mays</i>
Total Proteins	52,972	49,061	47,205	43,001	88,760
Proteins Aligned	30,354	23,929	23,231	23,107	38,653
Proteins Aligned (%)	57.30%	48.77%	49.21%	53.74%	43.55%
Protein Alignments	105,190	89,739	83,561	86,381	142,217

**Proteins** Protein sequences from 5 species (*Brachypodium distachyon*, *Oryza sativa*, *Sorghum bicolor*, *Setaria italica* and *Zea mays*) were soft masked for low complexity (segmasker from NCBI BLAST+ 2.3.0) and aligned to the soft masked genome (using PGSB repeats) with exonerate v2.2.0 (Slater and Birney, 2005) with parameters:  
`--model protein2genome --softmaskquery yes --softmasktarget yes --bestn 10 --minintron 20`

To identify a high confidence set of alignments, exonerate results were filtered at 50% identity and 80% coverage. Furthermore, alignments whose introns were either longer than 50kbps or that were not present in the set of Illumina RNA-Seq junctions were removed from further analysis (see Table S8.5).

**PacBio transcript classification** To generate high confidence evidence hints for gene prediction, Mikado filtered PacBio transcripts (Root, Seedling, Spike, Stem) were classified into the following three categories:

**Gold** : PacBio reads having a full length hit (complete/putative complete) with Full-LengtherNEXT and having a maximum of 2 complete 5'UTR exons and 1 complete 3'UTR exon;

**Silver** : Remaining models meeting the maximum 5'UTR and 3'UTR restrictions with an additional constraint of having at least 900bp CDS length;

**Bronze** : any remaining Mikado PacBio transcripts were assigned to the bronze category.

In addition, polished (Quiver high and low quality filtered) PacBio reads were filtered for splice sites that are concordant with Illumina RNA-Seq alignments and were used along with other evidences for the gene prediction.

**Classification of Mikado transcripts** The Mikado models (combining Illumina and PacBio assemblies) were classified into the following three categories:

**Gold** : Mikado transcripts having a full length hit (complete/putative complete) with Full-LengtherNEXT and having a maximum of 2 complete 5'UTR exons and 1 complete 3'UTR exon;

**Silver** : Remaining models meeting UTR restrictions with an additional constraint of having at least 300bp CDS length;

**Bronze** : Any remaining Mikado transcripts were assigned to bronze category if they had a maximum intron length of 50kbp.

**RNA-seq coverage hints** Individual RNA-Seq bam files from STAR were merged together and reads were extracted from merged bam using picardtools (SamToFastq.jar v1.84; Wysokar et al. (2016)). The extracted PE reads were then normalised using a Trinity utility (v2.0.2; Grabherr et al. (2011));

`insilico_normalization.pl --max\_cov 50 --pairs_together --KMER_SIZE 25`  
and were used to create the normalised bam with picardtools (FilterSamReads.jar v1.84; Wysokar et al. (2016)). The wig file was generated using RSeQC v2.3.7 (bam2wig.py; Wang et al. (2012)) and then converted to a hints file using a utility provided with AUGUSTUS (v2.7; (Stanke et al., 2006));  
`wig2hints.pl --width=10 --margin=10 --minthresh=2 --minscore=4 --prune=0.1 --radius=4.5`

### 8.3.2 Gene prediction

AUGUSTUS (v2.7) was used to predict gene models for the Wheat CS42 TGACv1 genome assembly by utilising the evidence hints generated from five sets of cross species protein alignments, PacBio models, Mikado PacBio models, PacBio plus Illumina Mikado models and RNA-Seq junctions (defining introns). Interspersed repeats were provided as “nonexonpart” hints and RNA-Seq read density was provided as “exonpart” hints. We assigned higher bonus scores and priority based on evidence type and classification (Gold, Silver, Bronze) to reflect the reliability of different evidence sets (see supplementary AUGUSTUS config file **S8**); Statistics of the generated models are presented in Table S8.6.

**Table S8.6:** AUGUSTUS gene prediction statistics.

Gene Count	224,994
Total transcripts	224,994
Transcripts per gene	1
Transcript mean size (incl. intron) (bp)	3547.89
Transcript mean size cDNA (bp)	1447.66
Transcript median size cDNA (bp)	1239
Min cDNA	8
Max cDNA	15,613
Total exons	833,929
Exons per transcript	3.71
Exon mean size (bp)	390.58
Total exons (distinct)	827,714
Exon mean size (distinct) (bp)	392.09
CDS mean size (bp)	302.18
CDS mean size (distinct) (bp)	302.22
Transcript mean size CDS (bp)	959.71
Transcript median size CDS (bp)	747
Min CDS	3
Max CDS	14,259
5UTR mean size (bp)	154.03
5UTR mean size (distinct) (bp)	153.96
3UTR mean size (bp)	249.69
3UTR mean size (distinct) (bp)	249.73

#### 8.4 Gene model refinement

The primary gene models generated by AUGUSTUS were corrected to remove long terminal introns spanning over 10kbp, identified from manual review as likely artefacts. To identify incorrectly split genes, AUGUSTUS gene models were compared against the high quality Mikado PacBio Gold and Silver set of gene models to identify cases where more than one AUGUSTUS model was contained within a PacBio model with at least 80% nucleotide precision (specificity), in which case we retained only the AUGUSTUS gene model with the highest nucleotide F1.

To add reliable alternative splice variants we ran PASA (Haas, 2003) with a filtered set of transcripts, removing from Mikado transcripts and PacBio reads those which had introns greater than 10kb, and retaining PacBio splice junctions that were consistent with RNA-Seq Illumina alignments. Transcripts were integrated into the annotation via a PASA utility:

```
validate_alignments_in_db.db --MIN_INTRON_LENGTH=20 --MAX_INTRON_LENGTH=50000  
--MIN_PERCENT_ALIGNED=70 --MIN_AVG_PER_ID=95 --NUM_BP_PERFECT_SPLICE_BOUNDARY=3
```

A second round of updates to the annotation was generated with PASA assemblies constructed from only PacBio reads. To identify and correct gene annotation artefacts, any incorrectly fused PASA models were replaced with a PacBio Gold gene model when the latter was found to overlap with a nucleotide recall of at least 30%. PASA transcripts associated with the incorrectly fused PASA gene but not found to overlap with the PacBio Gold gene model were clustered into new loci and retained. Transcript models with cDNAs shorter than 300bp were removed from further analysis.

#### 8.5 Assignment of gene biotypes and confidence classification

Gene models were classified as coding, non-coding and repeat associated and assigned as high or low confidence based on support from cross species protein similarity and wheat transcripts.

We decided to assign a confidence ranking to each transcript, in three levels:

**Protein ranking** : this rank is based on similarity - or lack thereof - of the transcript against publicly available protein datasets. The rankings go from 1 (best) to 5 (worst).

**Transcript ranking**: this rank is based on support for the model - or lack thereof - from our multiple sources of transcriptomic evidence. The rankings go from 1 (best) to 5 (worst).

**Confidence**: we assigned a general binary confidence tag ("High" vs "Low") for each transcript. To qualify to be considered a high-confidence *coding* transcript, a model has to fall in one of the following categories:

- Protein ranking P1 and transcript ranking T4 or better
- Protein ranking P2 and transcript ranking T4 or better
- Protein ranking P3 and transcript ranking T1

### 8.5.1 Cross species protein similarity ranking

Each gene model was assigned a protein rank (P1–P5) reflecting the level of coverage of the best identified homolog in a plant protein database. Protein ranks were assigned as:

**Protein Rank 1 (P1)** : proteins identified as full length in Full-LengtherNEXT with the UniProt database or at least 80% coverage in a supplementary BLAST database consisting of *A.thaliana*, *B. distachyon*, *O. Sativa*, *S. bicolor*, *S. italica* and *Z. mays* proteins

**Protein Rank 2 (P2)** : proteins with at least 60–80% coverage in the supplementary BLAST database;

**Protein Rank 3 (P3)** : proteins with at least 30–50% coverage in the supplementary BLAST database;

**Protein Rank 4 (P4)** : proteins with a low coverage hit (between 0–30%) in the supplementary BLAST database;

**Protein Rank 5 (P5)** : proteins with no hit in the supplementary BLAST database.

### 8.5.2 Wheat transcript support ranking

A transcript rank (T1–T5) was assigned based on the extent of support for the predicted gene model from either wheat PacBio reads or assembled wheat RNA-Seq data (all 10,943,015 transcripts assembled from all four transcript assembly methods).

We calculated a variant of annotation edit distance (*AED*) and used this to determine a transcript level ranking. First we define accuracy *AC* as:

$$AC = (SN + SP)/2$$

where *SN* is sensitivity and *SP* specificity, and then derived the *AED*:

$$AED = 1 - AC.$$

Rather than taking the union of all transcript evidence, we calculate *AED* at base, exon and splice junction level against all individual wheat transcripts used in our gene build (Illumina assemblies, cDNAs and PacBio reads), we then take the mean of base, exon and junction *AED* based on the transcript that best supported the gene model. *AED* statistics were calculated using the compare utility from Mikado (Venturini et al., 2016).

Transcript ranking was assigned based on:

**Transcript Rank 1 (T1)** : Full length support from cDNA or Pacbio read;

**Transcript Rank 2 (T2)** : full length support from Illumina assemblies;

**Transcript Rank 3 (T3)** : Best average *AED* less than 0.5;

**Transcript Rank 4 (T4)** : Best average *AED* between 0.5 and 1;

**Transcript Rank 4 (T5)** : No transcriptomic support (best average *AED* = 1).

### 8.5.3 Assignment of a locus biotype

Following the assignment of protein and transcript rankings, we assigned a locus biotype to each gene.

**Repeat associated biotypes** Genes were classified as repeat associated if all their transcripts aligned with at least 20% similarity and 30% coverage to the TransposonPSI library (v08222010; Haas (2010)) and had at least 40% coverage by PGSB interspersed repeats. In addition, genes with transcripts that had at least 20% similarity and 50% coverage to the TransposonPSI library or had at least 60% coverage by the PGSB interspersed repeats were also classified as repeat associated. In order to reduce the number of false positive calls, the combined set of putative repetitive transcripts identified above were further checked using a BLAST dataset (comprising protein sequences from *A. thaliana* TAIR10.31, *B. distachyon* v3.1, *H. vulgare* v1.31, *O. sativa* v7.0, *S. bicolor* v3.1, *S. italica* v2.2 and *Z. mays* v6a, all from Phytozome) filtered specifically for repeats, by excluding any sequence corresponding to one of the following parameters:

- Protein with a match for “retrotransposon”, “transposon” or both in their description
- At least 30% similarity and 60% coverage to a hit in TransposonPSI

Any assignment of repeat-associated status was judged a false positive call if the protein had a hit with at least 30% coverage against the filtered protein dataset above.

**Non-coding RNAs** Genes where all the transcript had a protein rank of P4 or P5 were checked to verify whether they could constitute putative non-coding RNAs. Transcript sequences were analysed with CPC v. 0.9.2 (Kong et al., 2007) in conjunction with Uniref90 from Uniprot (retrieved on 11th March 2016). Transcripts were called as putative non-coding RNAs if they met the following conditions:

- PR4 and CPC score lower or equal than -1
- PR5 and CPC score lower than 0

**Table S8.7:** Rankings and confidence of coding transcripts.

Protein Rank	Transcript Rank	Confidence	Transcript Count
P1	T1	High	66404
	T2	High	43423
	T3	High	20937
	T4	High	10013
	T5	Low	21469
P2	T1	High	3461
	T2	High	3545
	T3	High	3392
	T4	High	2084
	T5	Low	6213
P3	T1	High	1813
	T2	Low	4521
	T3	Low	3995
	T4	Low	3406
	T5	Low	12210
P4	T1	Low	781
	T2	Low	3116
	T3	Low	2846
	T4	Low	2494
	T5	Low	7484
P5	T1	Low	2079
	T2	Low	4638
	T3	Low	3944
	T4	Low	2915
	T5	Low	12364

**Protein-coding genes** Genes not assigned as non-coding were classified as protein coding; all the transcripts associated with them were assigned the same biotype.

#### 8.5.4 Removal of spurious genes

After assigning a biotype to each gene, we performed a final polish of the annotation by marking for removal loci where all the transcripts met the following criteria:

- Putative non-coding transcripts lacking transcript support (TR5)
- Putative coding transcript lacking transcript and protein similarity support (TR5,PR5)
- Protein coding transcripts harbouring an in-frame stop-codon

Before discarding these transcripts, we performed an expression estimation against all of our samples using Kallisto v 0.42.5 (Bray et al., 2016); in parallel, we aligned all high-confidence protein coding transcripts from the previous annotation (The International Wheat Genome Sequencing Consortium, 2014) using GMAPL v. 2015-11-20 (Wu and Watanabe, 2005) and asking for the best match with coverage over 90% and identity over 95% (excluding chimeric alignments). Genes were retained if one of their transcripts met at least one of the following conditions:

- Expression level over 0.5 TPM in at least one of our samples, as measured by Kallisto
- BLAST hit from the Full-LengtherNEXT analysis with the UniProt database.
- Match against the IWGSC set, with AED lower than 1, as measured by Mikado compare

Any gene whose transcripts were all marked for removal, even after these last checks, was excluded from the final annotation. Table S8.7 reports the final number of coding transcripts per each rank.

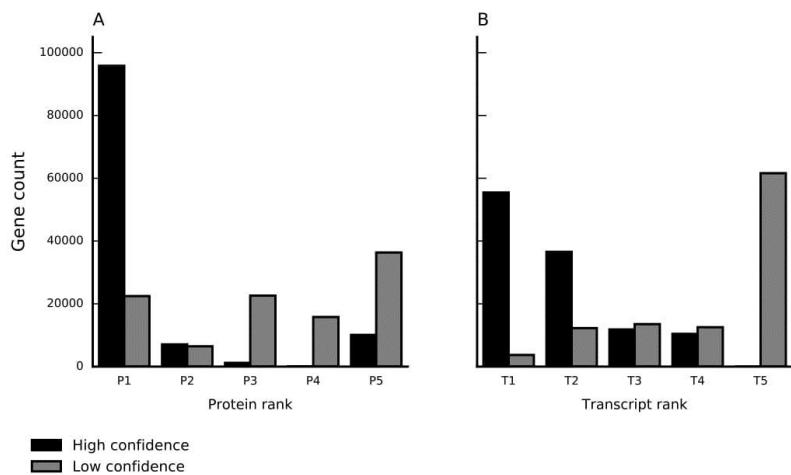
#### 8.5.5 Assignment of high and low confidence tags

Based on the above ranking, gene models were classified as high and low confidence as follows:

A **High confidence (biotype Protein\_coding)** - any protein coding gene where any of its associated gene models meet the following criteria:

**Table S8.8:** TGACv1 annotation biotype and gene confidence assignment.

Confidence Level	Biotype	Gene Count
High	protein_coding	104091
High	ncRNA	10156
Low	Protein_coding_repeat_associated	8556
Low	protein_coding	83217
Low	ncRNA_repeat_associated	1954
Low	ncRNA	9933



**Figure S8.1:** Assessment of confidence rankings for the protein coding portion of the wheat gene set. Protein (A) and transcript (B) classification for high and low confidence genes (gene level) based on classification of the representative gene model.

- PR1 and TR1 to TR4
- PR2 and TR1 to TR4
- PR3 and TR1

**B Low confidence (biotype Protein\_coding):** any protein coding gene where all of its associated transcript models do not meet the criteria to be considered as high confidence protein coding transcripts.

**C High confidence (biotype ncRNA):** any ncRNA gene where any of its associated gene models meet the following criteria:

- TR1
- TR2

**D Low confidence (biotype ncRNA):** any ncRNA gene where all of its associated transcript models do not meet the criteria to be considered as high confidence non-coding transcripts.

**E Low confidence (biotype Protein\_coding\_Repeat\_associated, ncRNA\_Repeat\_associated)** all repeat associated genes are classed as low confidence.

This classification defines four locus biotypes (protein\_coding, ncRNA, protein\_coding\_repeat\_associated and ncRNA\_repeat\_associated) and two locus level confidence classifications: “high” or “low”. Transcript classifications were harmonised within each gene so that each of them only harbours transcripts of one classification, following the order of rankings in the list above.

The number of genes within each category can be found in Table S8.8, and a graphical summary of the genes associated with each protein and transcript ranking can be found in Figure S8.1.

#### 8.5.6 Assignment of a representative gene model

We assigned a representative model for a gene by selecting a model with the highest confidence ranking (as described in Table S8.7, where a rank 1 is greater than a rank 5 model, i.e., PR1 is better than PR5, TR1 is better than TR5) and lowest *AED* by keeping the order:

1. highest protein rank
2. highest transcript rank
3. lowest *AED*.

For ncRNA genes, we assigned the representative model by considering the order:

1. highest transcript rank
2. lowest *AED*.

We compiled a summary of the annotation statistics in Table 3 of the manuscript.

#### 8.5.7 Assessment of the TGACv1 annotation

**Comparison with *B. distachyon* models.** We assessed the coherence in gene length between a selected set of TGACv1 *Triticum aestivum* and *Brachypodium distachyon* genes. We have downloaded 2707 *Brachypodium distachyon* proteins identified as single copy in *Brachypodium distachyon*, *Oryza sativa*, *Sorghum bicolor*, *Setaria italica* and *Zea mays* from Phytozome 11 (BioMart URL link: <https://go0.g1/8Ujnkj>). The *B. distachyon* proteins were blasted (ncbi-blast-2.3.0+, maximum evalue  $1 \times 10^{-5}$ ) against TGACv1 *T. aestivum* proteins and the reciprocal best hit was selected using a custom perl script. A high coherence in gene length was found between *B. distachyon* proteins and TGACv1 *T. aestivum* proteins (Figure S8.2).

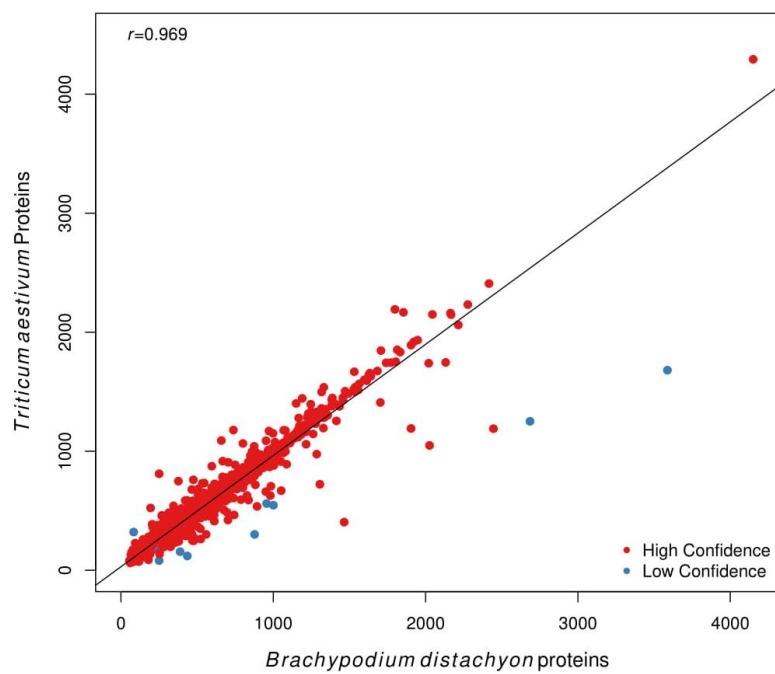
**Reconstruction of the gene space in multiple *T. aestivum* assemblies.** We assessed how completely the “gene space” was represented in TGACv1 relative to publicly available wheat assemblies by aligning the 1,509,322 PacBio transcripts to each assembly (minimum 95% identity; Figure S8.3). Of the PacBio transcripts 93% could be aligned with greater than 90% coverage to TGACv1, 19% more than to the synthetic W7984 assembly (74%; Chapman et al. (2015)).

**Comparison with IWGSC gene models** We compared the previous annotation with ours (The International Wheat Genome Sequencing Consortium, 2014; Choulet et al., 2014) by aligning the gene models onto our assembly with GMAPL (version 2015-11-20; Wu and Watanabe (2005)) with the following command line options:

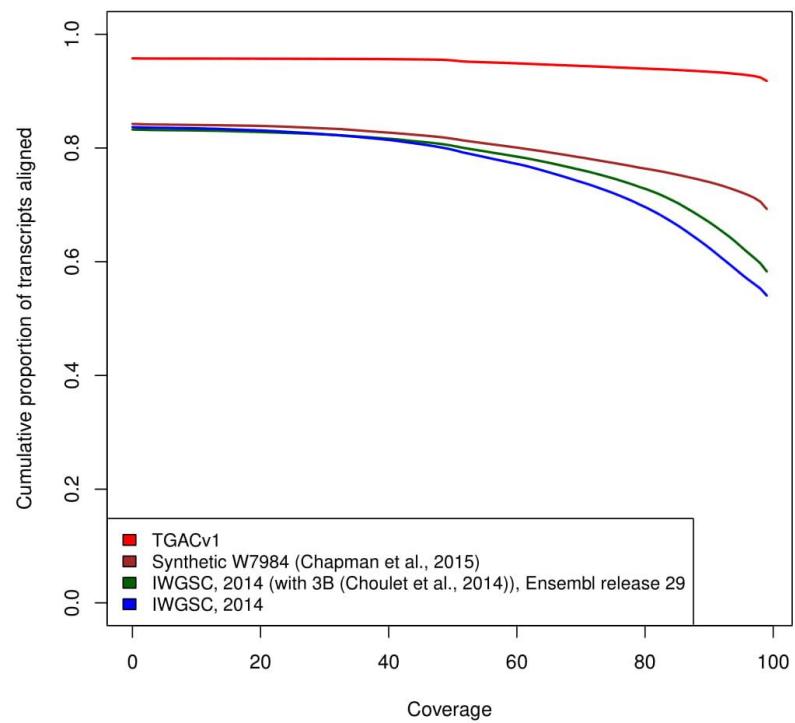
```
gmapl --no-chimeras -n 1 -f 2 --min-trimmed-coverage=0.90 --min-identity=0.95
```

The alignment has been effectuated separately for the high confidence genes and the low confidence set. The alignments were compared against our annotation with Mikado compare (v. 0.22.0; Venturini et al. (2016)), and binned into four different classes:

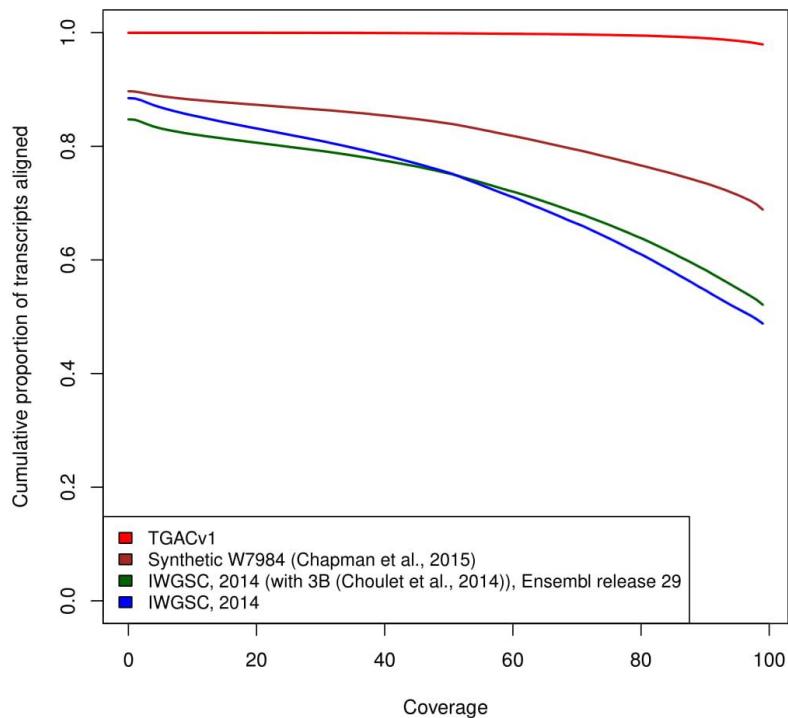
1. TGAC model missed (class code in the refmap file: NA, X, x, P, p, i, I, ri, rI, u).



**Figure S8.2:** Coherence in gene length between *Triticum aestivum* and *Brachypodium distachyon* proteins. Blast analysis ( $1 \times 10^{-5}$ ) identified 2686 proteins that had reciprocal best hits to 2707 *Brachypodium distachyon* proteins identified as single copy in *B. distachyon*, *O. sativa*, *S. bicolor*, *S. italica*, *Z. mays* (Phytozome). A high coherence in gene length was found between *Triticum aestivum* and *Brachypodium distachyon*, with a correlation coefficient  $r$  equal to 0.969.



**Figure S8.3:** Assessment of gene content in different wheat assemblies. PacBio transcripts (1,509,322) were aligned with GMAP (version 2015-11-20; Wu and Watanabe (2005)) to TGACv1 and three public assemblies. The plot shows cumulative proportion of aligned sequences in each assembly.



**Figure S8.4:** Assessment of TGACv1 gene content in public wheat assemblies. TGACv1 transcripts were aligned with GMAP (version 2015-11-20) to TGACv1 and three public assemblies. The plot shows cumulative proportion of aligned sequences in each assembly.

2. Structural difference between the TGAC model and the IWGSC model (class codes in the refmap file: f, j, J, n, h, O, C, mo, m, o, e).
3. IWGSC contained within the TGAC model (class codes in the refmap file: c).
4. Concordance between the two annotations (class codes in the refmap file: =, \_)

Results are reported in Figure 3 of the manuscript.

To assess how much of the TGACv1 gene content was contained in other publicly available wheat assemblies we aligned TGACv1 genes and assessed the proportion of TGACv1 models aligned relative to alignment coverage (Figure S8.4).

#### 8.5.8 Evaluation of non-coding RNAs

**Comparison with coding models in *T. aestivum*** We extracted the GFF3 of the 10,156 high-confidence ncRNA genes of the TGACv1 annotation using the grep utility from Mikado v0.24.0; only representative transcripts for each gene were retained. Likewise, we extracted the GFF3 of all coding genes (both high and low confidence). Mikado compare was then used to find the best match for each entry in the former GFF in the latter one. For the purposes of this evaluation, class codes in the TMAP file of u,p and P were considered as intergenic, X and x as matches on the opposite strand, and finally i and I as intronic.

**Alignment against the genomes of progenitors** We downloaded the genomes of two progenitors of *Triticum aestivum*, *Triticum urartu* and *Aegilops tauschii*, from Ensembl plants release 32. The representative transcripts of the 10,156 high-confidence ncRNA genes of the TGACv1 annotation were aligned against each of these genomes using GMAP v2015-11-20 (Wu and Watanabe, 2005), with the command line options:

```
gmap --no-chimeras -n 5 -f 2 --cross-species
```

The matches were then extracted from the GFF files, filtered for hits with identity and coverage greater than 90%, and merged into a unique list.

## 8.6 Alternative splicing analysis

RNA-Seq reads generated via the Illumina platform are often too short to cover a full transcript and unambiguously link alternative 5' and 3' splicing events. Furthermore, mapping of relatively short (100–300bp) reads can lead to misalignment and the identification of a substantial number of false positive splice junctions (Sturgill et al., 2013). With different assembly methods showing considerable variation in the number and structure of transcripts assembled we chose to take a conservative approach to annotating alternative splicing in the TGACv1 gene set, giving greater emphasis to long PacBio reads and excluding transcripts with severely truncated coding sequences. To provide a more comprehensive representation of alternative splicing we subsequently integrated transcripts assemblies generated from six strand specific Illumina libraries (Table S8.1, BioProject accession number PRJEB15048). RNA-Seq transcript assemblies were generated from the six samples using cufflinks (v2.2.1) and subsequently merged via cuffmerge (Roberts et al., 2011b), the TGACv1 gene models were provided as reference annotation. The merged transcripts assemblies were filtered to contain transcripts that are novel isoforms to the TGACv1 annotation, i.e. share at least one splice junction with the reference transcript. Splice variants identified from this additional analysis are provided as a separate track in the Ensembl wheat browser [http://plants.ensembl.org/Triticum\\_aestivum](http://plants.ensembl.org/Triticum_aestivum), and can be retrieved from the Earlham Institute server (see Section 8.8) In order to analyse different alternative splicing events and to identify transcripts that are susceptible to nonsense mediated decay (NMD), a bioconductor package, spliceR (Vitting-Seerup et al., 2014), was used with the output generated from running cuffdiff (Trapnell et al., 2012).

## 8.7 Functional annotation of protein coding transcripts

All the proteins of our annotation were annotated using AHRD v.3.1 (Hallab et al., 2014). Sequences were blasted against TAIR10 *A. thaliana* protein sequences (Lamesch et al., 2012) and the plant sequences of UniProt v. 2016\_05, both SwissProt and TREMBL datasets (The UniProt Consortium, 2014). Proteins were BLASTed using BLASTP+ v. 2.2.31 asking for a maximum e-value of 1. We adapted the standard example configuration file pathtest/resources/ahrd\_example\_input.yml, distributed with the AHRD tool, changing the following apart from the location of input and output files:

1. we included the GOA mapping from uniprot,
2. The regular expression used to analyse the TAIR header was amended to correct a parsing error to:

```
^>(?<accession>[aA][tT][0-9mCc][gG]\d+(\.\.\d+)?)\s+\|\| Symbols  
:[^\|\|]+\|\|\s+(?<description>([^\|\|]+))(\s*\|\|.*)?\$
```

Concurrently, we analysed the same set of sequences using InterProScan 5.18.57 (Jones et al., 2014). A custom Perl script was used to integrate the ranking, biotype, and functional classification from both tools into a unified file available at: [http://opendata.earlham.ac.uk/Triticum\\_aestivum/TGAC/v1/annotation/Triticum\\_aestivum\\_CS42\\_TGACv1\\_scaffold.annotation.gff3.functional\\_annotation.tsv.gz](http://opendata.earlham.ac.uk/Triticum_aestivum/TGAC/v1/annotation/Triticum_aestivum_CS42_TGACv1_scaffold.annotation.gff3.functional_annotation.tsv.gz).

## 8.8 Data Access

Sequencing reads generated for this study have been submitted to the European Nucleotide Archive under the accession code PRJEB15048. The annotation is available in Ensembl Plants genomic repository (release 32) at [http://plants.ensembl.org/Triticum\\_aestivum](http://plants.ensembl.org/Triticum_aestivum) and from the Earlham Institute server at [http://opendata.earlham.ac.uk/Triticum\\_aestivum/TGAC/v1/annotation](http://opendata.earlham.ac.uk/Triticum_aestivum/TGAC/v1/annotation). The latter repository contains the following files:

- TGACv1 annotation, in GFF3 format:
  - *Triticum\_aestivum\_CS42\_TGACv1\_scaffold.annotation.gff3.gz*
- Sequences for the transcript models of TGACv1 cDNAs, CDS and proteins:
  - *Triticum\_aestivum\_CS42\_TGACv1\_scaffold.annotation.gff3.cdna.fa.gz*
  - *Triticum\_aestivum\_CS42\_TGACv1\_scaffold.annotation.gff3.cds.fa.gz*
  - *Triticum\_aestivum\_CS42\_TGACv1\_scaffold.annotation.gff3.pep.fa.gz*
- Functional annotation of TGACv1 models:
  - *Triticum\_aestivum\_CS42\_TGACv1\_scaffold.annotation.gff3.functional\_annotation.tsv.gz*
- Annotation of alternative splicing events (see Section 8.6), in both GFF3 and GTF format:
  - *Triticum\_aestivum\_CS42\_TGACv1\_scaffold.annotation.AS.gff3.gz*
  - *Triticum\_aestivum\_CS42\_TGACv1\_scaffold.annotation.AS.gtf.gz*

## 9 Proteomics

Proteome profiling was conducted through reanalysis of Duncan et al. (2017). Briefly, organ and developmental stage samples were collected from both field and lab grown *Triticum aestivum* cv. Wyalkatchem. Frozen samples were crushed using mortar and pestle before protein extraction with the chloroform / methanol procedure (Wessel and Flügge, 1984) prior to tryptic digestion. A peptide level prefractionation was performed according to Yang et al. (2012) before reversed phase C18 LC/MS analysis on an Agilent 6550 Q-ToF. Spectra were matched against the combined high and low confidence protein coding peptide sequence set (249,547 sequences) with CometUI (2016.01 rev. 2; Eng et al. (2013)) precursor tolerance +/- 50 ppm, variable oxidation of methionine, fixed carbamidomethyl C. Results were validated through the Trans-Proteomic Pipeline, with the tools peptide and protein prophet (TPP v4.8.0; Deutsch et al. (2010)). A 2% peptide level FDR cutoff was calculated through the inclusion of reversed decoys of the protein sequences. Peptide matches to TGACv1 genes and transcripts are provided as Supplementary File **S9**.

## 10 Orthologous gene family analyses

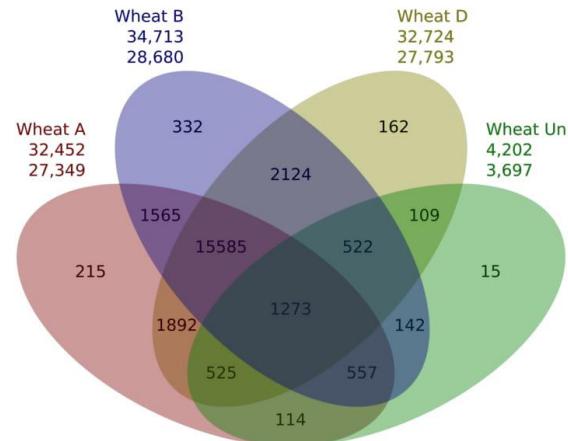
### 10.1 OrthoMCL gene family clustering of wheat subgenome genes

Gene family clusters were defined from the bread wheat high-confidence class genes, separated for their subgenome origin (A, B and D) and undefined origin ("U") using OrthoMCL software version 2.0 Li (2003). In a first step, pairwise sequence similarities between all input protein sequences were calculated using BLASTP with an e-value cut-off of  $1 \times 10^{-5}$ . Markov clustering of the resulting similarity matrix was used to define the ortholog cluster structure, using an inflation value (-I) of 1.5 (OrthoMCL default).

The input datasets were:

- Bread Wheat A genome (high-conf): 32,452 genes
- Bread Wheat B genome (high-conf): 34,713 genes
- Bread Wheat D genome (high-conf): 32,724 genes
- Bread Wheat genes of unknown origin (high-conf): 4202 genes

Splice variants were removed from the data sets, keeping the representative gene model, and data sets were filtered for internal stop codons and incompatible reading frames. A total of 87,519 coding sequences from these three datasets were clustered into 25,132 gene families (clusters). An overview of the cluster structure is shown in Figure S10.1. We identified 13,070  $\times$  3 genes found in a 1:1:1:0 ratio in the A,B,D and U subgenomes (triads); this set was filtered to 9642 triads with >90% identity in pairwise BLASTP alignments between A,B and D genes (Supplemental file S10). The same OrthoMCL analysis was also performed with all TGACv1 gene models, both high- and low-confidence in a separate run (Figure S10.2).

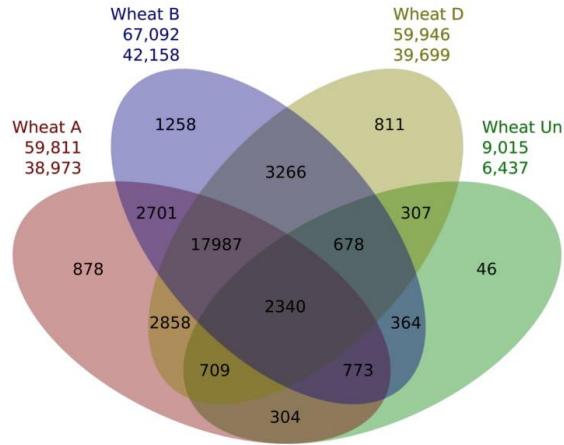


**Figure S10.1:** OrthoMCL clustering of bread wheat genes (HC class) from the A, B and D subgenome and unclassified origin ("Un"). The first number under the species name gives the total number of transcripts predicted, the second number is the number of transcripts in OrthoMCL clusters. The difference gives the number of singletons (genes not clustered).

### 10.2 OrthoMCL gene family clustering of the bread wheat genome and related species

Following the protocol used in section 10.1, OrthoMCL was used to define gene family clusters at a species level, using as datasets the bread wheat high-confidence class genes, the annotated gene sets of three grasses from diverse grass sub-families, and *Arabidopsis thaliana* (Figure S10.3). The input datasets were:

- Bread Wheat A genome (high-conf): 32,452 genes
- Bread Wheat B genome (high-conf): 34,713 genes
- Bread Wheat D genome (high-conf): 32,724 genes
- Bread Wheat genes of unknown origin (high-conf): 4,202 genes
- Sorghum bicolor v2.1: 33,032 genes
- Brachypodium distachyon v2.1: 31,694 genes



**Figure S10.2:** OrthoMCL clustering of bread wheat genes (all confidence classes) from the A, B and D subgenome and unclassified origin (“Un”). The first number under the species name gives the total number of transcripts predicted, the second number is the number of transcripts in OrthoMCL clusters. The difference gives the number of singletons (genes not clustered).

- Rice MSU7.0: 39,049 genes
- Arabidopsis thaliana TAIR10: 27,416 genes

Coding sequences from these five species were clustered into 29,862 gene families.

In a separate run, the same OrthoMCL analysis was performed with all bread wheat gene models given as a single species (Figure S10.3).

### 10.3 GO over-/under-representation for specific groups/singletons

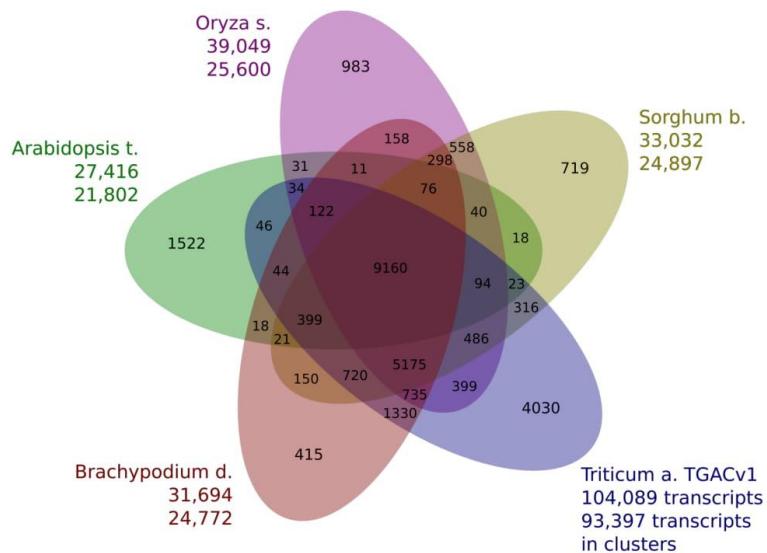
Over-/under-representation of gene ontology (GO) terms in specific gene families and subsets (see Section 10.4) were analysed via hypergeometric testing using the functions GOstats (Falcon and Gentleman, 2007) and GSEABase (Morgan et al., 2008) from the Bioconductor R package against a universe of all genes with GO annotations. Revigo (Sipek et al., 2011), which removes redundant and similar terms from long GO lists by semantic clustering was applied to visualise the enrichment results.

### 10.4 Expanded gene families in OrthoMCL and GO over-representation within

From the OrthoMCL analyses described in Sections 10.1 and 10.2, we extracted gene models from different distinct OrthoMCL subsets:

- A **“Subgenome-specific” set:** Wheat genes in groups/clusters which are subgenome-specific (cluster/group contains only genes from subgenome A, B or D) and cluster size greater than 1;
- B **“Subgenome-singletons” set:** Wheat genes which were not clustered within any of the OrthoMCL groups, termed “Singletons”, separated by their subgenome origin;
- C **“Wheat-subgenome-expanded” set:** Wheat genes in groups/clusters where the gene copy number is significantly (p-value less than 0.05) expanded in one of the subgenomes relative to the other subgenome including clusters (size greater than one) that only consist of the respective subgenome genes;
- D **“Wheat-expanded(A/B/D)” set:** Wheat genes, separated by subgenome origin, in groups/clusters where the Wheat gene copy number is significantly expanded (p-value less than 0.05) relative to any of the other species contained including clusters (size greater than one) that only consist of Wheat genes.

The individual gene sets were analysed for over-represented GO terms from all GO categories “biological process”, “molecular function” and “cellular component”. Results are summarized and visualized in Supplemental file S2.



**Figure S10.3:** OrthoMCL clustering of bread wheat genes (HC confidence class) from the A, B and D subgenome and unclassified origin ("Un") together as a single species, against the gene complements of Arabidopsis, Sorghum, Rice and Brachypodium. The first number under the species name gives the total number of transcripts predicted, the second number is the number of transcripts in OrthoMCL clusters. The difference gives the number of singletons (genes not clustered).

## 11 Gene expression analyses

### 11.1 Expression quantification and analysis

#### 11.1.1 Gene expression quantification

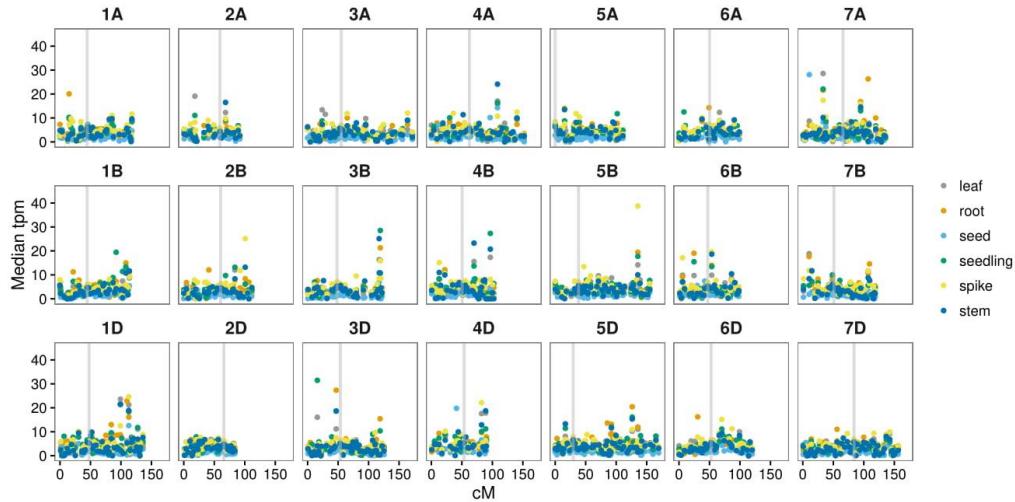
Wheat gene expression quantification was carried out as described in Borrill et al. (2016) using kallisto v0.42.3 (Bray et al., 2016) to pseudoalign reads to the complete TGAC transcriptome (including both high and low confidence genes) as a reference. The SRA studies included are listed in S11.1. For paired end reads *kallisto* was run using default parameters with 100 bootstraps (-b 100). For single end reads kallisto was run using 100 bootstraps (-b 100) in the single end read mode (-single), the average fragment length used was 150 bp (-l 150) with a standard deviation of 50 (-s 50) - these values were taken as an average of reported fragment lengths for studies included.

**Table S11.1:** SRA studies analysed with expVIP using TGAC gene models as a reference.

Study identifier	Summary	Total reads	Reads mapped to TGAC		Reference
			Reads mapped to IWGSC		
DRP000768	phosphate starvation in roots and shoots	118,053,746	104,886,994 (88%)	84,529,715 (72%)	Oono et al. (2013)
ERP003465	fusarium head blight infected spikelets	1,827,362,091	1,633,149,812 (89%)	1,357,197,955 (74%)	Kugler et al. (2013)
ERP004505	grain tissue-specific developmental timecourse	873,709,556	718,777,030 (54%)	475,184,621 (82%)	Pfeifer et al. (2014)
SRP004884	flag leaf downregulation of GPC	209,427,573	148,280,320 (72%)	121,855,143 (58%)	Cantu et al. (2011)
SRP013449	grain tissue-specific developmental timecourse	132,702,451	110,682,153 (83%)	82,417,257 (62%)	Gillies et al. (2012)
SRP017303	stripe rust infected seedlings	33,361,836	15,622,370 (47%)	13,732,210 (41%)	Cantu et al. (2013)
SRP022869	Septoria tritici infected seedlings	100,582,632	71,948,196 (72%)	63,155,877 (63%)	Yang et al. (2013)
SRP028357	shoots and leaves of nulli tetra group 1 and group 5	3,304,500,117	2,918,789,524 (88%)	2,258,692,000 (68%)	Leach et al. (2014)
SRP029372	grain tissue-specific developmental timecourse	101,477,759	26,992,810 (22%)	17,525,439 (17%)	Li et al. (2013)
SRP038912	comparison of stamen pistil and pistilliody expression	217,315,378	196,322,732 (90%)	153,009,134 (70%)	Yang et al. (2015)
SRP041017	stripe rust and powdery mildew infection timecourse	395,463,786	325,434,104 (82%)	272,228,560 (69%)	Zhang et al. (2014a)
SRP041022	developmental time-course of synthetic hexaploid	134,641,113	120,448,445 (90%)	84,583,556 (63%)	Li et al. (2014)
ERP008767	grain tissue-specific expression at 12 days post anthesis	45,213,827	36,971,938 (82%)	26,420,708 (58%)	Pearce et al. (2015)
SRP045409	drought and heat stress time-course in seedlings	921,578,806	592,272,829 (64%)	533,928,182 (58%)	Liu et al. (2015)
ERP004714	developmental time-course of Chinese Spring	1,536,051,415	1,340,790,669 (88%)	1,066,712,760 (69%)	Choulet et al. (2014)
SRP056412	grain developmental timecourse with 4A dormancy QTL	1,875,916,011	1,082,551,207 (57%)	808,809,053 (43%)	Barrero et al. (2015)
PR-JEB15048	developmental time-course of Chinese Spring	824,241,135	631,301,185 (77%)	N/A	This study.

#### 11.1.2 Differential gene expression analysis

Differential gene expression analysis was carried out on the kallisto output abundance files using sleuth (Pimentel et al., 2016). Default settings were used except that the maximum number of bootstraps considered was 30 (`max_bootstrap = 30`). For the integrated disease and stress analysis each sample was compared to the control sample from the study from which it originated. Genes with a FDR adjusted p-value (q-value) less than 0.001 were considered differentially expressed.



**Figure S11.1:** Median gene expression level per chromosome bin. cM position was determined by BLAST of TGAC scaffolds to the Chapman scaffold which had POPSEQ position information. Only bins with 3 or more genes were included. Outliers above expression level 45tpm were excluded from the graph. Grey vertical lines indicate centromere position.

### 11.1.3 Visualisation of gene expression

The quantified gene expression from kallisto were visualised using the expVIP platform (Borrill et al., 2016). It is displayed at [www.wheat-expression.com](http://www.wheat-expression.com).

## 11.2 Gene expression across 17 diverse RNA-seq studies

We used expVIP (Borrill et al., 2016) to analyse 16 wheat gene expression studies from the short read archive (SRA) from a range of tissues, developmental stages and stress conditions alongside the six RNA samples sequenced during the course of this study (Table S11.1). In total these 424 individual samples contained 12.6 billion reads of which 10 billion mapped to the TGAC transcriptome containing 273,739 genes. This average mapping rate of 75% of reads is higher than the 59% of reads which mapped to the previous IWGSC gene models suggesting that the TGAC transcriptome is more complete. We found that 95% of genes (260,079) had at least 1 read mapping to them, and 58% of genes (160,074) were expressed in at least one samples at over 2tpm which has been advocated as the cut-off for real expression over noise (Wagner et al., 2013). The percentage of genes expressed over the background noise level of 2tpm is relatively low (58%) which may be because the TGAC gene models also include non-coding RNAs which are generally expressed at very low expression levels and low confidence gene models which are not supported by evidence from other species. If we only include high confidence gene models 78% of genes are expressed at over 2tpm. To facilitate access to these RNA-seq datasets we have updated <http://www.wheat-expression.com> to show gene expression levels for each TGAC gene of interest across all the 17 different studies. The visualisation interface can be filtered and sorted by the viewer according to the origin of each sample in terms of tissue, age, stress and variety. One gene and its homoeologs can be displayed as a bar graph or multiple genes can be displayed as a heatmap.

## 11.3 Gene expression patterns across chromosome regions

To investigate whether specific chromosomal domains influence the gene expression level we examined gene expression across the length of the chromosomes using genetic map assignments described in Section 5. We found that in general the median expression of genes was similar throughout most chromosomes (Figure S11.1). However certain chromosomal regions had much higher expression across several or all of the six tissues examined and these “enhanced expression regions” were located outside of centromeric regions.

## 11.4 Analysis of homoeolog gene expression in stress conditions

We identified 9642 triads which had a 1-1-1 relationship between the A, B and D genome copies (Section 10.1). To understand the roles of the three homoeologous copies within triads to a range of stress conditions we leveraged existing RNA-seq data for seedlings (Table S11.2). Gene expression quantification and differential expression analysis was carried out as described in Sections 11.1.1 and 11.1.2. Within each triad we classified changes in response to stress in each homoeolog as either up-regulation (over 2-fold change), down-regulation (under 0.5-fold change) or flat (between 0.5 to 2 fold change); all tests were considered statistically significant with  $q$  lower than 0.001. Each triad was then classified according to the number of homoeologs differentially expressed and their direction of change Table S11.3.

**Table S11.2:** RNA-seq samples used to analyse the response of homoeologous genes to stress conditions.

Study	Age	Conditions	Replicates
SRP041017	7 days	Stripe rust 24 h	3
		Stripe rust 48 h	3
		Stripe rust 72 h	3
		Powdery mildew 24 h	3
		Powdery mildew 48 h	3
		Powdery mildew 72 h	3
SRP045409	7 days	Drought stress 1 h	2
		Drought stress 6 h	2
		Heat stress 1 h	2
		Heat stress 6 h	2
		Drought and heat stress 1 h	2
		Drought and heat stress 6 h	2

**Table S11.3:** The expression patterns of homoeologs within triads in response to stress treatments. \*For triads where two homoeologues are up or down regulated, the third homoeologue could not be expressed in the opposite direction to avoid double counting of “opposite” class triads.

Condition	0	1 up	1 down	2 up*	2 down*	3 up	3 down	Opposite
drought_1h	8,588	866	148	31	0	0	0	9
heat_1h	6,931	1,731	707	142	17	3	0	111
drought_heat_1h	5,941	2,008	1,129	214	68	10	1	271
drought_6h	6,248	1,521	1,354	148	110	1	1	259
heat_6h	5,288	1,780	1,728	195	211	5	3	432
drought_heat_6h	4,965	1,677	2,028	185	253	8	8	518
mildew_24h	8,793	607	218	15	2	0	0	7
mildew_48h	8,802	180	640	4	12	0	0	4
mildew_72h	9,184	24	425	0	6	0	0	3
yellow_rust_24h	9,069	267	290	3	7	0	0	6
yellow_rust_48h	9,455	13	172	0	2	0	0	0
yellow_rust_72h	9,342	41	257	0	1	0	0	1

In triads in which two homoeologs were up- or down-regulated, the A, B and D genome were represented equally (chi-squared test  $p = 0.517$  and  $p = 0.243$  respectively). Similarly in triads in which one homoeolog was down-regulated the three genomes were represented equally (chi-squared test  $p = 0.537$ ). However in triads in which one homoeolog was up-regulated the three genomes did not respond equally, with the D genome being more responsive to stress conditions (the numbers of triads with one homoeolog up-regulated in which the A, B and D genome homoeolog was upregulated were 3390, 3494 and 3831 respectively, chi-squared test  $p = 3.45 \times 10^{-7}$ ). In triads with opposite patterns of homoeolog expression the B genome was more frequently up-regulated than the other two genomes (the numbers of triads with opposite homoeolog expression patterns in which the A, B and D genome homoeolog was upregulated were 526, 606 and 538 respectively, chi-squared test  $p = 0.035$ ), however all three genomes were as likely as each other to be the down-regulated genome in triads with opposite homoeolog expression patterns (chi-squared test  $p = 0.0745$ ).

### 11.5 Homoeologous gene expression analysis

We decided to investigate expression of homoeologous genes using an ad-hoc approach similar to that described in Liu et al. (2015). We decided to focus on three studies for this analysis: SRP041017, SRP045409, and ERP004505. For each of our 9642 triads, we verified in each condition whether their expression was balanced by performing a paired fisher test between the A and B gene, B and D gene, and A and D gene; as in Liu et al. (2015), we compared the two expression values for the triads against the sum of all the expression values for the subgenome in the condition, minus the expression of the gene under analysis (equation (1)); Fisher test as implemented in Scipy v. 0.18.0 Jones et al. (2001)). We corrected our p-values using the standard Benjamini-Hochberg method for False Discovery Rate, as implemented in Stasmodels 0.6.1 (Seabold and Perktold, 2010).

$$F(x,y) = Fisher((tpm_x, tpm_y), ((\sum_{\chi=1}^{\Xi} tpm_{\chi}) - tpm_x, (\sum_{v=1}^{\Upsilon} tpm_v) - tpm_y)) \quad (1)$$

The probability for two homoeologous  $x,y$  genes to be expressed at an unbalanced level was calculated by performing a Fisher exact test of their expression, in TPM, versus the sum of all TPM values of the triads for their respective subgenomes  $\Xi$  and  $\Upsilon$  excluding the couple of genes themselves.

Expression values for the analysed triplets, and the accession codes for the RNA-Seq raw data, can be found in Supplementary file **S11**, while the Fisher test evaluation results are reposted in Supplementary file **S12**.

Subsequently, we considered a pairwise comparison within a replicate as significant if the following conditions verified:

1. at least one of the two genes compared had to have an expression level greater than 0.01 TPMs (to exclude lowly expressed loci).
2. the comparison had to have a corrected p-value lower than 0.05.

3. Either one of the two genes had an expression of 0, or the absolute log2 Fold Change between the two genes was 1 or above.

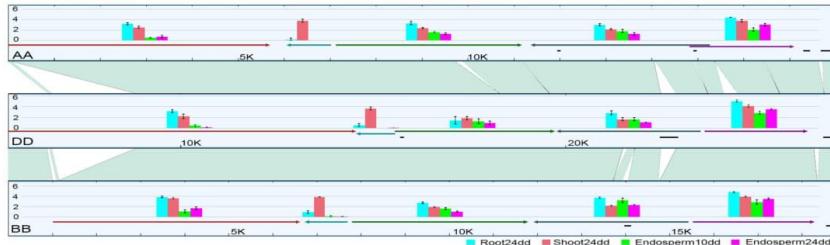
Each comparison was assigned one of three signed values (0 for no differential expression, 1 for an over-expression of the first gene compared to the second, -1 for an under-expression of the first gene compared to the first, NA if neither gene was expressed at a sufficiently high level). A pair of genes was considered as unbalanced if all the replicates were found to have a significant and coherent difference in expression between the two members (ie. if in a couple the first gene was significantly under-expressed in a sample and significantly over-expressed in another replicate or without evidence for a difference in expression, the comparison would have been called as inconclusive). A triad was called as unbalanced if at least one of its internal pairs was unbalanced.

Expression values for the analysed triplets can be found in supplementary file **S11**, while the Fisher test evaluation results are reported in supplementary file **S12**. The final evaluation is reported in supplementary file **S13**.

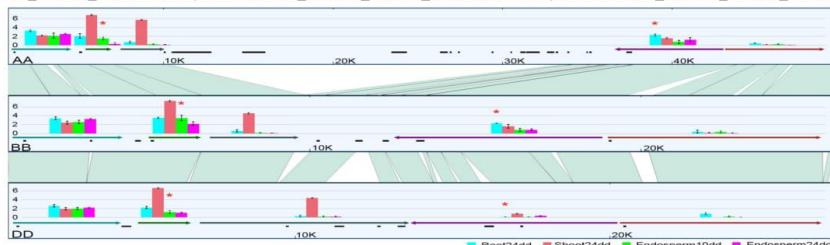
### 11.6 Gene expression in syntenic loci

Collinearity was detected between the high confidence genes annotated on the wheat sub-genomes using MCScanX (Wang et al., 2012). Protein sequences for the high confidence genes were used in all versus all BLASTP analysis (Section 10.1).

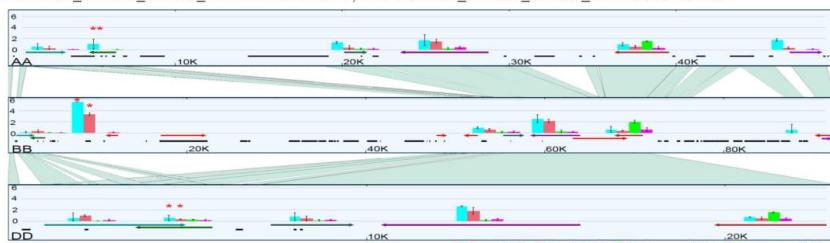
Conserved blocks were defined as a set of at least 5 genes (anchors) in the same order between 2 sub-genomes, with a maximum of 25 spacer genes between the anchors in a collinear block. A total of 91 pairwise collinear blocks were identified, from these 12 collinear blocks of the A, B and D sub-genomes were identified. (Supplemental file **S14**). Pairwise alignments between two syntenic blocks were calculated using LAST (Frith et al., 2010). Adjacent syntenic alignments were joined into single larger syntenic alignments using the UCSC Chain/Net pipeline (Kent et al., 2003). Expression levels of genes in the blocks were assessed using triplicated RNAseq data from Chinese Spring root and shoot tissues, and from 10day and 20 whole endosperm tissue (SRA studies DRP000768 and ERP004505). Gene expression levels were expressed as  $\log_2(TPM + 1)$ . Unbalanced expression was defined as a significant difference in expression between homoeologues in any of the four tissues measured, defined as the expression of any homoeologue having greater than  $4(TPM + 1)$  expression levels than another homoeolog. Collinear relationships, synteny links, and expression levels for the genes on four syntenic blocks selected to illustrate different patterns of gene expression were plotted using SyntenyPlot (<https://github.com/lufuhao/SyntenyPlot>). Promoter motifs were identified using PlantCARE, and transcription start sites were identified from the TGACv1 annotation. Synteny views of genes in AA, BB, and DD scaffolds in 4 selected blocks are shown in Figure S11.2, showing different patterns of conservation of gene and repeat order and gene expression.



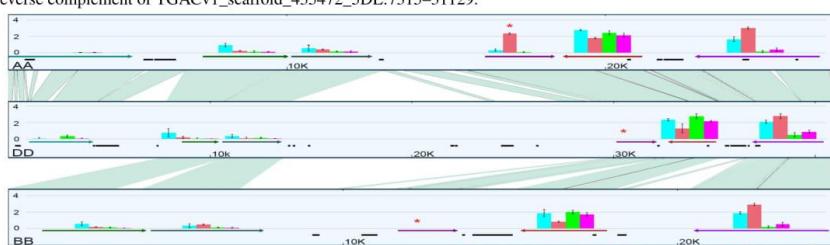
(A) This block illustrates a high degree of similarity in the A, B and D genomes, with similar patterns of gene expression. AA: reverse complement of TGACv1\_scaffold\_195481\_3AL:1-18115, BB: TGACv1\_scaffold\_224116\_3B:34006-53294, DD: TGACv1\_scaffold\_250027\_3DL:14626-41915.



(B) This block shows the interspersion of a tract of repeats in the A genome compared to the B and D genomic blocks. A gene encoding a histone-lysine N methyltransferase in the D genome is expressed at lower levels in root tissues. AA: TGACv1\_scaffold\_288349\_4AL:28586-78590, BB: reverse complement of TGACv1\_scaffold\_328157\_4BS:112255-138693, DD: TGACv1\_scaffold\_361457\_4DS:18746-46470.

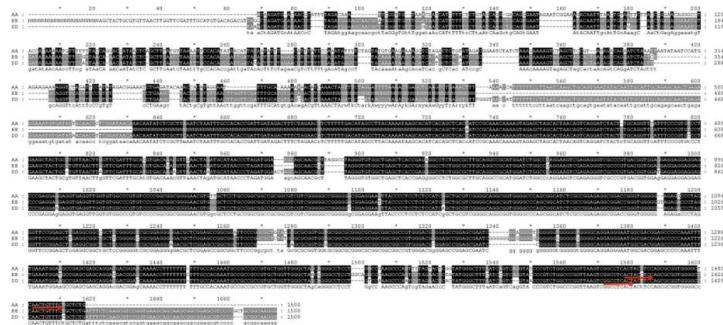


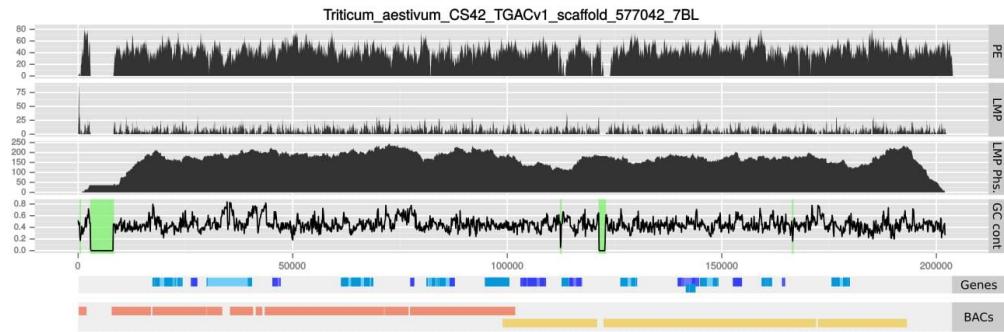
(C) This block shows unbalanced expression of an uncharacterised protein in the A, B and D genomes. There are major differences in the repeat composition in the A and B genomes compared to the D genome AA: TGACv1\_scaffold\_375286\_5AL:25956-75758, BB: TGACv1\_scaffold\_404593\_5BL:66316-159457, DD: reverse complement of TGACv1\_scaffold\_435472\_5DL:7315-31129.



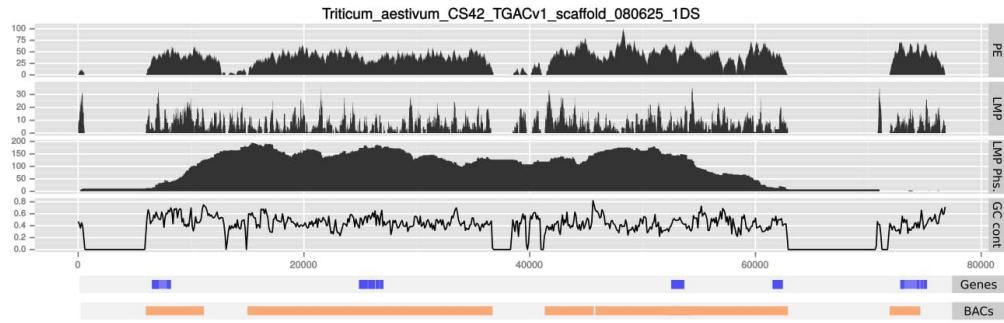
(D) A gene encoding a cytochrome P450 7A14-like protein is highly expressed in shoot tissues in the A genome, and the homoeologous B and D genes are not detectably expressed. AA: TGACv1\_scaffold\_392578\_5AS:182276-210005, BB: TGACv1\_scaffold\_424311\_5BS:12106-37625, DD: TGACv1\_scaffold\_456510\_5DS:23146-64545.

**Figure S11.2:** Four syntenic blocks showing conserved gene order and different patterns of gene expression, and repeat and gene interspersions. Genome segments are arranged to reveal patterns of maximum conservation. Unbalanced gene expression is identified by a red asterisk above the bar graph of expression levels. The x axis scale is in bp. Bar plots the  $\log_2(TPM + 1)$  of gene expression (0-6) on the y axis. Asterisks above gene expression bar plots indicate unbalanced expression. Arrows indicate genes, with homoeologous genes shown in the same colour. Red arrows mark no-syntenic genes. Black boxes show repeat-masked regions.





**Figure S12.1:** Scaffold 577042 of the TGACv1 assembly with resistance genes, aligned BACs and read data. The tracks from top to bottom show coverage of paired-end reads, coverage of mate-pair reads, coverage of mate-pair fragments, GC content and N regions (highlighted in green), resistance genes, and BACs. There are two BACs in 7 and 4 contigs, respectively, and 20 resistance genes.



**Figure S12.2:** Scaffold 080625 of the TGACv1 assembly with gluten genes, aligned BACs and read data. The tracks from top to bottom show coverage of paired-end reads, coverage of mate-pair reads, coverage of mate-pair fragments, GC content and N regions (highlighted in green), gluten genes, and BACs. There is one BAC in 5 contigs and 6 resistance genes.

## 12 Gene families of agronomic importance

### 12.1 Disease resistance genes

Disease resistance genes were predicted by analysing the domain architectures with previously established pipeline (Sarris et al., 2016) which utilises the Pfam annotation of functional domains. In addition, we scanned high confidence proteomes for previously identified NLR MEME motifs (Jupe et al., 2012) and analysed the results with NLR-parser (Steuernagel et al., 2015) to predict NLR-associated motifs and assess CC-NBS-LRR type disease resistance genes. The fragmented and complete transcript genes were also compared by the presence of start and stop codons in predicted transcripts. The sequences for the resistance genes are provided in Supplementary files **S15**, **S16**, **S17** and **S18**.

### 12.2 Gluten genes

Due to the challenges of annotating repeat rich gluten genes, we reviewed all regions with nucleotide similarity to publicly available gluten sequences (NCBI, Zhang et al. (2014b); Pfeifer et al. (2014)) with blastx (e-10) or GMAP (at least 95% identity, at least 40% coverage) via the Apollo browser (<http://genomearchitect.github.io/>).

Due to the challenges of annotating repeat rich gluten genes, we reviewed regions with nucleotide similarity to publicly available gluten sequences (NCBI, Zhang et al. (2014a), Pfeifer et al. (2014)) with blastx (e-10) or GMAP (at least 95% identity, at least 40% coverage) via the Apollo browser (<http://genomearchitect.github.io/>). The manually updated annotations are provided as supplementary files **S19**, **S20**, **S21**, **S22** and **S23**. Gluten pseudogenes are provided in a separate Supplementary file, **S24**.

### 12.3 Gibberellin genes

Wheat genomic sequences corresponding to genes from the gibberellin biosynthesis, inactivation and signalling pathways were identified in the TGACv1 assembly by BLASTN, using previously identified sequences from wheat (Pearce et al., 2015) or rice (Hirano et al., 2008) and aligned using Geneious (<http://www.geneious.com>).

### 12.4 BAC analysis

The two BACs we have used in our examples are from a larger set of BACs which we were sequenced and assembled. Briefly, BACs were selected for sequencing from a *HinDIII* partial digest BAC library (Allouis et al., 2003). BAC DNAs were miniprepped (Sambrook and Russell, 2006), treated with ATP dependent DNase to remove *E. coli* genomic DNA, and individually barcoded Illumina Nextera libraries prepared. Nextera libraries were sequenced using Illumina chemistry  $2 \times 250$  bp cycles (paired end). The reads were demultiplexed and filtered to remove the BAC vector, *E. coli* genome, and wheat chloroplast and mitochondria sequences. The remaining reads for each BAC were then assembled using DISCOVAR *de novo* (Weisenfeld et al., 2014) and then trimmed to remove any remaining vector sequence from the contigs. The assemblies had an average content of 111 kbp and an average contig N50 of 16.7 kbp. The assemblies of the BACs used in Figures S12.1 and S12.2 are given in Supplemental files **S25** and **S26**, respectively.

## 13 Authors' contributions

BJC, LV, DH, CF, DS, FDP and MDC designed the sequencing experiments. DNA and RNA was isolated by NMCK or GY, and sequencing libraries prepared by DH, TB and JL. BJC, GGA, JW and FDP designed and implemented the assembly strategy. GGA, JW and BJC performed the genome assembly. LV, GKA and DS designed and implemented the annotation strategy, including manual curation of gluten genes and alternative splicing analysis. HG (MIPS) performed the global analysis of repeats. CU performed detailed analysis of breakpoints versus previous assemblies. MS and GH (MIPS), PK (EBI) and DS performed global gene family analyses. CS, DR and KV.K designed and implemented the approach to anchor the assembly onto the genetic map, predicted translocations and designed the validation strategy. R-gene family analyses and gluten gene family analyses was performed by CS, DR and KV.K. Analyses of BACs was done by GKe and MDC. Validation of predicted translocations was performed by MDC, AC, NP and LPA. PB, CU, RRG, LV, DS, F-HL and MWB performed gene expression analyses and JT, OD, AHM proteome profiling and analysis. AP performed analyses of gibberellin pathways. DMB, GN, AK, GKA, DS, RPD, and PJK integrated assemblies, annotations and sequencing reads into public databases. CF and HC provided project management. MWB, LV, DS, GKe and MDC wrote the manuscript and all authors contributed to the text.

## 14 File list

- S1** Supplemental Information Additional information for the main article.
- S2** Gene ontology enrichment results for singleton and expanded OrthoMCL families (see Section 10)
- S3** List of all potential translocation events supported by OrthoMCL outlier triads as described in Section 6.1.
- S4** WGS map with corrected bins.
- S5** The TGACv1 map (scaffold id, chromosome, cM; only class 1 scaffolds)
- S6** Scaffold, Position on TGACv1 map (chromosome:cM), and Map classification for all TGACv1 scaffolds that had at least one matching WGS marker. For class 1 scaffolds (assigned to unique genetic bin), Position has only one entry, while class 2 (assigned to ambiguous bins on the same chromosome), class 3 (assigned to ambiguous bins on homoeologous chromosomes), and class 4 (assigned to at least two non-homoeologous chromosomes) scaffolds have multiple entries, separated by ';'. See Section 5.
- S7** Primers used for translocation confirmation.
- S8** AUGUSTUS config file used in gene prediction.
- S9** Peptide matches to TGACv1 genes and transcripts
- S10** List of the 9642 triads of homoeologous genes, as defined in Section 10.1
- S11** Expression values, in TPM, for the transcripts in CS42.triplets.tsv across the analysed samples. See Section 11.5
- S12** Pairwise Fisher test for the expression within a triad within each replicate. See Section 11.5
- S13** Evaluation of whether members of a triplet were expressed in a balanced or unbalanced way in a given sample. See Section 11.5.
- S14** Collinear blocks identified between A,B and D genomes. See Section 11.6.
- S15** Sequences of all NBS CDS in Fasta format. See Section 12.1
- S16** Sequences of all NBS-LRR CDS in Fasta format. See Section 12.1.
- S17** Sequences of all translated NBS-LRR CDS in Fasta format. See Section 12.1.
- S18** Sequences of all translated NBS CDS in Fasta format. See Section 12.1.
- S19** cDNA sequences of manually annotated gluten genes. See Section 12.2.
- S20** Coding sequences of manually annotated gluten genes. See Section 12.2.
- S21** Protein sequences of manually annotated gluten genes. See Section 12.2.
- S22** Gene position and structure of manually annotated gluten genes. See Section 12.2.
- S23** Summary of manual gluten gene annotation. See Section 12.2.
- S24** Sequences of manually annotated gluten pseudogenes. See Section 12.2.
- S25** Sequence of scaffold 577042 of the TGACv1 assembly containing resistance genes. See Section 12.4.
- S26** Sequence of scaffold 080625 of the TGACv1 assembly containing gluten genes. See Section 12.4.

## 15 References

Allouis, S., Moore, G., Bellec, A., Sharp, R., Rampant, P. F., Mortimer, K., Pateyron, S., Foote, T., Griffiths, S., Caboche, M., *et al.*, 2003. Construction and characterisation of a hexaploid wheat (*Triticum aestivum* L.) BAC library from the reference germplasm 'Chinese Spring'. *Cereal Research Communications*, **31**(3/4):331–338.

BabrahamLab, 2014. Trim Galore.

Barrero, J. M., Cavanagh, C., Verbyla, K. L., Tibbets, J. F., Verbyla, A. P., Huang, B. E., Rosewarne, G. M., Stephen, S., Wang, P., Whan, A., *et al.*, 2015. Transcriptomic analysis of wheat near-isogenic lines identifies PM19-A1 and A2 as candidates for a major dormancy QTL. *Genome Biology*, **16**(1):93.

Borrell, P., Ramirez-Gonzalez, R., and Uauy, C., 2016. expVIP: a Customizable RNA-seq Data Analysis and Visualization Platform. *Plant Physiology*, **170**(4):2172–2186.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L., 2016. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, **34**(5):525–527.

Cantu, D., Pearce, S. P., Distelfeld, A., Christiansen, M. W., Uauy, C., Akhunov, E., Fahima, T., and Dubcovsky, J., 2011. Effect of the down-regulation of the high Grain Protein Content (GPC) genes on the wheat transcriptome during monocarpic senescence. *BMC Genomics*, **12**(1):492.

Cantu, D., Segovia, V., MacLean, D., Bayles, R., Chen, X., Kamoun, S., Dubcovsky, J., Saunders, D. G., and Uauy, C., 2013. Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. sp. *tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC Genomics*, **14**(1):270.

Chapman, J. A., Mascher, M., Buluç, A., Barry, K., Georganas, E., Session, A., Strnadova, V., Jenkins, J., Sehgal, S., Oliker, L., *et al.*, 2015. A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biology*, **16**(1):26.

Choulet, F., Alberti, A., Theil, S., Glover, N., Barbe, V., Daron, J., Pingault, L., Sourdille, P., Couloux, A., Paux, E., *et al.*, 2014. Structural and functional partitioning of bread wheat chromosome 3B. *Science*, **345**(6194):1249721–1249721.

Clavijo, B., Garcia Accinelli, G., Wright, J., Heavens, D., Barr, K., Yanes, L., and Di Palma, F., 2017. W2RAP: a pipeline for high quality, robust assemblies of large complex genomes from short read data. *bioRxiv*, **10.1101/110999**.

Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., *et al.*, 2010. A guided tour of the Trans-Proteomic Pipeline. *PROTEOMICS*, **10**(6):1150–1159.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R., 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**(1):15–21.

Duncan, O., Trösch, J., Fenske, R., Taylor, N. L., and Millar, A. H., 2017. Resource: Mapping the *Triticum aestivum* proteome. *The Plant Journal*, **89**(3):601–616.

Eddy, S. R., 2011. Accelerated Profile HMM Searches. *PLoS computational biology*, **7**(10):e1002195.

EdicoGenome, 2014. Dragen Bio-IT processor.

Ellinghaus, D., Kurtz, S., and Willhöft, U., 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC bioinformatics*, **9**:18.

Eng, J. K., Jahan, T. A., and Hoopmann, M. R., 2013. Comet: An open-source MS/MS sequence database search tool. *PROTEOMICS*, **13**(1):22–24.

Falcon, S. and Gentleman, R., 2007. Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23**(2):257–258.

Fernandez, N. and Guerrero, D., 2012. Full Lengther Next.

Frith, M. C., Hamada, M., and Horton, P., 2010. Parameters for accurate genome alignment. *BMC Bioinformatics*, **11**(1):80.

Gillies, S. A., Futardo, A., and Henry, R. J., 2012. Gene expression in the developing aleurone and starchy endosperm of wheat. *Plant Biotechnology Journal*, **10**(6):668–679.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., *et al.*, 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, **40**(D1):D1178–D1186.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., *et al.*, 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**(7):644–652.

Haas, B. J., 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, **31**(19):5654–5666.

Haas, B. J., 2010. TransposonPSI.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., *et al.*, 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, **8**(8):1494–1512.

Hallab, A., Klee, K., Boecker, F., Girish, S., and Schoof, H., 2014. Automated assignment of Human Readable Descriptions (AHRD).

Heavens, D., Accinelli, G. G., Clavijo, B., and Clark, M. D., 2015. A method to simultaneously construct up to 12 differently sized Illumina Nextera long mate pair libraries with reduced DNA input, time, and cost. *BioTechniques*, **59**(1):42–45.

Hirano, K., Aya, K., Hobo, T., Sakakibara, H., Kojima, M., Shim, R. A., Hasegawa, Y., Ueguchi-Tanaka, M., and Matsuoka, M., 2008. Comprehensive Transcriptome Analysis of Phytohormone Biosynthesis and Signaling Genes in Microspore/Pollen and Tapetum of Rice. *Plant and Cell Physiology*, **49**(10):1429–1450.

Jones, E., Oliphant, T., Peterson, P., and Others, 2001. SciPy: Open source scientific tools for Python.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., *et al.*, 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**(9):1236–1240.

Jupe, F., Pritchard, L., Etherington, G. J., MacKenzie, K., Cock, P. J., Wright, F., Sharma, S. K., Bolser, D., Bryan, G. J., Jones, J. D., *et al.*, 2012. Identification and localisation of the NB-LRR gene family within the potato genome. *BMC Genomics*, **13**(1):75.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D., 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences*, **100**(20):11484–11489.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, **14**(4):R36.

Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., and Gao, G., 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research*, **35**(Web Server issue):W345–9.

Kopylova, E., Noe, L., and Touzet, H., 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, **28**(24):3211–3217.

Kugler, K. G., Siegwart, G., Nussbaumer, T., Ametz, C., Spannagl, M., Steiner, B., Lemmens, M., Mayer, K. F., Buerstmayr, H., and Schweiger, W., *et al.*, 2013. Quantitative trait loci-dependent analysis of a gene co-expression network associated with Fusarium head blight resistance in bread wheat (*Triticum aestivum* L.). *BMC Genomics*, **14**(1):728.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., *et al.*, 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, **40**(D1):D1202–D1210.

Langmead, B. and Salzberg, S. L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**(4):357–359.

Leach, L. J., Belfield, E. J., Jiang, C., Brown, C., Mithani, A., and Harberd, N. P., 2014. Patterns of homoeologous gene expression shown by RNA sequencing in hexaploid bread wheat. *BMC Genomics*, **15**(1):276.

Leggett, R. M., Clavijo, B. J., Clissold, L., Clark, M. D., and Caccamo, M., 2014. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics*, **30**(4):566–568.

Li, A., Liu, D., Wu, J., Zhao, X., Hao, M., Geng, S., Yan, J., Jiang, X., Zhang, L., Wu, J., *et al.*, 2014. mRNA and Small RNA Transcriptomes Reveal Insights into Dynamic Homoeolog Regulation of Allopolyploid Heterosis in Nascent Hexaploid Wheat. *The Plant cell*, **26**(5):1878–1900.

Li, H.-Z., Gao, X., Li, X.-Y., Chen, Q.-J., Dong, J., and Zhao, W.-C., 2013. Evaluation of Assembly Strategies Using RNA-Seq Data Associated with Grain Development of Wheat (*Triticum aestivum* L.). *PLoS ONE*, **8**(12):e83530.

Li, L., 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, **13**(9):2178–2189.

Liu, Z., Xin, M., Qin, J., Peng, H., Ni, Z., Yao, Y., and Sun, Q., 2015. Temporal transcriptome profiling reveals expression partitioning of homeologous genes contributing to heat and drought acclimation in wheat (*Triticum aestivum* L.). *BMC Plant Biology*, **15**(1):152.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., *et al.*, 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, **1**(1):18.

Magoc, T. and Salzberg, S. L., 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**(21):2957–2963.

Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., and Clavijo, B. J., 2017. Kat: a k-mer analysis toolkit to quality control ngs datasets and genome assemblies. *Bioinformatics*, **33**(4):574.

Mapleson, D. L., Venturini, L., and Swarbreck, D., 2016. Portcullis. <https://github.com/maplesond/portcullis>.

Morgan, M., Falcon, S., and Gentleman, R., 2008. GSEABase: Gene set enrichment data structures and methods.

Oono, Y., Kobayashi, F., Kawahara, Y., Yazawa, T., Handa, H., Itoh, T., and Matsumoto, T., 2013. Characterisation of the wheat (*Triticum aestivum* L.) transcriptome by de novo assembly for the discovery of phosphate starvation-responsive genes: gene expression in Pi-stressed wheat. *BMC Genomics*, **14**(1):77.

Pearce, S., Huttly, A. K., Prosser, I. M., Li, Y.-d., Vaughan, S. P., Gallova, B., Patil, A., Coghill, J. A., Dubcovsky, J., Hedden, P., *et al.*, 2015. Heterologous expression and transcript analysis of gibberellin biosynthetic genes of grasses reveals novel functionality in the GA3ox family. *BMC Plant Biology*, **15**(1):130.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, **33**(3):290–295.

Pfeifer, M., Kugler, K. G., Sandve, S. R., Zhan, B., Rudi, H., Hvidsten, T. R., The International Wheat Genome Sequencing Consortium (IWGSC), Mayer, K. F. X., and Olsen, O.-A., 2014. Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science*, **345**(6194):1250091.

Pimentel, H. J., Bray, N., Puente, S., Melsted, P., and Pachter, L., 2016. Differential analysis of RNA-Seq incorporating quantification uncertainty. *Preprint*.

Ramirez-Gonzalez, R. H., Uauy, C., and Caccamo, M., 2015. PolyMarker: A fast polyploid primer design pipeline: Fig. 1. *Bioinformatics*, **31**(12):2038–2039.

Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L., 2011a. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, **27**(17):2325–2329.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., and Pachter, L., 2011b. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, **12**(3):R22.

Sambrook, J. and Russell, D. W., 2006. Preparation of Plasmid DNA by Alkaline Lysis with SDS: Minipreparation. *CSH protocols*, **2006**(1).

Sarris, P. F., Cevik, V., Dagdas, G., Jones, J. D. G., and Krasileva, K. V., 2016. Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens. *BMC Biology*, **14**(1):8.

Seabold, S. and Perktold, J., 2010. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, pages 57–61.

Sears, E. R., 1966. Nullisomic-Tetrasomic Combinations in Hexaploid Wheat. In *Chromosome Manipulations and Plant Genetics*, pages 29–45. Springer US, Boston, MA.

Slater, G. S. C. and Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, **6**(1):31.

Song, L., Sabuncuyan, S., and Florea, L., 2016. CLASS2: accurate and efficient splice variant annotation from RNA-seq reads. *Nucleic Acids Research*, **44**(10):e98–e98.

Spannagl, M., Nussbaumer, T., Bader, K. C., Martis, M. M., Seidel, M., Kugler, K. G., Gundlach, H., and Mayer, K. F., 2016. PGSB PlantsDB: updates to the database framework for comparative plant genome research. *Nucleic Acids Research*, **44**(D1):D1141–D1147.

Stanke, M., Tzvetkova, A., and Morgenstern, B., 2006. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome biology*, **7 Suppl 1**(May 2005):S11.1–8.

Steuernagel, B., Jupe, F., Witek, K., Jones, J. D. G., and Wulff, B. B. H., 2015. NLR-parser: rapid annotation of plant NLR complements. *Bioinformatics*, **31**(10):1665–1667.

Sturgill, D., Malone, J. H., Sun, X., Smith, H. E., Rabinow, L., Samson, M.-L., and Oliver, B., 2013. Design of RNA splicing analysis null models for post hoc filtering of Drosophila head RNA-Seq data with the splicing analysis kit (Spanki). *BMC Bioinformatics*, **14**(1):320.

Šupek, F., Bošnjak, M., Škunca, N., and Šmuc, T., 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS one*, **6**(7):e21800.

The UniProt Consortium, 2014. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, **42**(D1):D191–D198.

The International Wheat Genome Sequencing Consortium, 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science*, **345**(6194):1251788–1251788.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L., 2012. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, **31**(1):46–53.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**(5):511–515.

Venturini, L., Caim, S., Mapleson, D. L., Kaithakottil, G. G., and Swarbreck, D., 2016. Mikado. <https://github.com/lucventurini/mikado>.

Vitting-Seerup, K., Porse, B., Sandelin, A., and Waage, J., 2014. spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. *BMC Bioinformatics*, **15**(1):81.

Wagner, G. P., Kin, K., and Lynch, V. J., 2013. A model based criterion for gene expression calls using RNA-seq data. *Theory in Biosciences*, **132**(3):159–164.

Wang, L., Wang, S., and Li, W., 2012. RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, **28**(16):2184–2185.

Weisenfeld, N. I., Yin, S., Sharpe, T., Lau, B., Hegarty, R., Holmes, L., Sogoloff, B., Tabbaa, D., Williams, L., Russ, C., et al., 2014. Comprehensive variation discovery in single human genomes. *Nature Genetics*, **46**(12):1350–1355.

Wessel, D. and Flügge, U., 1984. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Analytical Biochemistry*, **138**(1):141–143.

Wu, T. D. and Watanabe, C. K., 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**(9):1859–1875.

Wysokar, A., Tibbetts, K., McCown, M., Homer, N., and Fennell, T., 2016. Picard: A set of Java command line tools for manipulating high-throughput sequencing data (HTS) data and formats.

Yang, F., Li, W., and Jørgensen, H. J. L., 2013. Transcriptional Reprogramming of Wheat and the Hemibiotrophic Pathogen *Septoria tritici* during Two Phases of the Compatible Interaction. *PLoS ONE*, **8**(11):e81606.

Yang, F., Shen, Y., Camp, D. G., and Smith, R. D., 2012. High-pH reversed-phase chromatography with fraction concatenation for 2D proteomic analysis. *Expert Review of Proteomics*, **9**(2):129–134.

Yang, Z., Peng, Z., Wei, S., Liao, M., Yu, Y., and Jang, Z., 2015. Pistillody mutant reveals key insights into stamen and pistil development in wheat (*Triticum aestivum* L.). *BMC Genomics*, **16**(1):211.

Zhang, H., Yang, Y., Wang, C., Liu, M., Li, H., Fu, Y., Wang, Y., Nie, Y., Liu, X., and Ji, W., et al., 2014a. Large-scale transcriptome comparison reveals distinct gene activations in wheat responding to stripe rust and powdery mildew. *BMC Genomics*, **15**(1):898.

Zhang, W., Ciclitira, P., and Messing, J., 2014b. PacBio sequencing of gene families - a case study with wheat gluten genes. *Gene*, **533**(2):541–6.

# A chromosome conformation capture ordered sequence of the barley genome

Martin Mascher<sup>1,2\*</sup>, Heidrun Gundlach<sup>3\*</sup>, Axel Himmelbach<sup>1</sup>, Sebastian Beier<sup>1</sup>, Sven O. Twardziok<sup>3</sup>, Thomas Wicker<sup>4</sup>, Volodymyr Radchuk<sup>1</sup>, Christoph Dockter<sup>5</sup>, Pete E. Hedley<sup>6</sup>, Joanne Russell<sup>6</sup>, Micha Bayer<sup>6</sup>, Luke Ramsay<sup>6</sup>, Hui Liu<sup>6</sup>, Georg Haberer<sup>3</sup>, Xiao-Qi Zhang<sup>7</sup>, Qisen Zhang<sup>8</sup>, Roberto A. Barrero<sup>9</sup>, Lin Li<sup>10</sup>, Stefan Taudien<sup>11</sup>, Marco Groth<sup>11</sup>, Marius Felder<sup>11</sup>, Alex Hastic<sup>12</sup>, Hana Šimková<sup>13</sup>, Helena Staňková<sup>13</sup>, Jan Vrána<sup>13</sup>, Saki Chan<sup>12</sup>, María Muñoz-Amatriain<sup>14</sup>, Rachid Ounit<sup>15</sup>, Steve Wanamaker<sup>14</sup>, Daniel Bolser<sup>16</sup>, Christian Colmsee<sup>1</sup>, Thomas Schmutz<sup>1</sup>, Lala Aliyeva-Schnorr<sup>1</sup>, Stefano Grasso<sup>17</sup>, Jaakko Tanskanen<sup>18</sup>, Anna Chailyan<sup>5</sup>, Dharanya Sampath<sup>19</sup>, Darren Heavens<sup>19</sup>, Leah Clissold<sup>19</sup>, Sujie Cao<sup>20</sup>, Brett Chapman<sup>9</sup>, Fei Dai<sup>21</sup>, Yong Han<sup>21</sup>, Hua Li<sup>20</sup>, Xuan Li<sup>20</sup>, Chongyun Lin<sup>20</sup>, John K. McCooke<sup>9</sup>, Cong Tan<sup>9</sup>, Penghao Wang<sup>7</sup>, Songbo Wang<sup>20</sup>, Shuya Yin<sup>21</sup>, Gao Feng Zhou<sup>7</sup>, Jesse A. Poland<sup>22</sup>, Matthew I. Bellgard<sup>9</sup>, Ljudmilla Borisjuk<sup>1</sup>, Andreas Houben<sup>1</sup>, Jaroslav Doležel<sup>13</sup>, Sarah Ayling<sup>19</sup>, Stefano Lonardi<sup>15</sup>, Paul Kersey<sup>16</sup>, Peter Langridge<sup>23</sup>, Gary J. Muehlbauer<sup>10,24</sup>, Matthew D. Clark<sup>19,25</sup>, Mario Caccamo<sup>19,26</sup>, Alan H. Schulman<sup>18</sup>, Klaus F. X. Mayer<sup>3,27</sup>, Matthias Platzer<sup>11</sup>, Timothy J. Close<sup>14</sup>, Uwe Scholz<sup>1</sup>, Mats Hansson<sup>28</sup>, Guoping Zhang<sup>21</sup>, Ilka Braumann<sup>5</sup>, Manuel Spannagl<sup>3</sup>, Chengdao Li<sup>7,29,30</sup>, Robbie Waugh<sup>6,31</sup> & Nils Stein<sup>1,32</sup>

Cereal grasses of the Triticeae tribe have been the major food source in temperate regions since the dawn of agriculture. Their large genomes are characterized by a high content of repetitive elements and large pericentromeric regions that are virtually devoid of meiotic recombination. Here we present a high-quality reference genome assembly for barley (*Hordeum vulgare* L.). We use chromosome conformation capture mapping to derive the linear order of sequences across the pericentromeric space and to investigate the spatial organization of chromatin in the nucleus at megabase resolution. The composition of genes and repetitive elements differs between distal and proximal regions. Gene family analyses reveal lineage-specific duplications of genes involved in the transport of nutrients to developing seeds and the mobilization of carbohydrates in grains. We demonstrate the importance of the barley reference sequence for breeding by inspecting the genomic partitioning of sequence variation in modern elite germplasm, highlighting regions vulnerable to genetic erosion.

Barley remains dated to the dawn of agriculture have been found at several archaeological sites<sup>1,2</sup>. In addition to indications that barley was an important food crop, recent excavations have fuelled speculation that beverages from fermented grains may have motivated early Neolithic hunter-gatherers to erect some of humankind's oldest monuments<sup>3,4</sup>. Moreover, brewing beer may also have played a role in the eastward spread of the crop after its initial domestication in the Fertile Crescent<sup>5,6</sup>.

Since 2012, both genetic research and crop improvement in barley have benefited from a partly ordered draft sequence assembly<sup>7</sup>. This community resource has underpinned gene isolation<sup>8,9</sup> and population genomic studies<sup>10</sup>. However, these and other efforts have also revealed limitations of the current draft assembly. The limitations are often direct consequences of two characteristic genomic features: the extreme abundance of repetitive elements, and the severely reduced frequency of meiotic recombination in pericentromeric regions<sup>11</sup>.

These factors have limited the contiguity of whole-genome assemblies to kilobase-sized sequences originating from low-copy regions of the genome. Thus, a detailed investigation of the composition of the repetitive fraction of the genome—including expanded gene families—and of the distribution of targets of selection and crop improvement in (genetically defined) pericentromeric regions has been beyond reach.

Here we present a map-based reference sequence of the barley genome including the first comprehensively ordered assembly of the pericentromeric regions of a Triticeae genome. The resource highlights a conspicuous distinction between distal and proximal regions of chromosomes that is reflected by the intranuclear chromatin organization. Moreover, chromosomal compartments are differentiated by an exponential gradient of gene density and recombination rate, striking contrasts in the distribution of retrotransposon families, and distinct patterns of genetic diversity.

<sup>1</sup>Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, 06466 Seeland, Germany. <sup>2</sup>German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103 Leipzig, Germany. <sup>3</sup>PGSB - Plant Genome and Systems Biology, Helmholtz Center Munich - German Research Center for Environmental Health, 85764 Neuherberg, Germany. <sup>4</sup>Department of Plant and Microbial Biology, University of Zurich, 8008 Zurich, Switzerland. <sup>5</sup>Carlsberg Research Laboratory, 1799 Copenhagen, Denmark. <sup>6</sup>The James Hutton Institute, Dundee DD2 5DA, UK. <sup>7</sup>School of Veterinary and Life Sciences, Murdoch University, Murdoch, WA6150, Australia. <sup>8</sup>Australian Export Grains Innovation Centre, South Perth, WA6151, Australia. <sup>9</sup>Centre for Comparative Genomics, Murdoch University, WA6150, Murdoch, Australia. <sup>10</sup>Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108, Minnesota, USA. <sup>11</sup>Leibniz Institute on Aging - Fritz Lipmann Institute (FLI), 07745 Jena, Germany. <sup>12</sup>BioNano Genomics Inc., San Diego, CA 92121, California, USA. <sup>13</sup>Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, 78371 Olomouc, Czech Republic. <sup>14</sup>Department of Botany & Plant Sciences, University of California, Riverside, Riverside, CA 92521, California, USA. <sup>15</sup>Department of Computer Science and Engineering, University of Hinxton CB10 1SD, UK. <sup>16</sup>Department of Agricultural and Environmental Sciences, University of Udine, 33100 Udine, Italy. <sup>18</sup>Green Technology, Natural Resources Institute (Luke), Viikki Plant Science Centre, and Institute of Biotechnology, University of Helsinki, 00014, Helsinki, Finland. <sup>19</sup>Earlham Institute, Norwich NR4 7UH, UK. <sup>20</sup>BGI-Shenzhen, Shenzhen, 518083, China. <sup>21</sup>College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, 310058, China. <sup>22</sup>Kansas State University, Wheat Genetics Resource Center, Department of Plant Pathology and Department of Agronomy, Manhattan, KS 66506, Kansas, USA. <sup>23</sup>School of Agriculture, University of Adelaide, Urrbrae, SA5064, Australia. <sup>24</sup>Department of Plant and Microbial Biology, University of Minnesota, St. Paul, MN 55108, Minnesota, USA. <sup>25</sup>School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, UK. <sup>26</sup>National Institute of Agricultural Botany, Cambridge CB3 0LE, UK. <sup>27</sup>Wissenschaftszentrum Weihenstephan (WZW), Technical University Munich, 85354 Freising, Germany. <sup>28</sup>Department of Biology, Lund University, 22362 Lund, Sweden. <sup>29</sup>Department of Agriculture and Food, Government of Western Australia, South Perth WA 6151, Australia. <sup>30</sup>Hubei Collaborative Innovation Centre for Grain Industry, Yangtze University, Jingzhou, Hubei, 434023, China. <sup>31</sup>School of Life Sciences, University of Dundee, Dundee DD2 5DA, UK. <sup>32</sup>School of Plant Biology, University of Western Australia, Crawley, WA6009, Australia.

\*These authors contributed equally to this work.

**Table 1 | Assembly and annotation statistics**

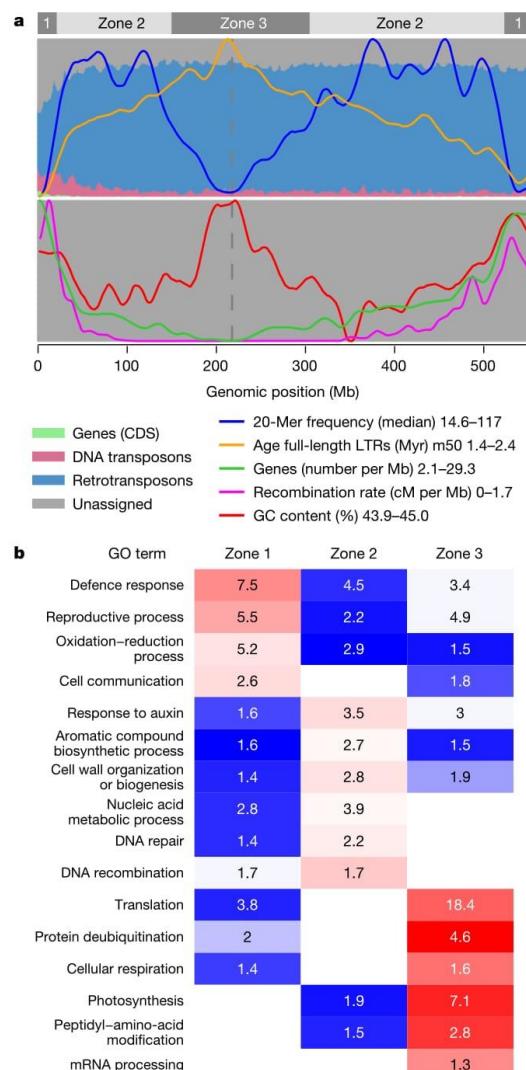
Number and cumulative length of sequenced BACs	87,075 (11.3 Gb)
Length of non-redundant sequence	4.79 Gb
Number of sequence contigs	466,070
BAC sequence contig N50	79 kb
Number and cumulative length of BAC super-scaffolds	4,235 (4.58 Gb)
Number and cumulative length of singleton BACs	2,123 (205 Mb)
Super-scaffold N50	1.9 Mb
Sequence anchored to the POPSEQ genetic map	4.63 Gb (97%)
Sequence anchored to the Hi-C map	4.54 Gb (95%)
Number of annotated high-confidence genes	39,734
Annotated coding sequence	65.3 Mb (1.4%)
Annotated transposable elements	3.70 Gb (80.8%)

### A chromosome-scale assembly of the barley genome

We adopted a hierarchical approach to generate a high-quality reference genome sequence of the barley cultivar Morex, a US spring six-row malting barley. First, a total of 87,075 bacterial artificial chromosomes (BACs) were sequenced, mainly using Illumina paired-end and mate-pair technology and assembled individually from 4.5 terabases of raw sequence data<sup>12–14</sup> (Supplementary Note 1). In a second step, overlaps between adjacent clones<sup>15</sup> were detected and validated by physical map information<sup>16</sup>, a genetic linkage<sup>17</sup> and a highly contiguous optical map<sup>18</sup> to construct super-scaffolds composed of merged assemblies of individual BACs (Table 1 and Extended Data Table 1). This increased the contiguity as measured by the N50 value (the scaffold size above which 50% of the total length of the sequence was included in the assembly) from 79 kb to 1.9 Mb. Scaffolds were assigned to chromosomes using a population sequencing (POPSEQ) genetic map<sup>17</sup>. Finally, we used three-dimensional proximity information obtained by chromosome conformation capture sequencing (Hi-C) to order and orient BAC-based super-scaffolds (Supplementary Note 2 and ref. 22). The final chromosome-scale assembly of the barley genome consists of 6,347 ordered super-scaffolds composed of merged assemblies of individual BACs, representing 4.79 Gb (~95%) of the genomic sequence content, of which 4.54 Gb have been assigned to precise chromosomal location in the Hi-C map (Table 1). Mapping of transcriptome data and reference protein sequences from other plant species to the assembly identified 83,105 putative gene loci including protein-coding genes, non-coding RNAs, pseudogenes and transcribed transposons (Fig. 1, Extended Data Fig. 1, Extended Data Table 2 and Supplementary Note 3). These loci were filtered further and divided into 39,734 high-confidence genes (with four different sub-categories) and 41,949 low-confidence genes on the basis of sequence homology to related species (Methods and Supplementary Note 3.4). Moreover, we predicted 19,908 long non-coding RNAs (Supplementary Note 3.7) and 792 microRNA precursor loci (Supplementary Note 3.8). The high co-linearity between the Hi-C-based pseudomolecules and linkage and cytogenetic maps<sup>22</sup> as well as the conserved order of syntenic genes in pericentromeric regions compared with model grass *Brachypodium distachyon* (Extended Data Fig. 2a) corroborated the quality of the assembly. Extrapolating from a set of conserved eukaryotic core genes<sup>23</sup>, we estimate that the predicted gene models represent 98% of the cultivar Morex barley gene complement (Extended Data Fig. 2b).

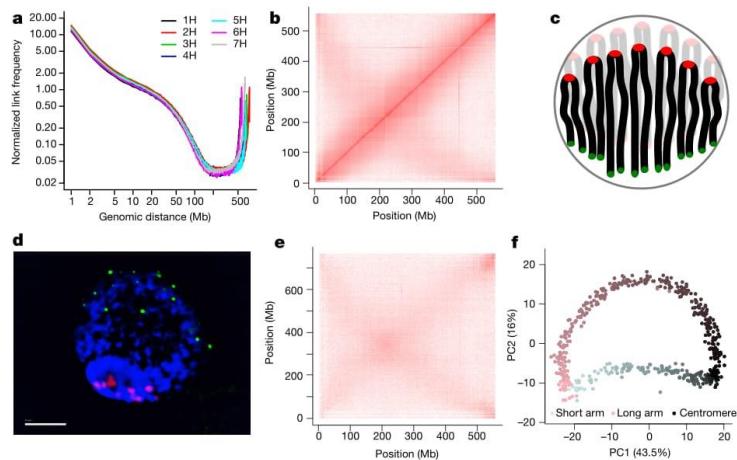
### Organization of chromatin

Barley has served as a model for traditional cytogenetics<sup>11</sup>, but relating chromosomal features to unique sequences has been challenging, requiring the cloning of repeat-free probes<sup>24</sup>. The reference sequence allowed us to employ the Hi-C data to interrogate the three-dimensional organization of chromatin in the nucleus. As in other eukaryotes<sup>20,25,26</sup>, the spatial proximity of genomic loci as measured by Hi-C link frequency is highly dependent on their distance in the linear genome (Fig. 2a). However, we observed an elevated link frequency at



**Figure 1 | Characteristics of genomic compartments in barley chromosomes.** **a**, The distribution of genomic features in 4 Mb windows is plotted along chromosome 1H. Analogous panels for the other chromosomes are found in Extended Data Fig. 5a. The left column in the legend refers to the background shading in the top panel; the right column indicates the colour code for lines in both panels. CDS, predicted coding sequences; cM, centimorgans. **b**, Enrichment of Gene Ontology (GO) terms in genomic compartments. Coloured rectangles indicate enrichment factors ranging from -2 (dark blue) to 2 (dark red). Numbers inside the rectangles indicate  $-\log_{10}$ -transformed *P* values.

distances above 200 Mb and a pronounced anti-diagonal pattern in the intrachromosomal Hi-C contact matrices (Fig. 2b and Extended Data Fig. 3a), indicating an increased adjacency of regions on different chromosome arms. We interpret this pattern as reflective of the so-called Rabl configuration<sup>27</sup> of interphase nuclei, where individual chromosomes fold back to juxtapose the long and short arms, with centromeres and telomeres of all chromosomes clustering at opposite poles of the nucleus (Fig. 2c and Supplementary Fig. 2.2). Fluorescence



**Figure 2 | Chromosome conformation capture analysis.** **a**, Distance-dependent decay of contact probability. **b**, Intrachromosomal contact matrix. The intensity of pixels represents the normalized count of Hi-C links between 1 Mb windows on chromosome 1H on a logarithmic scale. **c**, Schematic model of the Rabl configuration of interphase chromosomes. Centromeres and telomeres are presented by red and green circles, respectively. **d**, Leaf interphase nucleus of barley. Chromatin was stained blue with 4',6-diamidino-2-phenylindole (DAPI). Fluorescence *in situ* hybridization was performed with probes specific for centromeres (red)

and telomeres (green). Scale bar, 5  $\mu$ m. **e**, Interchromosomal contact matrix. The intensity of pixels represents the normalized count of Hi-C links between 1 Mb windows on chromosomes 1H (x axis) and 2H (y axis) on a logarithmic scale. A principal component analysis of the normalized contact matrix at 1 Mb resolution of chromosome 1H was conducted. **f**, The first and second eigenvectors are plotted against each other. Each point represents a 1 Mb window. Closer proximity to the centromere is indicated by a darker colour. Windows from the short and long arms are coloured blue and red, respectively.

*in situ* hybridization (Fig. 2d) supported this hypothesis. Principal component analysis of the intrachromosomal proximity matrix showed that the first three principal components cumulatively explained  $\sim$ 70% of the variation and differentiated (1) distal from proximal regions, (2) interstitial from both distal and proximal regions and (3) the long arms from the short arms (Fig. 2f and Extended Data Fig. 4a). A linear model taking into account the genomic distance between two loci, as well as their relative distance from the centromere, accounted for 79% of the variation (Extended Data Fig. 4b) in the intrachromosomal proximity matrix at 1 Mb resolution.

Contacts between loci on different chromosomes followed a similar pattern (Fig. 2e and Extended Data Fig. 3b): a prominent cross pattern supporting a juxtaposition of long and short arms. In contrast to intrachromosomal matrices, contact probabilities between loci on, for instance, the short arm of one chromosome are equal for loci on both the short and the long arm on another chromosome having the same relative distance to the centromeres: that is, facing each other in the interphase nucleus. We also observed a higher contact frequency between telomere-near regions, as has been observed in *Arabidopsis*<sup>25</sup>.

To test whether pairs of homologous chromosomes are positioned closer to each other than to non-homologues, we performed diploid Hi-C<sup>28</sup> on leaf tissue from F<sub>1</sub> hybrids between the cultivars Morex and Barke, and assigned the resultant Hi-C links to the haplotypes of both inbred parents by mapping reads to a diploid reference. We did not observe any preferential interaction between homologues. Rather, contacts between the maternal and paternal copies of the same chromosome occurred as frequently as between non-homologues (Extended Data Fig. 4c).

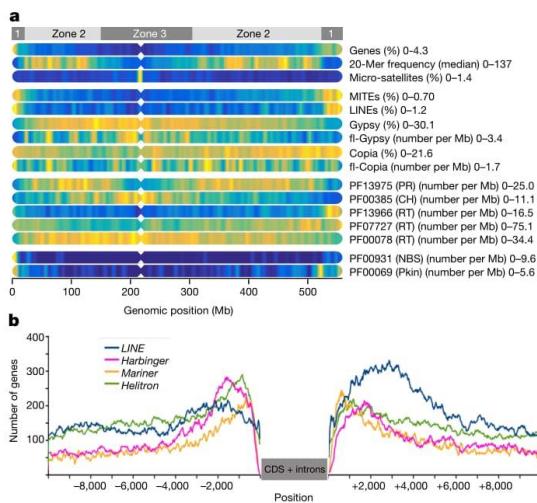
We conclude that the frequency with which loci juxtapose in three-dimensional space is predominantly determined by their position in the linear genome. This is in sharp contrast to the organization of chromatin in human nuclei where two compartments corresponding to open and closed chromatin domains are evident at megabase resolution<sup>20</sup>, but is consistent with cytogenetic mapping of histone marks associated with heterochromatin in large, repeat-rich genomes<sup>29</sup>.

### The genomic context of repetitive elements

Large plant genomes consist mainly of highly similar copies of repetitive elements such as long terminal repeat (LTR) retrotransposons and DNA transposons<sup>30,31</sup>. Our hierarchical sequencing strategy reduced the algorithmic complexity of assembling a highly repetitive genome from short reads. Instead of resolving complex repeat structures on the whole-genome level, we reconstructed the sequences of 100–150 kb BACs. This allowed us to disentangle nearly identical copies of highly abundant repetitive elements, as evidenced by the good representation of both mathematically defined repeats and retrotransposon families (Extended Data Fig. 2c, d). Homology-guided repeat annotation with a Triticeae-specific repeat library<sup>32</sup> identified 3.7 Gb (80.8%) of the assembled sequence as derived from transposable elements (Table 1, Fig. 1a and Extended Data Table 3), most of which were present as truncated and degenerated copies, with only 10% of mobile elements intact and potentially active.

Median 20-mer frequencies were used to partition the seven barley chromosomes into three zones (Fig. 1 and Extended Data Fig. 5a), reminiscent of the three compartments of wheat chromosome 3B<sup>33</sup>. The distal zone 1 was characterized by an enrichment of low-copy regions, a high gene content and frequent meiotic recombination. Zone 2, occupying the interstitial regions of chromosomes, had the highest 20-mer frequencies and intermediate gene density. Surprisingly, the abundance of repetitive 20-mers decreased in the proximal zone 3, where older mobile elements with diverged, and thus unique, sequences predominated (Fig. 1). The three zones also differed in the composition of the gene space (Extended Data Table 2b and Supplementary Note 3). For example, genes involved in defence response and reproductive processes were preferentially found in distal regions, while proximal regions contained more genes related to housekeeping processes, such as photosynthesis and respiration, compared with other parts of the genome (Fig. 1b).

Transposable element groups exhibited pronounced variation in their insertion site preferences (Fig. 3a and Extended Data Fig. 5b). On a global scale, most miniature inverted-repeat transposable elements



**Figure 3 | The genomic context of repetitive elements.** **a**, Abundance of key genomic features, different transposon superfamilies and common Pfam domains across chromosome 1H. Analogous panels for the other chromosomes are found in Extended Data Fig. 5b. The colour scale of the heatmaps ranges from blue (0) to yellow (maximum across all chromosomes per track). Minimum and maximum values are indicated to the right of each track. MITEs, miniature inverted-repeat transposable elements; LINEs, long interspersed elements; fl, full-length; PR, protease; CH, chromodomain; RT, reverse transcriptase; NBS, NB-ARC; Pkin, protein kinase. **b**, Transposable elements up- and downstream of genes. Coding sequences of high-confidence genes were used as anchor points. Transposable element composition was determined 10 kb up- and downstream of each gene. The x axis indicates the position relative to the gene, while the y axis indicates how many genes had a transposable element of the respective superfamily at the respective position in their upstream/downstream region.

and long interspersed elements were found in gene-rich distal regions, as has been reported in other grass species<sup>34,35</sup>. By contrast, zone 3 was populated by *Gypsy* retrotransposons, while *Copia* elements favoured zones 1 and 2. These differences in the relative abundance of retrotransposon families were reflected by distinct distributions of functional domains. For example, sequences encoding the chromodomain (PF00385) are concentrated in the vicinity of the centromere and may be involved in the target specificity through incorporation in the integrase of *Gypsy* elements<sup>36</sup> (Fig. 3a and Extended Data Fig. 5b).

At a local scale, different types of elements also occupy different niches in the proximity of genes (Fig. 3b). *Mariner* transposons preferably reside within 1 kb up- or downstream of the coding regions of genes, while *Harbinger* and long interspersed elements are found further away. The observed distribution of different types of transposable elements around genes may reflect selective pressures, allowing only the smallest elements, namely *Mariners*, to be tolerated closest to genes. Intriguingly, *Helitrons* as well as elements of the *Harbinger* superfamily have a clear preference for promoter regions, while long interspersed elements have a preference for downstream regions (Fig. 3b). At greater distances from genes, large elements such as LTR retrotransposons and CACTA elements dominate.

#### Expansion of gene families

The barley reference sequence enabled us to disentangle complex gene duplications that may shed light on gene family expansion specific to barley or the Triticeae. A total of 29,944 genes belonged to families with multiple members (Fig. 4a and Supplementary

Note 4.1). Gene families expanded in barley were tested for overrepresentation of Gene Ontology<sup>37</sup> terms compared with sorghum, rice, *Brachypodium* and *Arabidopsis*. Among the most significant results were terms related to defence response and disease resistance (NBS-LRR and thionin genes), as well as thioredoxin genes (Supplementary Note 4.1).

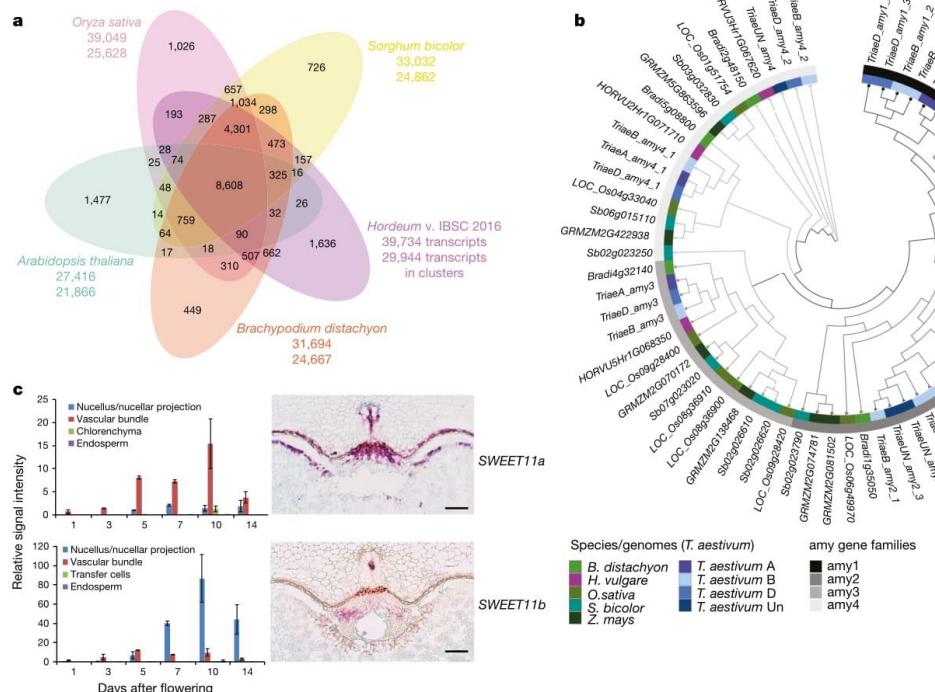
In the following, we focused on a detailed analysis of gene families having particular importance for malting quality. Germinating barley grains possess high diastatic power: that is, the combined ability of a complex of enzymes to mobilize fermentable sugars from starch. Key diastatic enzymes include  $\alpha$ -amylases. The genome of barley cultivar Morex contains 12  $\alpha$ -amylase (amy) family sequences (Supplementary Note 4.2 and Extended Data Table 4a), which can be classified into four subfamilies<sup>38</sup>. Gene duplication events have occurred in the subfamilies amy1 and amy2 (Fig. 4b), located on chromosomes 6H and 7H, respectively. The existence of these duplications had been speculated earlier, but could not be analysed further because of high sequence similarity between the copies. The reference assembly contained five full-length amy1 subfamily genes, four of which, here designated as *amy1\_1a-d*, shared >99.8% identity at the nucleotide level including introns. Locus-specific PCR confirmed earlier suggestions<sup>39,40</sup> of multiple, highly similar *amy1\_1* genes (Extended Data Fig. 6 and Supplementary Note 4.2). Given the relevance of  $\alpha$ -amylase activity to the brewing process, the high variability of the *amy1\_1* multiple gene locus (Extended Data Fig. 6) observed in landraces and elite lines, including modern malting cultivars, is remarkable.

The accumulation of fermentable carbohydrates in the grain depends on the transfer of sugars from maternal tissue into the developing seeds. In contrast to the two routes of nutrient transfer in rice seeds—the nucellar projection and nucellar epidermis—delivery of assimilates into barley grains occurs predominantly via the nucellar projection<sup>41</sup> and requires active transporters. The family of SUGARS WILL EVENTUALLY BE EXPORTED TRANSPORTER (SWEET) transmembrane proteins mediating sugar efflux<sup>42</sup> consists of 23 members in barley (Extended Data Table 4b and Supplementary Note 4.3). There is a small extension of the sugar-transporting SWEET11, SWEET13, SWEET14 and SWEET15 subfamilies, with two or more genes for each subgroup compared with only a single orthologue in rice and *Arabidopsis* (Extended Data Table 4b). Duplication of SWEET11 was most likely followed by neofunctionalization as evidenced by divergent expression patterns. Both *SWEET11a* and *SWEET11b* were highly expressed in maternal seed tissue, but differed in the distribution of expression domains (Fig. 4c and Extended Data Fig. 7). Genes encoding a family of vacuolar processing enzymes, which are essential for programmed cell death in maternal tissue<sup>43</sup> and starch accumulation in the grain (Supplementary Note 4.3 and V.R., unpublished observations) showed a similar expansion in barley (Extended Data Table 4c), pointing to the central role of the nucellar projection for grain filling in the Triticeae.

These examples of genes involved in sugar transport and metabolism illustrate that the high-quality reference genome sequence can serve as a springboard for the in-depth analysis of the evolutionary history of gene duplications, their relation to morphological and physiological innovations, and their impact on crop performance.

#### Molecular diversity and haplotype analysis

To explore how the new barley genome assembly could be exploited for genetics and breeding, we generated exome sequence data from 96 European elite barley lines, half with a spring growth habit, half with a winter one (Supplementary Table 5.1). We investigated the extent and partitioning of molecular variation within and between these groups using 71,285 single-nucleotide polymorphisms (SNPs). Plotting diversity values in 100 SNP windows both in linear order (Fig. 5a) and according to physical distance (Fig. 5b) revealed marked contrasts in the levels and distribution of diversity both within and between gene pools. In spring types, extensive regions on



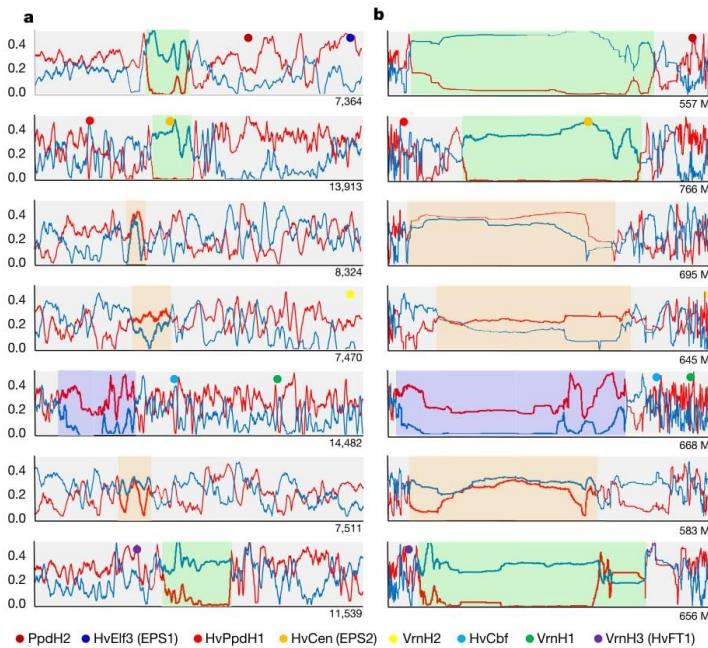
**Figure 4 | Expansion of agronomically important gene families.**  
**a**, OrthoMCL clustering of the barley high-confidence gene complement with *B. distachyon*, rice, sorghum and *Arabidopsis thaliana* genes. Numbers in the sections of the Venn diagram correspond to numbers of clusters (gene groups). The first number below the species name denotes the total number of proteins that were included into the OrthoMCL analysis for each species. The second number indicates the number of genes in clusters for a species. **b**, Phylogenetic tree of 68 full-length  $\alpha$ -amylase protein sequences derived from amy genes identified in the genomes of barley, hexaploid wheat, *B. distachyon*, rice, sorghum and maize. Each wheat subgenome was considered separately to facilitate the comparison of gene copy numbers and duplication events across species. Note that for the amy4 subfamily, two to three genes per genome were identified in all genomes. These genes are located on distinct chromosomes and hence most probably did not originate from tandem gene duplications. While most species further contain only a single amy3 gene copy per genome, moderate copy number extension was observed in sorghum and rice where a potential tandem gene duplication resulted in two amy3 gene copies.

Three genes of the amy2 subfamily were identified on chromosome 7H in barley and on chromosomes 7A, 7B, 7D in wheat. No similar copy number extension was observed in *B. distachyon*, *Sorghum bicolor* or *Oryza sativa*. In maize, two amy2 genes were identified. The amy1 subfamily shows the highest level of copy number extension. Tandem duplications are present in sorghum and rice. Two to three full-length genes were identified per genome in hexaploid wheat on group 6H chromosomes and five full-length amy1 genes on chromosome 6H and unanchored scaffolds in barley. Notably four of these barley genes share 99.8–100% sequence identity on protein and nucleotide level, indicating very recent duplication events. *T. aestivum*, *Triticum aestivum*; *Z. mays*, *Zea mays*. **c**, Expression of the SWEET11 gene subfamily in the developing barley grains. Left, expression profiles of SWEET11a and SWEET11b as determined by quantitative real-time PCR (qPCR) on total RNA isolated from micro-dissected developing grains. Right, localization of SWEET11a and SWEET11b expression in cross-section of immature seeds by RNA *in situ* hybridization. Hybridizations with sense probes are shown as negative controls in Extended Data Fig. 7a. Scale bars, 100  $\mu$ m.

chromosomes 1H, 2H and 7H were virtually devoid of diversity, as was a large region on 5H in the winter gene pool. For these chromosomes, this results in a single gene-pool-specific haplotype across the extensive pericentromeric regions. Chromosomes 3H, 4H and 6H maintain higher diversity across these regions owing to the presence of multiple similarly extensive haplotypes. This is even more evident when diversity is plotted on a physical scale (Fig. 5b). We presume that the lack of observed variation in elite germplasm is a signature of intense selection during breeding for different end-use sectors (principally malting versus feed barley), and the virtual absence of allelic re-assortment during meiosis owing to restricted recombination in the pericentromeric regions.

Crosses between spring and winter barleys are rarely performed as they are considered to disrupt the gene-pool-specific gene complexes required for general performance (such as phenological adaptations) and end-use quality. Contrasting local patterns of diversity outside the pericentromeric regions therefore also most likely reflect

the outcome of selection within alternative gene pools. We explored this further by comparing diversity in eight characterized genes whose variant alleles are important for conditioning barley's seasonal growth habit (Supplementary Note 5). Of the eight genes, *HvCEN* is uniquely 'locked' in the pericentromeric region of chromosome 2H where alternative alleles at a single SNP confer both differences in days-to-heading<sup>44</sup> and strong latitudinal differentiation<sup>10</sup>. The extensive pericentromeric haplotype in spring barleys (Fig. 5) may stem from selection for this single *HvCEN* SNP. While strong selection for other favourable alleles locked in the same region in spring barley cannot be ruled out, the virtual absence of recombination severely restricts exploitation of diversity across the entire region. Despite our focus here on life-history traits, strong selection for other traits mapping to pericentromeric regions<sup>45,46</sup>, including good malting quality in the spring gene pool on chromosomes 1H and 7H, would probably also reduce diversity in these regions. Interestingly, we are unaware of any phenotypic trait in the winter gene pool that would



**Figure 5 | Distribution of genetic diversity across the barley genome.** Ninety-six elite barley cultivars, including 48 from the winter gene pool (blue line) and 48 from the spring gene pool (red line), were used. Diversity (unbiased heterozygosity, *y* axis) is plotted as the rolling average of 100 adjacent SNPs along each chromosome. For improved visualization, all chromosomes have been normalized to a standard length. **a**, Patterns of diversity on chromosomes 1H–7H (top to bottom). The distance between each SNP has been normalized (that is, this does not show genetic distance). The number of SNPs included on each chromosome is given at the bottom right of each plot. **b**, The same diversity values normalized according to physical distance. Extensive peri-centromeric regions of very low diversity in the spring gene pool are highlighted in green and low diversity in the winter gene pool in purple. Regions with similar levels of diversity in both gene pools are highlighted in orange. Coloured dots show the position of eight loci previously identified as being differentiated between the winter and spring gene pools.

result in strong selection for a single pericentromeric haplotype on chromosome 5H.

We next explored patterns of linkage disequilibrium across the entire genome. As expected for two highly inbred and elite crop gene pools, we observed extensive linkage disequilibrium on all chromosomes in both spring and winter barleys (Extended Data Fig. 8). The number of discrete haplotype blocks in this germplasm set varied from 86 to 161 per chromosome (Extended Data Fig. 8). Surprisingly, the two-row spring gene pool, generally considered to be narrowest owing to intense selection for malting quality, exhibited a greater number of haplotype blocks than the winter lines for most chromosomes.

## Discussion

To assemble a highly contiguous reference genome sequence for barley, we combined hierarchical shotgun sequencing, a strategy previously used for assembling large and complex plant genomes<sup>33,47</sup>, with novel technologies such as optical mapping<sup>18</sup> and chromosome-scale scaffolding with Hi-C<sup>21</sup>. The latter technology was key to resolving the linear order of sequence scaffolds in pericentromeric regions. We anticipate the adoption of Hi-C-based genome mapping in other Triticeae species, such as bread and durum wheat and their wild relatives. Now that the quality of whole-genome shotgun assemblies is on a par with map-based assemblies<sup>48,49</sup>, we believe that the barley genome project will be one of the last such efforts to follow the laborious BAC-by-BAC approach.

The barley reference genome sequence constitutes an important community resource for cereal genetics and genomics. It will facilitate positional cloning, provide a better contextualization of population genomic datasets and enable comparative genomic analysis with other Triticeae in non-recombining regions that have been inaccessible to analysis of gene collinearity until now. The exciting methodological advances in sequence assembly and genome mapping have enabled even large and repeat-rich genomes to be unlocked<sup>48,50</sup> and hold the promise of constructing reference-quality genome sequences, not only for a single cultivar, but also for representatives of major germplasm groups.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 26 August 2016; accepted 3 March 2017.**

1. van Zeist, W. & Bakker-Heeres, J. A. H. Archaeological studies in the Levant 1. Neolithic sites in the Damascus basin: Aswad, Ghorafe, Ramad. *Palaeohistoria* **24**, 165–256 (1985).
2. Riehl, S., Zeidi, M. & Conard, N. J. Emergence of agriculture in the foothills of the Zagros Mountains of Iran. *Science* **341**, 65–67 (2013).
3. Dietrich, O., Heun, M., Notroff, J., Schmidt, K. & Zarnkow, M. The role of cult and feasting in the emergence of Neolithic communities. New evidence from Göbekli Tepe, south-eastern Turkey. *Antiquity* **86**, 674–695 (2012).
4. Hayden, B., Canuel, N. & Shanse, J. What was brewing in the Natufian? An archaeological assessment of brewing technology in the Epipaleolithic. *J. Archaeol. Method Theory* **20**, 102–150 (2013).
5. Wang, J. *et al.* Revealing a 5,000-y-old beer recipe in China. *Proc. Natl. Acad. Sci. USA* **113**, 6444–6448 (2016).
6. Zohary, D., Hopf, M. & Weiss, E. *Domestication of Plants in the Old World: The Origin and Spread of Domesticated Plants in Southwest Asia, Europe, and the Mediterranean Basin* (Oxford Univ. Press, 2012).
7. International Barley Genome Sequencing Consortium. A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**, 711–716 (2012).
8. Yang, P. *et al.* PROTEIN DISULFIDE ISOMERASE LIKE 5-1 is a susceptibility factor to plant viruses. *Proc. Natl. Acad. Sci. USA* **111**, 2104–2109 (2014).
9. Pourkheirandish, M. *et al.* Evolution of the grain dispersal system in barley. *Cell* **162**, 527–539 (2015).
10. Russell, J. *et al.* Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation. *Nat. Genet.* **48**, 1024–1030 (2016).
11. Künzel, G., Korzun, L. & Meister, A. Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. *Genetics* **154**, 397–412 (2000).
12. Beier, S. *et al.* Multiplex sequencing of bacterial artificial chromosomes for assembling complex plant genomes. *Plant Biotechnol. J.* **14**, 1511–1522 (2016).
13. Muñoz-Amatriain, M. *et al.* Sequencing of 15 622 gene-bearing BACs clarifies the gene-dense regions of the barley genome. *Plant J.* **84**, 216–227 (2015).
14. Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative *k*-mers. *BMC Genomics* **16**, 236 (2015).
15. Colmsee, C. *et al.* BARLEX - the Barley Draft Genome Explorer. *Mol. Plant* **8**, 964–966 (2015).

16. Ariyadasa, R. *et al.* A sequence-ready physical map of barley anchored genetically by two million single-nucleotide polymorphisms. *Plant Physiol.* **164**, 412–423 (2014).
17. Mascher, M. *et al.* Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *Plant J.* **76**, 718–727 (2013).
18. Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30**, 771–776 (2012).
19. Kalhor, R., Tjøn, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* **30**, 90–98 (2011).
20. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
21. Burton, J. N. *et al.* Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
22. Beier, S. *et al.* Construction of a map-based reference genome sequence for barley, *Hordeum vulgare*. *L. Sci. Data* **4**, 170044 (2017).
23. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
24. Fuchs, J., Houben, A., Brandes, A. & Schubert, I. Chromosome ‘painting’ in plants – a feasible technique? *Chromosoma* **104**, 315–320 (1996).
25. Grob, S., Schmid, M. W. & Grossniklaus, U. Hi-C analysis in *Arabidopsis* identifies the KNOT, a structure with similarities to the flamenco locus of *Drosophila*. *Mol. Cell* **55**, 678–693 (2014).
26. Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465**, 363–367 (2010).
27. Tang, C. L., He, Y. & Pawlowski, W. P. Chromosome organization and dynamics during interphase, mitosis, and meiosis in plants. *Plant Physiol.* **158**, 26–34 (2012).
28. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
29. Houben, A. *et al.* Methylation of histone H3 in euchromatin of plant chromosomes depends on basic nuclear DNA content. *Plant J.* **33**, 967–973 (2003).
30. Flavell, R. B., Bennett, M. D., Smith, J. B. & Smith, D. B. Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochem. Genet.* **12**, 257–269 (1974).
31. SanMiguel, P. *et al.* Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**, 765–768 (1996).
32. Wicker, T., Mattheus, D. E. & Keller, B. TREP: a database for Triticeae repetitive elements. *Trends Plant Sci.* **7**, 561–562 (2002).
33. Choulet, F. *et al.* Structural and functional partitioning of bread wheat chromosome 3B. *Science* **345**, 1249721 (2014).
34. Bureau, T. E. & Wessler, S. R. Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell* **6**, 907–916 (1994).
35. Bureau, T. E. & Wessler, S. R. Mobile inverted-repeat elements of the Tourist family are associated with the genes of many cereal grasses. *Proc. Natl. Acad. Sci. USA* **91**, 14111–1415 (1994).
36. Malik, H. S. & Eickbush, T. H. Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. *J. Virol.* **73**, 5186–5190 (1999).
37. Harris, M. A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
38. Huang, N., Sutliff, T. D., Litts, J. C. & Rodriguez, R. L. Classification and characterization of the rice  $\alpha$ -amylase multigene family. *Plant Mol. Biol.* **14**, 655–668 (1990).
39. Muthukrishnan, S., Gill, B. S., Swegle, M. & Chandra, G. R. Structural genes for  $\alpha$ -amylases are located on barley chromosomes 1 and 6. *J. Biol. Chem.* **259**, 13637–13639 (1984).
40. Khurshed, B. & Rogers, J. C. Barley  $\alpha$ -amylase genes. Quantitative comparison of steady-state mRNA levels from individual members of the two different families expressed in aleurone cells. *J. Biol. Chem.* **263**, 18953–18960 (1988).
41. Melkus, G. *et al.* Dynamic  $^{13}\text{C}/^1\text{H}$  NMR imaging uncovers sugar allocation in the living seed. *Plant Biotechnol. J.* **9**, 1022–1037 (2011).
42. Chen, L. Q. *et al.* Sucrose efflux mediated by SWEET proteins as a key step for phloem transport. *Science* **335**, 207–211 (2012).
43. Tran, V., Weier, D., Radchuk, R., Thiel, J. & Radchuk, V. Caspase-like activities accompany programmed cell death events in developing barley grains. *PLoS ONE* **9**, e109426 (2014).
44. Comadran, J. *et al.* Natural variation in a homolog of *Antirrhinum CENTRORADIALIS* contributed to spring growth habit and environmental adaptation in cultivated barley. *Nat. Genet.* **44**, 1388–1392 (2012).
45. Schmalenbach, I., Léon, J. & Pillen, K. Identification and verification of QTLs for agronomic traits using wild barley introgression lines. *Theor. Appl. Genet.* **118**, 483–497 (2009).
46. Han, F. *et al.* Dissection of a malting quality QTL region on chromosome 1 (7H) of barley. *Mol. Breed.* **14**, 339–347 (2004).
47. Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
48. Zimin, A. V. *et al.* Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the mega-reads algorithm. Preprint at <http://biorxiv.org/content/early/2016/07/26/066100> (2016).
49. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
50. Hirsch, C. *et al.* Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. *Plant Cell* **28**, 2700–2714 (2016).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This work was performed in the frame of the International Barley Genome Sequencing Consortium and was supported by German Ministry of Education and Research grants 0314000 and 0315954 to K.F.X.M., M.P., U.S. and N.S., and 031A536 to U.S. and K.F.X.M.; Leibniz ‘Pakt f. Forschung und Innovation’ grant ‘sequencing barley chromosome 3H’ to N.S. and U.S.; Scottish Government/UK Biotechnology and Biological Sciences Research Council (BBSRC) grant BB/100663X/1 to R.W., P.E.H. and J.R.; BBSRC grants BB/I008375/1 to M.D.C. and M.C., and BB/I008071/1 to P.K.; Finland grant 266430 and a BioNano grant to A.H.S.; Carlsberg Foundation grant 2012\_01\_0461 to the Carlsberg Research Laboratory; Grains Research and Development Corporation (GRDC) grant DAW00233 to C.L. and P.L.; Department of Agricultural and Food, Government of Western Australia grant 681 to C.L.; National Natural Science Foundation of China (NSFC) grant 31129005 to C.L. and G. Zhang; NSFC grant 31330055 to G. Zhang.; Czech Ministry of Education, Youth and Sports grant LO1204 to J.D.; US National Science Foundation (NSF) grant DBI 0321756 to T.J.C. and S.L.; US Department of Agriculture–Cooperative State Research, Education, and Extension Service–National Institute of Food and Agriculture (USDA–CSREES–NIFA) grants 2009-65300-05645 and 2011-68002-30029 to T.J.C., S.L. and G.J.M.; NSF Advances in Biological Informatics grant DBI-1062301 to T.J.C. and S.L.; University of California grant CA-RBPS-5306-H to T.J.C. and S.L.; NSF grant DBI 0321756 to S.L. BBSRC National Capability in Genomics (BB/J010375/1) and BBSRC Institute Strategic Programme funding for Bioinformatics (BB/J004669/1) to M.D.C., S.A. and M.C.; winter and spring barley accessions were a subset of genotypes selected from BBSRC and Agriculture and Horticulture Development Board projects AGOUEB and IMPROMALT (RD-2012-3776). We acknowledge (1) the technical assistance of S. König, M. Knauft, U. Beier, A. Kusserow, K. Trnka, I. Walde, S. Driesslein and C. Voss; (2) D. Stengel, A. Fiebig, T. Münch, D. Schüler, D. Arend, M. Lange and P. Rapazote-Flores for data management and submission; (3) K. Lipfert for artwork; (4) H. Berges, A. Bellec and S. Vautrin (CNRGV) for management and distribution of BAC libraries; (5) A. Grainer and D. Marshall for scientific discussions.

**Author Contributions** Project coordination: M.S., I.B., C. Li, R.W. (co-leader), N.S. (leader); BAC sequencing and assembly (1H, 3H, 4H): S.B., A. Himmelbach, S.T., M.F., M.G., M.M., U.S. (co-leader), M.P. (co-leader), N.S. (leader); BAC sequencing and assembly (2H, unassigned): D.S., D.H., S.A. (co-leader), M.D.C. (co-leader), M.C. (co-leader), R.W. (leader); BAC sequencing and assembly (5H, 7H): X.Z., R.A.B., Q.Z., C.T., J.K.M., B.C., G. Zhou, F.D., Y.H., S.Y., S. Cao, S. Wang, X.L., M.B., P.L., G. Zhang (co-leader), C. Li (leader); BAC sequencing and assembly (6H): S.B., S. Wang, C. Lin, H. Li, U.S., M.H. (co-leader), I.B. (leader); BAC sequencing (gene-bearing): M.M.-A., R.O., S. Wanamaker, S.L. (co-leader), T.J.C. (leader); optical mapping: A. Hastie, H. S., J.T., H. S., J.V., S. Chan, M. N., N.S., J.D., A.H.S. (leader); data integration: M.M. (leader), S.B., C.C., D.B., LL., T.S., J.A.P., P.K., N.S. (co-leader); transcriptome sequencing and analysis: P.E.H., M.B., J.R., H. Liu, S.T., M.F., M.G., M.P., R.W. (leader); annotation of transcribed regions: S.O.T., G.H., R.A.B., L.L., G.J.M., K.F.X.M. (co-leader), M.S. (leader); repetitive DNA analysis: T.W. (co-leader), J.T., K.F.X.M., A.H.S., H.G. (leader); gene family analysis: Q.Z., M.S., V.R., C.D., G.H., A.C., D.B., P.W., L.B., N.S., P.K., C. Li (co-leader), I.B. (leader); chromosome conformation capture: A. Himmelbach, S.G., L.A.-S., A. Houben, M.M. (co-leader), N.S. (leader); resequencing and diversity analysis: J.R., M.B., P.E.H., L.R., L.C., R.W. (leader); writing: M.M. (co-leader), M.S., A.H.S., G.J.M., R.W., N.S. (leader). All authors read and commented on the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to N.S. (stein@ipk-gatersleben.de), R.W. (robbie.waugh@hutton.ac.uk), C. Li (li@murdoch.edu.au), G. Zhang (zhanggg@zju.edu.cn), I.B. (ilka.braumann@carlsberg.com) or M.S. (manuel.spannagl@helmholtz-muenchen.de).

**Reviewer Information** *Nature* thanks M. Bevan, B. Keller and the other anonymous reviewer(s) for their contribution to the peer review of this work.

 This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

**Sequencing and assembly of individual BAC clones.** Barley genome sequencing relied exclusively on shotgun sequencing of 88,731 BAC clones using high-throughput next-generation sequencing-by-synthesis<sup>22</sup>. This comprised 15,661 so-called gene-bearing BAC clones preselected mainly by overgo-probe hybridization for the presence of transcribed genes and fingerprinted for definition of a minimum tiling path of the barley gene space. These gene-space minimum tiling path BAC clones were sequenced as combinatorial pools by Illumina short-read technology and, after quality trimming of de-convoluted reads, were assembled using Velvet version 1.2.09 as previously described<sup>13</sup>. The remaining 73,070 BACs were selected from a minimum tiling path representing the physical map of the barley genome<sup>16</sup>. Minimum tiling path BAC clones assigned to different barley chromosomes were sequenced at one of four sequencing centres, relying on highly multiplexed paired-end and mate-pair sequencing libraries using either the Roche 454 Titanium or the Illumina MiSeq, HiSeq2000 and HiSeq2500 platforms (Supplementary Note 1 and ref. 51). In brief, sequencing reads were de-convoluted on the basis of the used BAC-specific barcode sequence tags and assembled with sequencing centre-specific assembly pipelines. BAC clones sequenced on the Roche 454 Titanium platform were assembled with MIRA<sup>51</sup> according to previously described procedures<sup>52,53</sup>. Illumina HiSeq2000 paired-end sequencing data ( $2 \times 100$  nucleotides) of BAC clones were assembled either with CLC Assembly Cell version 4.0.6 beta (<http://www.clebio.com/products/clc-assembly-cell/>) set to default parameters<sup>12</sup>, SOAPdenovo version 2.01 (ref. 54) or the ABySS assembler (version 1.5.1)<sup>55</sup>. Sequence contigs of the *de novo* BAC assemblies larger than 500 base pairs (bp) were scaffolded using mate-pair sequencing information either generated from BAC DNA-derived 3 kbp insert mate-pair sequencing libraries or from 2 kbp, 5 kbp or 10 kbp genomic DNA-derived mate-pair libraries. This was achieved by either using BWA mem version 0.7.4 (ref. 56) with default parameters for read mapping, followed by scaffolding individual BACs using SSPACE version 3.0 Standard<sup>57</sup>, or with SOAPAligner/soap2 version 2.21 and using SOAPdenovo<sup>54</sup> scaffolder version 2.01.

**Genome-wide three-dimensional chromatin conformation capture sequencing.** To generate physical scaffolding information for the BAC sequence based genome assembly, as proposed in ref. 21, Hi-C and tethered conformation capture (TCC) sequencing data were generated from 7-day-old leaf tissue of greenhouse-grown barley plantlets by adapting previously published procedures (Supplementary Note 2). In brief, for Hi-C, freshly harvested leaves were cut into 2 cm pieces and vacuum infiltrated in nuclei isolation buffer supplemented with 2% formaldehyde. Crosslinking was stopped by adding glycine and additional vacuum infiltration. Fixed tissue was frozen in liquid nitrogen and ground to powder before re-suspending in nuclei isolation buffer to obtain a suspension of nuclei. About  $10^7$  purified nuclei were digested with 400 units of HindIII as described previously<sup>58</sup>. Digested chromatin was marked by incubating with biotin-14-dCTP and Klenow enzyme using a fill-in reaction<sup>20</sup> resulting in blunt-ended repaired DNA strands. Biotin-14-dCTP from non-ligated DNA ends was removed owing to the exonuclease activity of T4 DNA polymerase, followed by phenol-chloroform extraction and washing of the precipitated DNA as described<sup>20</sup>. As an alternative to Hi-C, the TCC protocol was also adapted for barley. Nuclei were prepared from barley leaf tissue as described above for Hi-C, before biotinyling the isolated chromatin using EZlink Iodoacetyl-PEG2-Biotin. The samples were neutralized with SDS, and DNA was digested with HindIII, dialysed, followed by immobilization to low surface coverage using streptavidin-coated magnetic beads<sup>19</sup>. Open DNA ends were labelled with biotin-14-dCTP using Klenow enzyme, and blunt-ended, labelled DNA products were collected from the magnetic beads by reversing the formaldehyde crosslink using proteinase K<sup>19</sup>. Biotin-14-dCTP from non-ligated DNA ends was removed by using Exonuclease III<sup>19</sup>. Hi-C and TCC products were mechanically sheared to fragment sizes of 200–300 bp by applying ultrasound using a Covaris S220 device followed by size-fractionation using AMPure XP beads. DNA fragments in the range between 150 and 300 bp were blunt-end repaired and A-tailed before purification through biotin-streptavidin-mediated pull-down<sup>58</sup>. Illumina paired-end adapters were ligated to the Hi-C and TCC products, respectively, followed by PCR amplification, pooling of PCR products and purification with AMPure XP beads before quantification of Hi-C/TCC libraries by qPCR for Illumina HiSeq2500 PE100 sequencing<sup>20</sup>.

**Nanochannel-based genome mapping.** Long-range scaffolding of genome sequence assemblies was facilitated by BioNano genome maps generated by nanochannel electrophoresis of fluorescently labelled high-molecular mass DNA obtained from flow-sorted chromosomes<sup>59</sup>. High-molecular mass DNA was

prepared from  $3.5 \times 10^6$  purified chromosomes (whole genome) of barley cultivar Morex essentially following published procedures<sup>60,61</sup>. The purified chromosomes were embedded in agarose miniplugs to achieve approximate concentrations of 1 million chromosomes per  $40\mu\text{l}$  volume before being treated with proteinase K as described previously<sup>61</sup>. DNA was labelled at *Nt.BspQI* nicking sites (GCTCTTC) by incorporation of fluorescent-dUTP nucleotide analogues using *Taq* polymerase as described previously<sup>59</sup>. The labelled DNA was analysed on the Irys platform (BioNano Genomics) in 191 cycles in total, generating 243 Gb of data exceeding 150 kb. On the basis of the label positions on single DNA molecules, *de novo* assembly was performed by a pairwise comparison of all single molecules and graph building<sup>62</sup>. The parameter set for large genomes was used for assembly with the IrysView software. A *P* value threshold of  $10^{-9}$  was used during the pairwise assembly,  $10^{-10}$  for extension and refinement steps and  $10^{-14}$  for merging contigs. A whole-genome map of 4.3 Gb was obtained (Extended Data Table 1).

**Data integration for constructing pseudomolecules.** The construction of pseudomolecules representing the seven barley chromosomes followed an iterative, mainly automated procedure which involved the integration of the following major datasets: (1) sequence assemblies of 87,075 unique, successfully sequenced and assembled BAC clones; (2) BAC assembly information from a genome-wide physical map of barley<sup>16</sup>; (3) 571,814 end-sequences of BAC clones<sup>7</sup>; (4) a dense linkage map assigning genetic positions to 791,177 contigs of a whole-genome shotgun assembly of barley cultivar Morex<sup>17</sup>; (5) Hi-C/TCC sequence information; and (6) the optical map of the genome of barley cultivar Morex. A schematic outline of the procedure is presented elsewhere<sup>22</sup>. In the first step, overlaps between individual BAC assemblies were searched with Megablast<sup>63</sup> by either applying 'stringent' or 'permissive' alignment criteria<sup>22</sup> and by combining with the high density genetic map information. On the basis of this initial analysis, a BAC overlap graph was constructed by use of the R package igraph<sup>64</sup> considering the above-listed additional datasets in subsequent iterative steps. Building the overlap graph focused first on overlaps obtained under 'stringent' search criteria for BACs within individual physical map contigs (FP contigs) and then subsequently also between independent FP contigs. Subsequently, overlaps obtained under 'permissive' criteria were evaluated while checking for cumulative evidences provided by the additional datasets supporting the overlap information<sup>22</sup>. Ordering and orienting of the resultant sequence scaffolds were achieved by integrating the overlap graph with Hi-C/TCC data<sup>22</sup>. Before the construction of pseudomolecules, we (1) identified genes incomplete or missing in the non-redundant sequence, but represented by (a) BAC sequence that had been excluded from the construction of the non-redundant sequence, or by (b) Morex WGS contigs, and (2) performed a final scan for contaminant sequences. Then a single FASTA file containing a single entry for each barley chromosome (a 'pseudomolecule') and an additional entry combining all sequences not anchored to chromosomes was constructed<sup>22</sup>.

**Three-dimensional chromatin conformation analysis.** Mapping of Hi-C/TCC reads and assignment to restriction fragments were performed as described elsewhere<sup>22</sup>. Briefly, raw reads were trimmed with cutadapt<sup>65</sup>. Trimmed Hi-C reads were mapped to the barley pseudomolecule sequence with BWA mem (version 0.7.12)<sup>66</sup>. Duplicate removal and sorting were performed with NovoSort (<http://www.novocraft.com/products/novosort/>). Mapped reads were assigned to restriction fragments with BEDtools<sup>67</sup>, tabulated with custom AWK scripts and imported into R (<http://www.r-project.org/>). Raw counts of Hi-C links were aggregated in 1 Mb bins and normalized separately for intra- and interchromosomal contacts using HiCNorm<sup>68</sup>. Contact probability matrices were plotted using standard R functions<sup>69</sup>. Principal component analysis was performed with the R function pcomp() on the matrix of log-transformed normalized Hi-C link counts between 1 Mb fragments.

We fitted the linear model  $\log_{10}(\text{nl}) \sim \log_{10}(\text{dist}) + \text{abs}(\text{cen\_dist1} - \text{cen\_dist2}) + \text{arm1:arm2} + \text{apos1:apos1}$  using the R function lm(). Here, nl is the normalized link count between two 1 Mb bins, dist is their distance in the linear genome, cen\_dist1 and cen\_dist2 are the relative distances from the centromere of both loci, arm1 and arm2 are the chromosome arm assignment of both loci, and apos1 and apos2 are the relative distances of both loci from the ends of the chromosome arm (that is, apos1 is close to zero if locus 1 is either near the centromere or the telomere, and close to one if locus 1 resides in interstitial regions). TCC reads of Morex  $\times$  Barke F<sub>1</sub> hybrids were mapped to a synthetic reference representing the parental genomes. An *in silico* Barke assembly was created by inserting SNPs discovered by aligning Barke WGS reads to the Morex reference assembly with BWA MEM<sup>66</sup> and calling variants with SAMtools<sup>70</sup>. SNPs were then inserted into the Morex reference using the FastaAlternateReferenceMaker of GATK<sup>71</sup>. TCC reads of the hybrid were then mapped to the synthetic reference as described above. Only uniquely alignable read pairs were considered. Hi-C link counts were tabulated at the level of chromosomes.

Fluorescence *in situ* hybridization was performed with *H. vulgare* nuclei as described earlier<sup>72</sup> using *Arabidopsis*-type telomere and barley centromere-specific [AGGGAG]<sub>5</sub> repeat probes<sup>73</sup>.

**Automated annotation of transcribed regions.** Automated gene annotation of the barley reference sequence assembly was based on four datasets providing independent gene evidence information (Supplementary Note 3). This included (1) RNA sequencing (RNA-seq) data; (2) reference protein predictions from barley<sup>7</sup>, rice<sup>4</sup>, *B. distachyon*<sup>75</sup> and *S. bicolor*<sup>76</sup>; (3) published barley full-length complementary DNA (fl-cDNA) sequences<sup>77</sup>; and (4) newly generated barley PacBio Iso-Seq data. Previously published<sup>7</sup> and newly generated RNA-seq datasets were derived from a total of 16 different tissues, each with three biological replicates, including seven vegetative, six inflorescence, two developing grain and one germinating grain tissues. RNA-seq libraries were sequenced on Illumina HiSeq2000 in paired-end 2 × 100 nucleotides (PE100) mode (Supplementary Note 3). To support gene calling in general, and the identification of alternative splice forms in particular, enriched full-length transcript information was generated by the Iso-Seq method using the PacBio RS II system and DNA Sequencing Chemistry 4.0 version 2 (Supplementary Note 3). RNA-seq-based transcript structures, reference-based gene model predictions, structure information from Iso-Seq alignments as well as structure information from flcDNA sequence alignments were clustered into a consensus transcript set using Cuffcompare<sup>78</sup> (Supplementary Note 3). Predicted transcript sequences were automatically extracted into a single FASTA file on the basis of respective coordinates in the genome assembly. Putative open reading frames and corresponding peptide sequences, including prediction of Pfam domains, were obtained by applying TransDecoder (<https://transdecoder.github.io>), which also resulted in reports about predicted alternative peptides per transcript (Supplementary Note 3). A single best translation per transcript was selected on the basis of BLASTP<sup>79</sup> comparison of all predicted peptides to a comprehensive protein database containing high-confidence protein sequences from *A. thaliana*<sup>80</sup>, maize<sup>47</sup>, *B. distachyon*<sup>75</sup>, rice<sup>75</sup> and *S. bicolor*<sup>76</sup>, followed by additional filtering procedures (Supplementary Note 3). Functional descriptions ('human readable descriptions') were generated for all potential genes using the AHRD pipeline (<https://github.com/groupschoof/AHRD>) on the basis of one representative protein sequence for each gene locus. Gene candidates were then classified into high- and low-confidence genes and further subdivided into nine classes, each supported by different levels of gene evidence (Supplementary Note 3). High-confidence protein-coding genes either showed significant sequence homology to a reference protein or were associated with a predicted function. Low-confidence genes were characterized by (1) having no or only weak sequence homology to reference proteins and no predicted function, (2) they were candidates for transposons or (3) they lacked an open reading frame of a minimal length (Supplementary Note 3). Completeness of gene-space representation was evaluated with the BUSCO pipeline<sup>83</sup> (Extended Data Fig. 2b).

**Feature distributions along the chromosomes.** A sliding window approach with a window size of 4 Mb and a shift of 0.8 Mb was used to display the distribution of different genome components and other features such as GC content or recombination rate along the chromosomes. The resulting data were smoothed with the python function `scipy.signal.gaussian` ( $p1 = 40$ ,  $p2 = 10$  for Fig. 1a;  $p1 = 15$ ,  $p2 = 3$  for Fig. 2a). The boundaries of genomic compartments (Fig. 1) are given in Supplementary Table 4.4.

**Annotation of the non-genic part of the genome.** Transposable elements were detected and classified by homology search with Vmatch (<http://www.vmatch.de>) against the REdat\_9.7\_Triticeae section of the PGSB transposon library<sup>81</sup>. The following parameter settings were used: identity  $\geq 70\%$ , minimal hit length 75 bp, seed length 12 bp (exact commandline: `-d -p -l 75 -identity 70 -seedlength 12 -exdrop 5`). The Vmatch output was filtered for redundant hits by prioritizing higher-scoring matches and then either shortening ( $< 90\%$  coverage and  $\geq 50$  bp rest length) or removing lower-scoring overlaps.

The identification of full-length LTR retrotransposons with LTRharvest<sup>82</sup> resulted in 143,957 non-overlapping candidate sequences using the following parameter settings: `overlaps best -seed 30 -minlenlr 100 -maxlenlr 2000 -mindistlitr 3000 -maxdistlitr 25000 -similar 85 -minstd 4 -maxstd 20 -motif tgca -motifmis 1 -vic 60 -xdrop 5 -mat 2 -mis -2 -ins -3 -del -3`. All candidates were annotated for PfamA domains with hmmer3 software<sup>83</sup> and stringently filtered for false positives by several criteria, the main ones being the presence of at least one typical retrotransposon domain (for example, RT, RH, INT, GAG) and a tandem repeat content below 25%. This resulted in a final set of 24,952 high-confidence full-length LTR retrotransposons. Insertion ages of the LTR retrotransposons were calculated according to the method of ref. 84 by the divergence of 5' and 3' LTRs that had been identical at the time of transposition. We used a grass-specific mutation rate of  $1 \times 10^{-8}$ . The average age of all full-length LTR elements was

calculated in 4 Mb windows and plotted in Fig. 1a. The frequencies of 20-mers were determined using Tallymer<sup>85</sup>.

Phylogenetic analysis of Gypsy elements was performed on predicted protein sequences deposited at the TREP database<sup>32</sup>. Protein domains in predicted open reading frames were identified with Pfam<sup>86</sup>, SignalP<sup>87</sup> and COILS<sup>88</sup>.

For the analysis of transposable element content in up- and downstream regions of genes, 10 kb immediately flanking the predicted coding sequences of all high-confidence genes were extracted from the genome assembly. The genomic segments were then used in BLASTN searches<sup>79</sup> against the TREP database<sup>32</sup>. After an initial annotation, previously unclassified or poorly characterized transposable element families were re-analysed and new consensus sequences were constructed. Analysis of up- and downstream regions was then repeated with the updated TREP database. The transposable element family producing the longest BLASTN hit was determined for every 20th base position of each 10 kb segment, resulting in 500 data points for each up- and downstream region of the high-confidence genes.

**Gene family analysis.** Gene family clusters were defined from 39,734 barley high-confidence class genes and the annotated gene sets of Rice MSU7.0 (39,049 genes, <http://rice.plantbiology.msu.edu/>), *B. distachyon* version 3.1 (31,694 genes, [https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org\\_Bdistachyon](https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Bdistachyon)), *S. bicolor* version 3.1 (33,032 genes, [https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org\\_Sbicolor](https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Sbicolor)) and *A. thaliana* TAIR10 (27,416 genes, <http://www.arabidopsis.org/>) using OrthoMCL<sup>89</sup> software version 2.0. Splice variants were removed from the datasets, keeping only the representative/longest protein sequence prediction, and datasets were filtered for internal stop codons and incompatible reading frames. In the first step, pairwise sequence similarities between all input protein sequences were calculated using BLASTP<sup>79</sup> with an e-value cut-off of  $10^{-5}$ . Markov clustering of the resulting similarity matrix was used to define the orthologue cluster structure, using an inflation value (-I) of 1.5 (OrthoMCL default). Gene families with barley-specific gene duplications, compared with other plant species, were extracted from the ENSEMBL Compara pipeline<sup>90</sup>. Over- and under-representation of Gene Ontology terms between barley and other plant species (Supplementary Tables 4.1–4.3) and between genomic compartments (Supplementary Table 4.5) were analysed with a hypergeometric test using the functions GOstats and GSEABase from the Bioconductor R package<sup>91</sup> against a universe of all genes with Gene Ontology annotations. REViGO<sup>92</sup>, which removes redundant and similar terms from long Gene Ontology lists by semantic clustering, was applied to visualize the enrichment results. Expansion of three barley gene families encoding  $\alpha$ -amylases, the vacuolar processing enzyme VPE2 protein subfamily and the sugar transporters SWEET11 subfamily, with specific importance in barley grain filling/seed development or barley germination/malting, were analysed in greater detail using BLAST searches (versus genome and gene prediction) as well as GenomeThreader mappings to the barley genome assembly. Further details are provided in Supplementary Note 4. *In situ* hybridizations for SWEET genes were performed as described previously<sup>93</sup>.

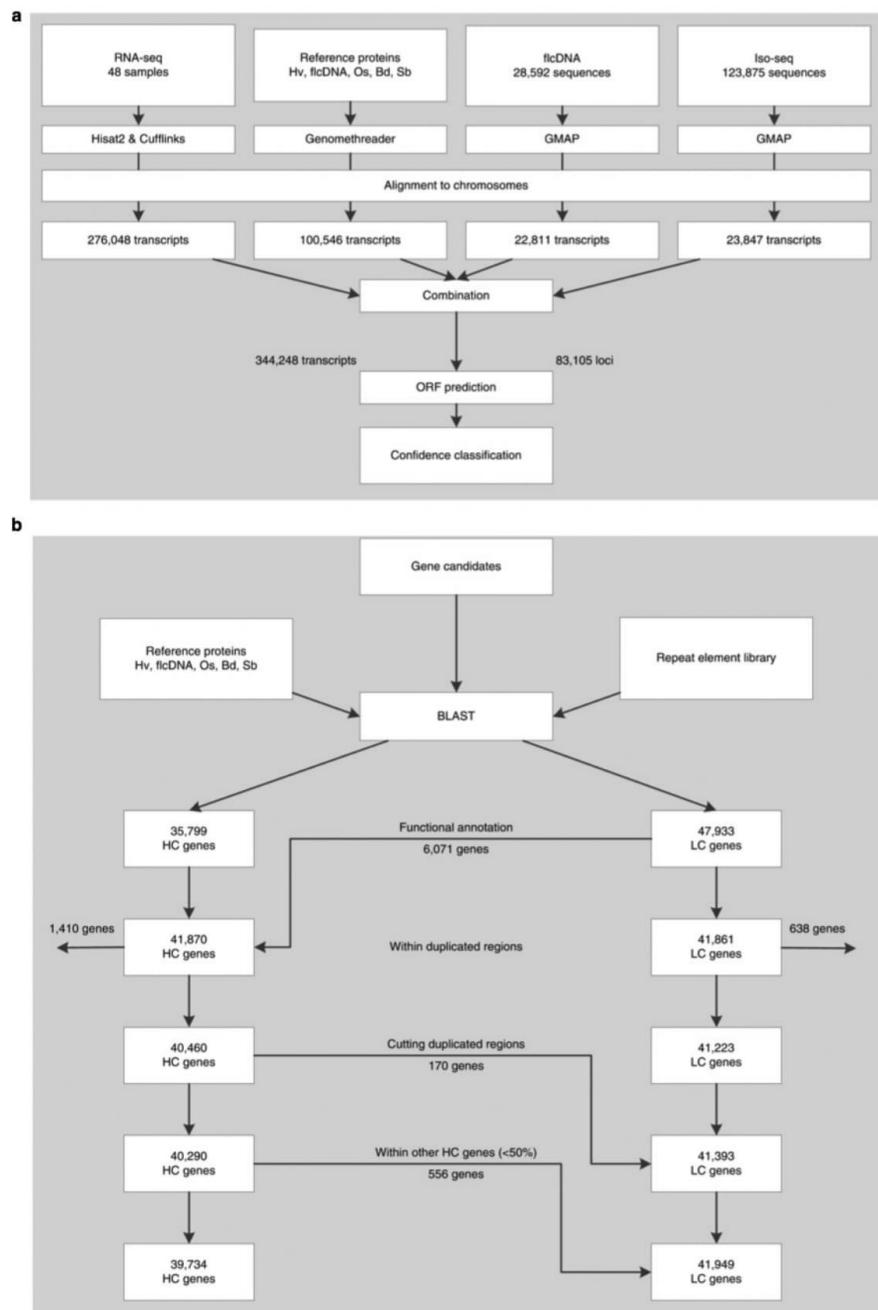
**Analysis of sequence and haplotype diversity.** Ninety-six two-row spring ( $n = 48$ ) and winter ( $n = 48$ ) homozygous inbred elite barley lines (Supplementary Table 5.1) were subjected to exome capture using the barley Roche NimbleGen exome capture liquid array<sup>94</sup> and sequenced on the Illumina HiSeq 2500 platform. An average of  $2 \times 21,876,780$  paired-end Illumina reads per sample was generated. This corresponds to approximately 72% coverage of the 61 Mb exome capture space.

The raw Illumina reads were mapped to the reference sequence with BWA-MEM version 0.7.10 (ref. 66), using a stringent mismatch setting of  $\leq 2\%$  mismatches per read. Variant calling was performed with the Genome Analysis Tool Kit (GATK)<sup>71</sup> version 3.4.0, following the GATK Best Practices pipeline (<https://www.broadinstitute.org/gatk/guide/best-practices.php>). This included read de-duplication, indel realignment, base quality score recalibration and variant calling with the latest version of the HaplotypeCaller. The workflow was implemented in a BASH script. The Tablet assembly viewer<sup>95</sup> was used for visual spot checks of mappings and SNPs calls.

Variant discovery resulted in 15,982,580 variants in total, of which 943,959 were multi-nucleotide polymorphisms or short insertions/deletions (indels), while the remainder represented SNPs. For subsequent genetic analysis, we first reduced the total variant dataset by applying rigorous filtering criteria to produce a highly robust subset of 72,563 SNPs distributed across all seven barley chromosomes. The filtering applied was as follows: (1)  $\geq 8 \times$  coverage for  $\geq 50\%$  of the samples; (2)  $\geq 95\%$  of samples represented at each SNP locus; (3)  $\geq 5\%$  minor allele frequency at the level of the sample; that is, counting sample genotypes rather than individual reads; (4) a VCF SNP quality score  $\geq 30$ ; and (5)  $\geq 98\%$  of samples homozygous. These filters reduced false-positive variant calls by removing spurious variant calls resulting from systematic read mis-mapping. Of this filtered dataset, a subset of 3,500 randomly sampled markers from each chromosome was analysed with the Haplovie software<sup>96</sup>. This subsampling was required as Haplovie was

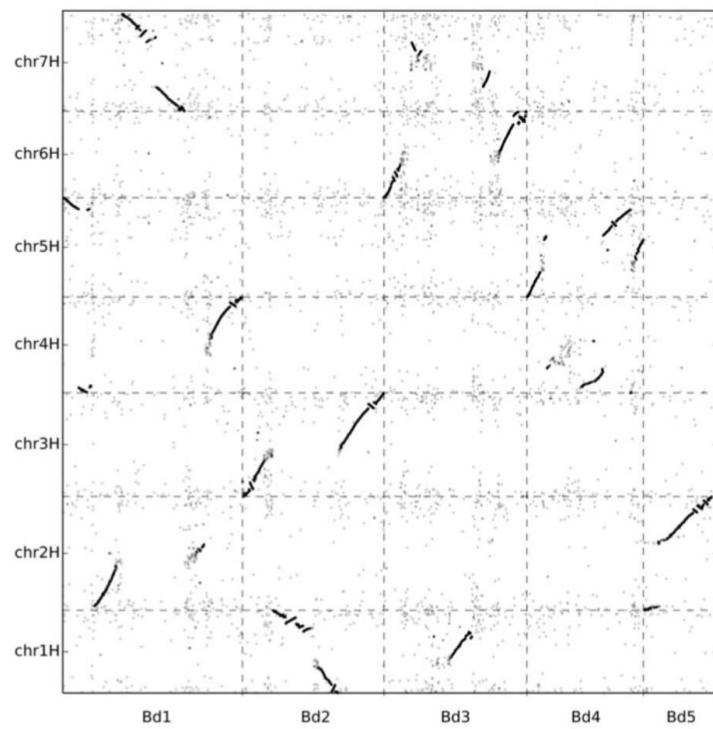
unable to generate the required plots when larger data volumes were used as input. Haploview was run on defaults, using the GABRIEL blocks method. The genotype calls were also imported into the genotype visualization software Flapjack<sup>97</sup> to produce chromosome-scale images of haplotype diversity within the spring and winter pools. Diversity statistics were calculated in GenAIEx version 6.502 (ref. 98) and rolling averages based on 100 adjacent SNPs were plotted in Microsoft Excel 2010. **Data availability.** The genome assembly for barley has been deposited in the Plant Genomics and Phenomics Research Data Repository under digital object identifier <http://dx.doi.org/10.5447/IPK/2016/34>. Accession numbers for all deposited datasets are listed in Supplementary Note 1. The barley genome assembly has been deposited on the IPK Barley Blast Server ([http://webblast.ipk-gatersleben.de/barley\\_ibsc/](http://webblast.ipk-gatersleben.de/barley_ibsc/)). All other data are available from the corresponding authors upon reasonable request.

- Chevreux, B., Wetter, T. & Suhai, S. Genome sequence assembly using trace signals and additional sequence information. In *Computer Science and Biology: Proc. 99th German Conference on Bioinformatics* (eds Hofstaedt, R. et al. 45–56 (GCB, 1999).
- Steuernagel, B. et al. *De novo* 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley. *BMC Genomics* **10**, 547 (2009).
- Taudien, S. et al. Sequencing of BAC pools by different next generation sequencing platforms and strategies. *BMC Res. Notes* **4**, 411 (2011).
- Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1**, 18 (2012).
- Simpson, J. T. et al. ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
- Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
- Belton, J. M. et al. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
- Štarková, H. et al. BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. *Plant Biotechnol. J.* **14**, 1523–1531 (2016).
- Lysák, M. A. et al. Flow karyotyping and sorting of mitotic chromosomes of barley (*Hordeum vulgare L.*). *Chromosome Res.* **7**, 431–444 (1999).
- Šimková, H., Čihálková, J., Vrána, J., Lysák, M. & Doležel, J. Preparation of HMW DNA from plant nuclei and chromosomes isolated from root tips. *Biol. Plant.* **46**, 369–373 (2003).
- Cao, H. et al. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *Gigascience* **3**, 34 (2014).
- Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**, 203–214 (2000).
- Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst.* 1695 (2006).
- Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
- Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://arxiv.org/abs/1303.3997> (2013).
- Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
- Hu, M. et al. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* **28**, 3131–3133 (2012).
- R Core Team. R: a language and environment for statistical computing (R Foundation for Statistical Computing, 2015).
- Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
- DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
- Aliyeva-Schnorr, L. et al. Cytogenetic mapping with centromeric bacterial artificial chromosomes contigs shows that this recombination-poor region comprises more than half of barley chromosome 3H. *Plant J.* **84**, 385–394 (2015).
- Hudakova, S. et al. Sequence organization of barley centromeres. *Nucleic Acids Res.* **29**, 5029–5035 (2001).
- International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
- International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763–768 (2010).
- Paterson, A. H. et al. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
- Matsumoto, T. et al. Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. *Plant Physiol.* **156**, 20–28 (2011).
- Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
- Altshul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
- Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
- Spannagi, M. et al. PGSB PlantsDB: updates to the database framework for comparative plant genome research. *Nucleic Acids Res.* **44** (D1), D1141–D1147 (2016).
- Ellinghaus, D., Kurtz, S. & Willhöft, U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
- Eddy, S. R. Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
- SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y. & Bennetzen, J. L. The paleontology of intergenic retrotransposons of maize. *Nat. Genet.* **20**, 43–45 (1998).
- Gremme, G., Steinbiss, S. & Kurtz, S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* **10**, 645–656 (2013).
- Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–D141 (2004).
- Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011).
- Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164 (1991).
- Li, L., Stoeckert, C. J. Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
- Bolser, D., Staines, D. M., Pritchard, E. & Kersey, P. Ensembl Plants: integrating tools for visualizing, mining, and analyzing plant genomics data. *Methods Mol. Biol.* **1374**, 115–140 (2016).
- Gentleman, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
- Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of Gene Ontology terms. *PLoS ONE* **6**, e21800 (2011).
- Radchuk, V., Weier, D., Radchuk, R., Weschke, W. & Weber, H. Development of maternal seed tissue in barley is mediated by regulated cell expansion and cell disintegration and coordinated with endosperm growth. *J. Exp. Bot.* **62**, 1217–1227 (2011).
- Mascher, M. et al. Barley whole exome capture: a tool for genomic research in the genus *Hordeum* and beyond. *Plant J.* **76**, 494–505 (2013).
- Milne, I. et al. Tablet—next generation sequence assembly visualization. *Bioinformatics* **26**, 401–402 (2010).
- Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
- Milne, I. et al. Flapjack—graphical genotype visualization. *Bioinformatics* **26**, 3133–3134 (2010).
- Peakall, R. & Smouse, P. E. GenAIEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics* **28**, 2537–2539 (2012).
- The International Wheat Genome Sequencing Consortium. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* **345**, 1251788 (2014).

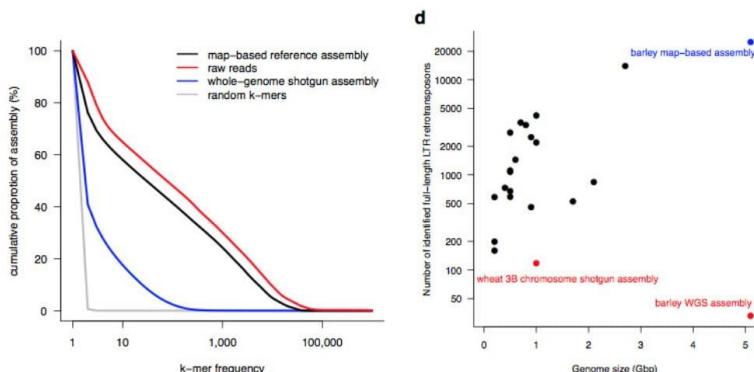


**Extended Data Figure 1 | Gene annotation pipeline.** **a**, Gene annotation pipeline combined gene evidence information from four data sources. Open reading frames were then predicted for 83,105 gene candidates. **b**, Gene candidates were classified into high-confidence (HC) and low-confidence (LC) genes on the basis of homology to reference proteins and

alignment to library of repeat elements. Additional filtering procedures were applied before defining the final gene sets. Arrows between boxes with counts of high-confidence and low-confidence genes in each step indicate re-classifications (high-confidence to low-confidence, or low-confidence to high-confidence).

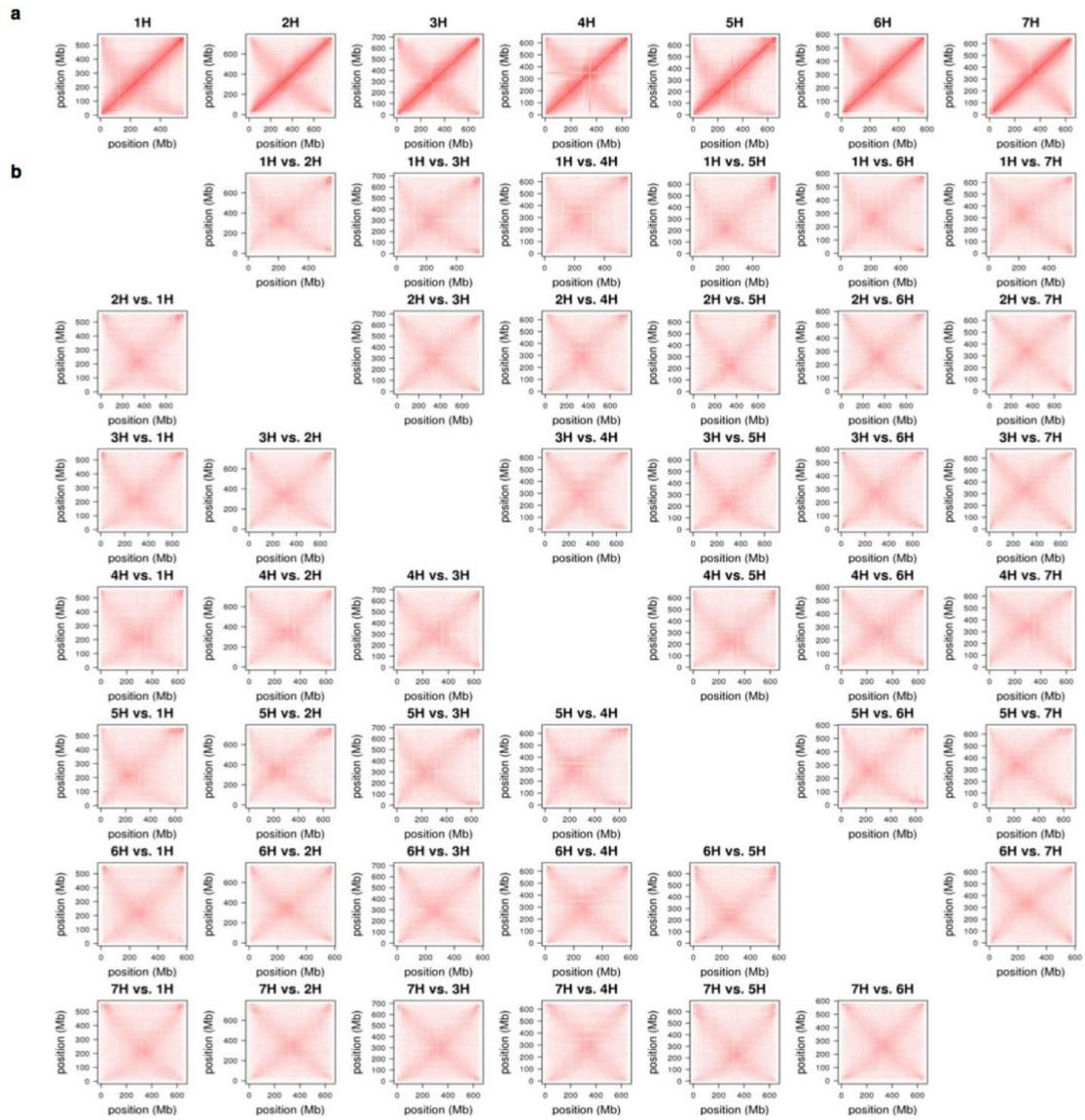


Gene set	Complete genes	Fragmented genes	Missing genes
All gene models	97.6%	1.3%	1.2%
High-confidence genes only	94.7%	1.9%	3.5%
Low-confidence genes only	10.5%	4.5%	85.0%
Gene models annotated on WGS assembly	95.3%	2.9%	1.8%

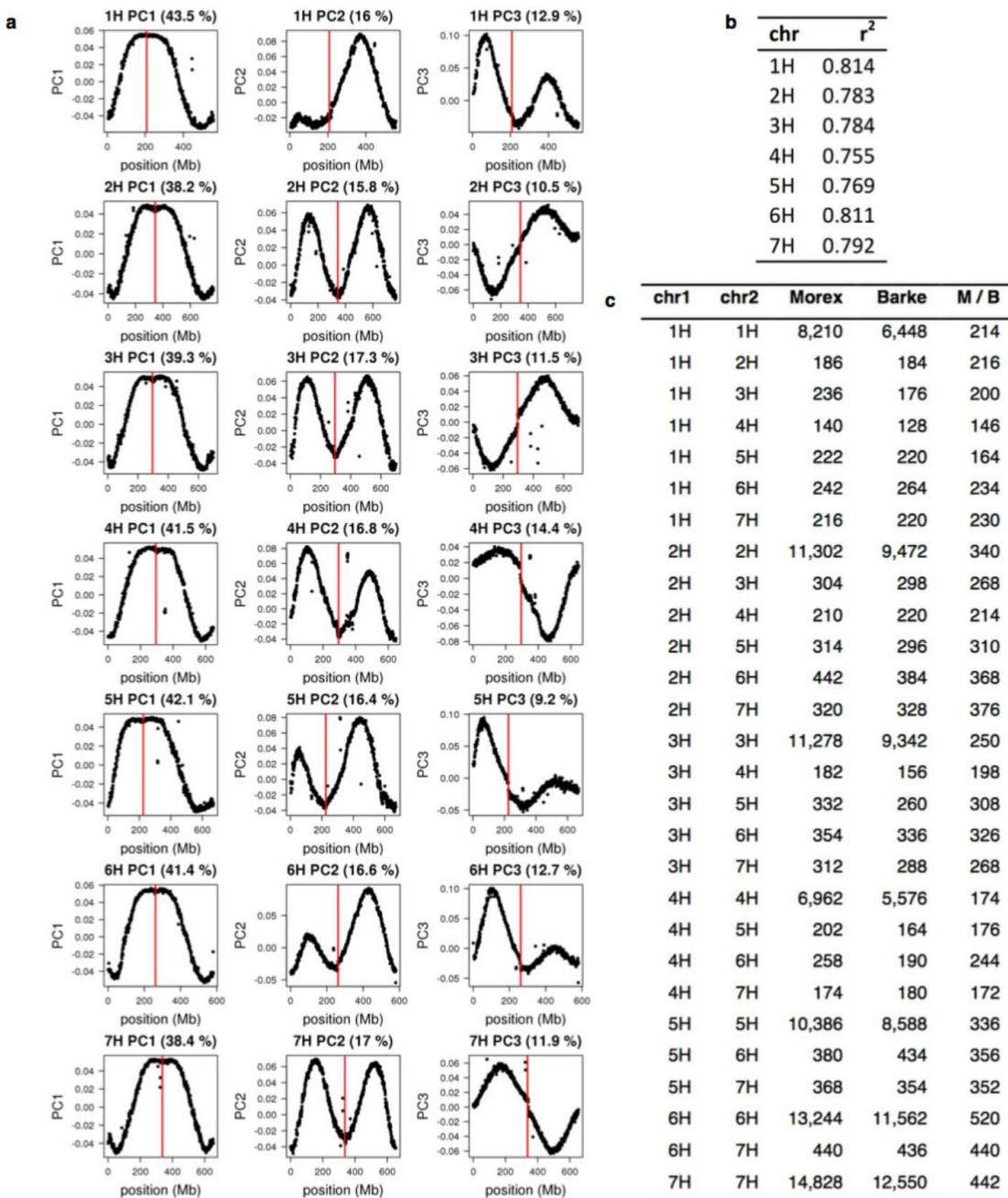


**Extended Data Figure 2 | Assembly validation.** **a**, Conserved gene order between barley (y axis) and *B. distachyon* (x axis). **b**, Completeness of the gene annotation as assessed by BUSCO. **c**, Representation of repetitive k-mers in reads and assemblies. **d**, Representation of full-length LTR

retrotransposons in sequence assemblies of plant genomes with different sizes (represented by black points). The map-based reference sequence of barley reported in the present paper is shown in blue. Red dots correspond to shotgun assemblies of the barley genome<sup>7</sup> and wheat chromosome 3B<sup>99</sup>.

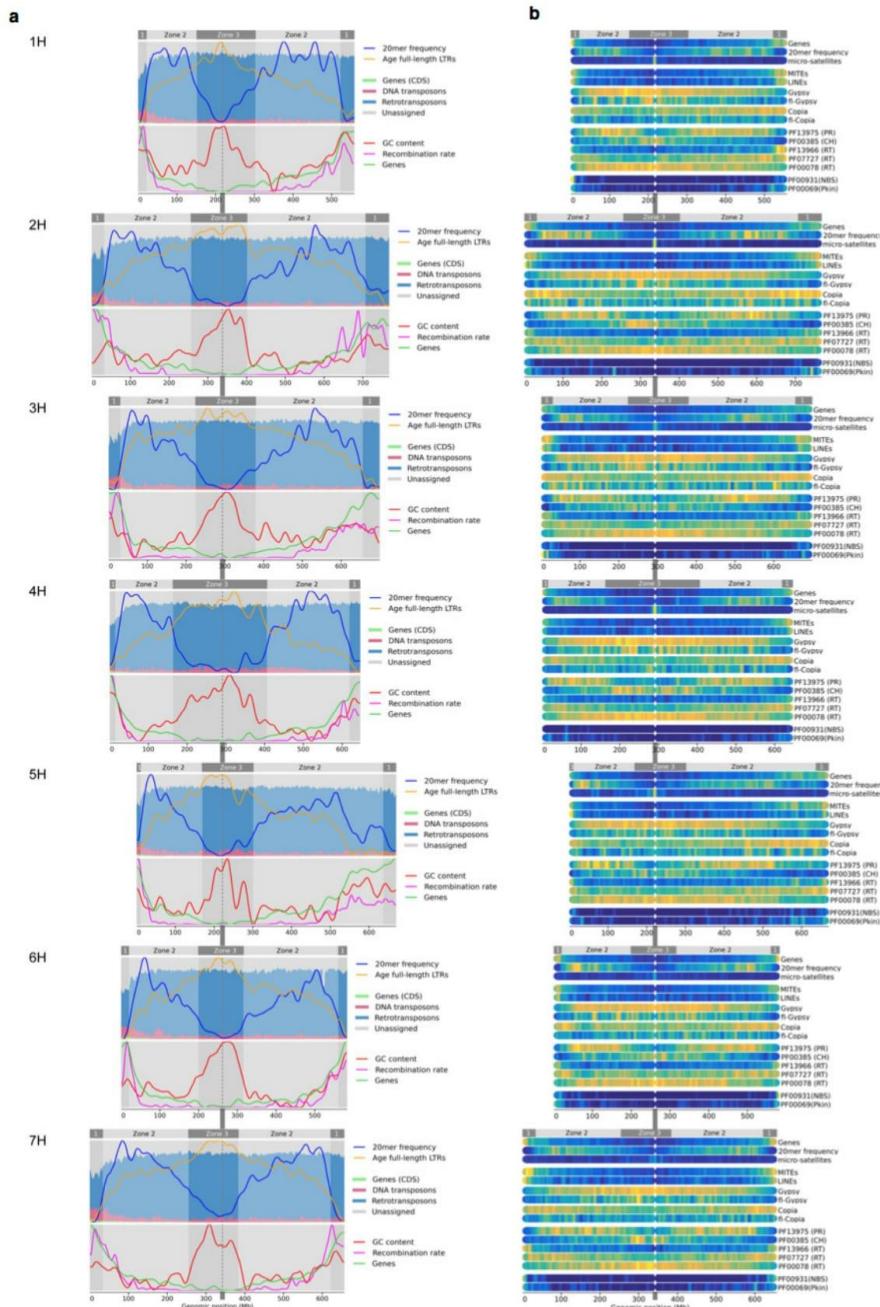


Extended Data Figure 3 | Hi-C contact matrices. **a**, Intrachromosomal contacts. **b**, Interchromosomal contacts. Darker red indicates a higher contact probability.

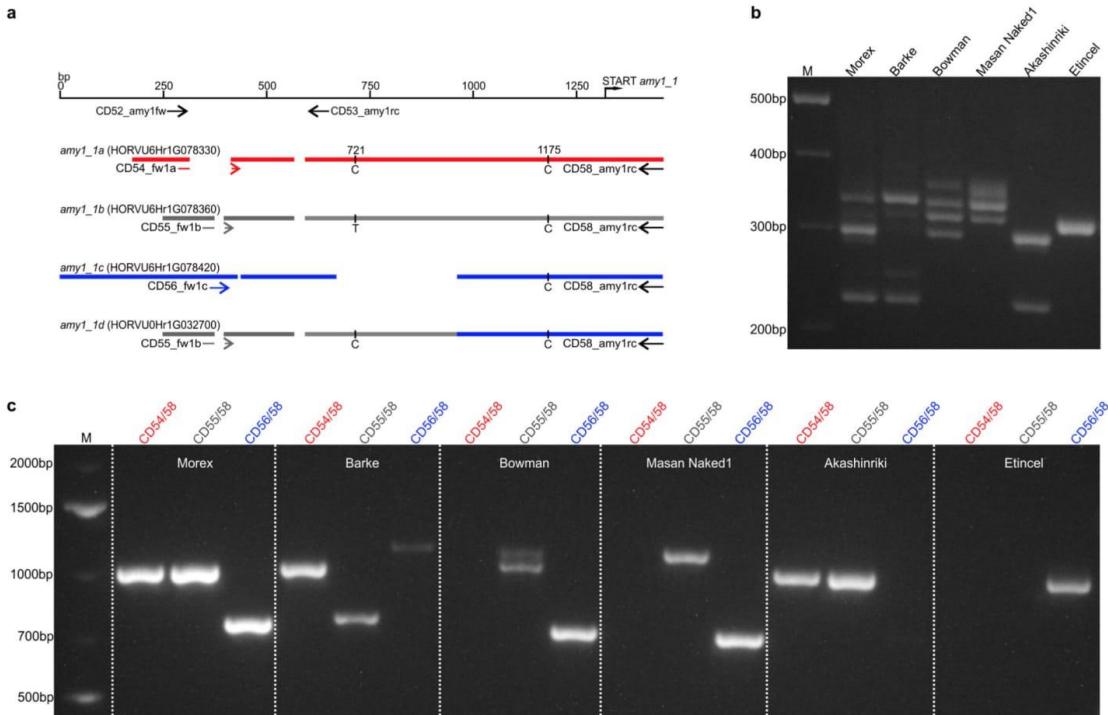


**Extended Data Figure 4 | Global patterns in Hi-C contact matrices.**  
**a**, Principal component analysis of intrachromosomal Hi-C contact matrices. The eigenvectors of the first three principal components are plotted. Centromere positions are marked with a red line. **b**, Proportion of variance explained by linear models incorporating position informational

in the linear genome fitted to the Hi-C contact matrices. **c**, Hi-C link counts in Morex  $\times$  Barke F<sub>1</sub> hybrids within the same chromosome, between homologous chromosomes and between non-homologous chromosomes.

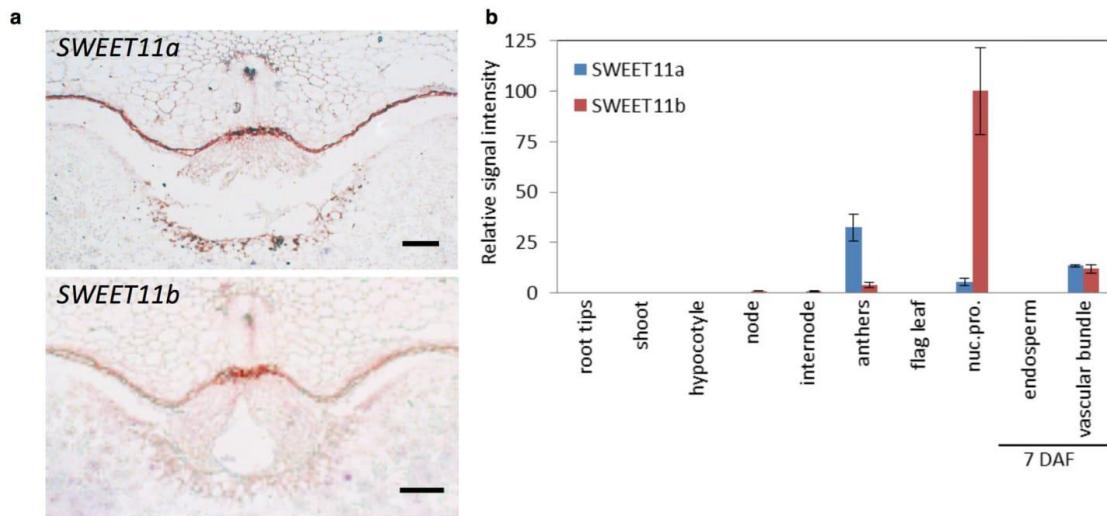


**Extended Data Figure 5 | Distributions of genomic features and the context of repetitive elements.** **a, b,** Panels **a** and **b** are analogous to Figs 1a and 2a. Grey vertical connector bars and dashed lines inside sub-panels for each chromosome indicate centromere positions.

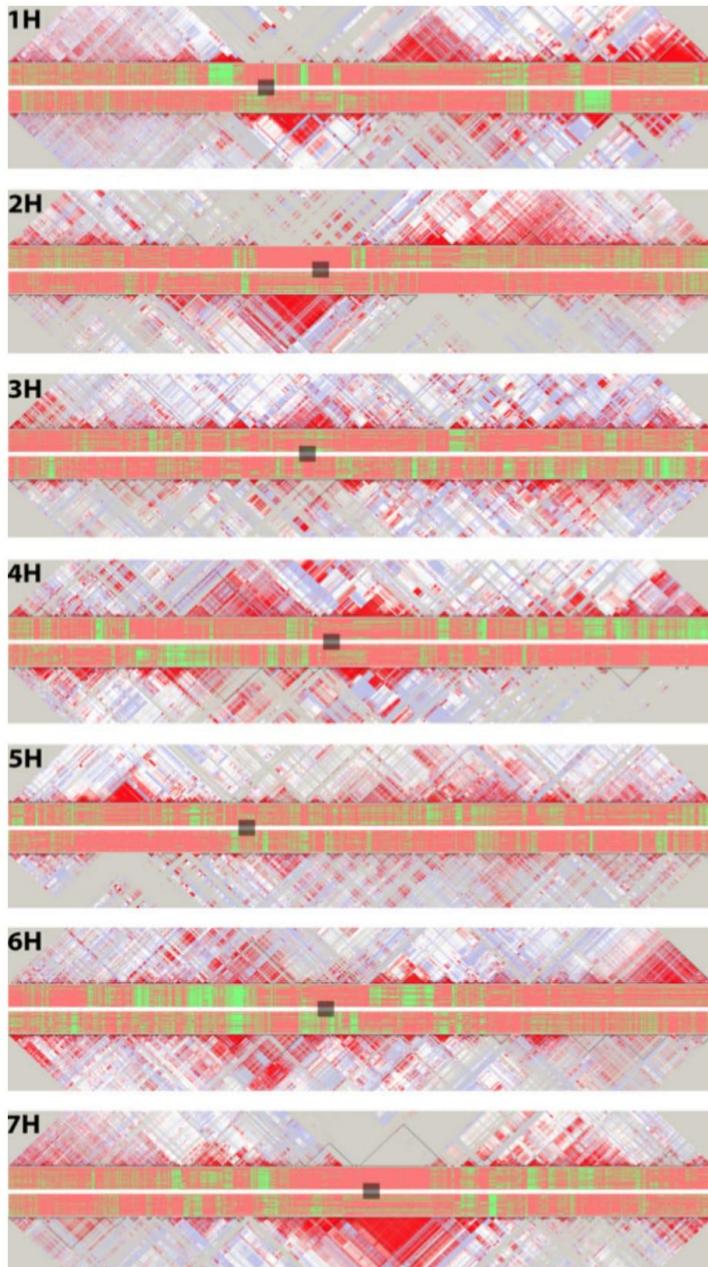


**Extended Data Figure 6 | Experimental strategy to distinguish individual *amy1\_1* copies by PCR from genomic DNA through polymorphisms in the extended promoter regions of *amy1\_1* full-length copies.** **a**, Experimental strategy, primers CD52\_amy1fw and CD53\_amy1rc bind in the extended promoter region of all full-length *amy1\_1* copies (expected amplicon sizes are 225 bp for *amy1\_1a*, 299 bp for *amy1\_1b* and *amy1\_1d* and 336 bp for *amy1\_1c*). Forward primers CD54\_fw1a, CD55\_fw1b and CD56\_fw1c are designed to specifically amplify copies *amy1\_1a*, *amy1\_1b* and *amy1\_1c*, respectively when used with reverse primer CD58\_amy1rc, which binds in the coding region of all *amy1\_1* copies. Expected amplicon sizes are 1,024 bp (*amy1\_1a*), 1,026 bp (*amy1\_1b*) and 757 bp (*amy1\_1c*). Primer pair (CD55\_fw1b–CD58\_amy1rc) further binds to copy *amy1\_1d*; here, sequences of the expected amplicons contain sufficient polymorphisms to distinguish these copies from each other. Positions of selected sequence polymorphisms and deleted regions suitable to distinguish single copies are indicated as black vertical bars and gaps, respectively. Numbering was done in respect of copy *amy1\_1b*. **b**, PCR amplification of *amy1\_1* promoter regions in six barley cultivars and landraces. As expected, a PCR for cultivar Morex, using universal primers CD52\_amy1fw and CD53\_amy1rc, resulted in three amplicons of the expected sizes 225, 299 and 336 bp (compare **a**), which was confirmed by Sanger sequencing. Further primers CD52\_amy1fw and CD53\_amy1rc were used to amplify the *amy1\_1* extended promoter region in various barley cultivars. These experiments

indicate polymorphic variation in, or even absence of, single promoters of *amy1\_1* in the different cultivars. The cultivars analysed differ in row type (six-rowed: cultivars Morex, Masan Naked 1, Akashinriki, Etincel; two-rowed: cultivars Barke, Bowman), growth habit (spring barley: cultivars Morex, Barke, Bowman, Masan Naked 1, Akashinriki; winter barley: cultivar Etincel) and geographic origin (North America: cultivars Morex, Bowman; Europe: cultivars Barke, Etincel; Asia: cultivars Masan Naked 1, Akashinriki). The cultivars Masan Naked 1 and Akashinriki depict landraces used for food, Bowman was classified as non-malting barley, while Morex, Barke and Etincel represent modern malting barley. **c**, Copy-specific PCR amplification of *amy1\_1* extended promoter regions. PCR amplification and Sanger sequencing identified three *amy1\_1* copies in barley cultivar Morex: *amy1\_1a* (CD54\_fw1a–CD58\_amy1rc), *amy1\_1b* (CD55\_fw1b–CD58\_amy1rc) and *amy1\_1c* (CD56\_fw1c–CD58\_amy1rc). Additionally, sequencing revealed two polymorphic sites in PCR amplicon *amy1\_1b* (CD55\_fw1b–CD58\_amy1rc) at positions 721 bp (T/C) and 1175 bp (C/T) (see **a**), indicating the presence of one or two additional *amy1\_1b*-like copies in the genome of the analysed individual. The presence of copy *amy1\_1d* could not be confirmed. The reason for that might have been sequence deviations in the cultivar Morex accession used for BAC library construction versus that used for the presented experiments, or differences in PCR efficiency for amplification of copies *amy1\_1b* and *amy1\_1d*.



**Extended Data Figure 7 | SWEET gene expression.** **a**, Control experiment for mRNA *in situ* hybridizations shown in Fig. 3c. *In situ* hybridization with sense probes for *SWEET11a* (top) and *SWEET11b* (bottom). Scale bars, 100  $\mu$ m. **b**, Expression of *SWEET11a* and *SWEET11b*. Results of qPCR in different plant organs and in the developing grains at 7 days after flowering (DAF).



**Extended Data Figure 8 | Haplotype blocks in sets of 48 samples each of elite two-row spring barley lines (top half of each chromosome's figure) and winter barley lines (bottom half), separately for each chromosome.** We restricted the number of SNPs per chromosome by randomly choosing 3,500 to fit with the maximum permitted by the software. The red and green plots in the centre of each chromosome figure represent whole-canvases dumps produced with the Flapjack software<sup>97</sup>. Markers are arranged in columns in linear order along the chromosome; red pixels represent reference alleles, while green pixels represent alternative alleles. Each row represents a barley cultivar; these have been sorted top to

bottom by year of introduction (ascending). The Flapjack plots are framed by cropped linkage disequilibrium plots generated with the HaploView software<sup>96</sup>. Colour intensity conveys the extent of linkage between pairs of markers (red, highest). Approximate centromere positions are indicated by semi-opaque grey squares. The triangles with the thin black outline represent haplotype blocks as computed by HaploView. In some regions, extensive stretches exist where no blocks were detected (for example, chr2H, spring lines in top half, near centromere). These generally present highly monomorphic regions where there is no evidence for multiple haplotypes, and consequently blocks were not called.

Extended Data Table 1 | Hi-C and optical map datasets for chromosome-scale assembly

**a Summary of Hi-C libraries**

Library	Number of all reads	Number of mapped reads	Links between restriction fragments
<b>HiC1</b>	229,672,122	63,133,030	7,449,949
<b>HiC2</b>	334,742,791	79,745,191	7,663,777
<b>HiC4</b>	183,044,989	53,818,372	4,983,859
<b>HiC5</b>	178,785,306	58,212,813	2,439,898
<b>HiC6</b>	219,294,615	63,853,743	5,594,744
<b>TCC2</b>	260,968,878	55,242,411	7,431,165
<b>TCC4</b>	182,033,300	35,964,622	6,336,274
<b>TCC5</b>	204,856,338	42,544,941	7,913,758
<b>TCC7</b>	236,976,831	65,188,433	7,197,767
<b>TCC8</b>	226,042,216	71,397,037	4,380,187
<b>TCC9</b>	237,059,303	49,879,999	8,877,701
<b>TOTAL</b>	2,493,476,689	638,980,592	70,269,079

**b Raw data and assembly statistics of the optical map.**

<b>Number of molecules &gt; 150 kb</b>	774,557
<b>Molecule N50</b>	340 kb
<b>Number of contigs</b>	2,875
<b>Assembly length</b>	4,289 Mb
<b>Average contig coverage</b>	57-fold
<b>Fraction of molecules aligned to assembly</b>	85 %

Extended Data Table 2 | Statistics on gene annotation and genomic compartments

**a** Gene annotation statistics for high-confidence (HC) and low-confidence (LC) genes.

	1H	2H	3H	4H	5H	6H	7H	Un	TOTAL
No. of HC genes	4,634	6,518	5,760	4,380	6,165	4,544	5,576	2,157	39,734
No. of LC genes	4,911	6,259	6,035	4,720	6,420	4,994	6,712	1,898	41,949
No. of HC transcripts	30,711	40,432	38,322	29,388	37,877	28,293	35,709	7,538	248,270
No. of LC transcript	10,754	13,287	12,589	10,331	12,471	10,354	12,795	3,275	85,856
Mean length of HC genes	5,450	7,533	5,835	5,472	6,013	6,091	6,319	3,195	6,010
Mean length of LC genes	2,460	2,561	2,145	2,253	2,381	2,322	2,286	1,982	2,328
Median no. of transcript per HC gene	3	3	3	3	3	3	3	2	3
Median no. of transcript per LC gene	1	1	1	1	1	1	1	1	1
Mean length of HC transcripts	1,990	1,876	1,992	1,983	1,926	1,961	1,888	1,475	1,927
Mean length of LC transcripts	1,595	1,484	1,532	1,487	1,534	1,453	1,360	1,156	1,478
Median no. of exon per HC transcript	6	5	6	6	5	5	5	4	5
Median no. of exon per LC transcript	2	2	2	2	2	2	2	1	2
Mean length of HC proteins	380	351	364	366	357	361	362	298	360
Mean length of LC proteins	191	173	184	166	179	164	165	164	174

**b** Genomic compartments across all chromosomes

	ZONE 1 distal	ZONE 2 interstitial	ZONE 3 proximal
Size	433 Mb (9 %)	3,075 Mb (63.6 %)	1,076 (Mb) (22.3 %)
Number of genes	9,725 (24.5 %)	24,516 (61.7 %)	3,336 (8.4 %)
Gene density per Mb	22.5	8.0	3.1
Transposon content	64.2 %	82.1 %	83.7 %
LTR/DNA-TE ratio	6.1	18.7	16.8
Gypsy/Copia ratio	0.6	1.3	1.8

Extended Data Table 3 | Repeat annotation statistics

	% of genome	% of TE bp	number	number %	size (Mb)	average length (bp)
<b>Mobile Element (TXX)</b>	<b>80.8</b>	<b>100.0</b>	<b>3,408,238</b>	<b>100</b>	<b>3,695</b>	<b>1,084</b>
<b>Class I: Retroelement (RXX)</b>	<b>75.2</b>	<b>93.1</b>	<b>2,881,139</b>	<b>84.5</b>	<b>3,439</b>	<b>1,194</b>
LTR Retrotransposon (RLX)	75.0	92.7	2,859,922	83.9	3,427	1,198
<i>Copia</i> (RLC)	16.0	19.8	588,579	17.3	732	1,243
<i>Gypsy</i> (RLG)	21.3	26.3	765,584	22.5	972	1,270
unclassified LTR (RLX)	37.7	46.6	1,505,759	44.2	1,723	1,144
non-LTR Retrotransposon (RXX)	0.3	0.3	21,217	0.6	12	581
LINE (RIX)	0.3	0.3	19,173	0.6	12	605
SINE (RSX)	0.0	0.0	2,044	0.1	1	355
<b>Class II: DNA Transposon (DXX)</b>	<b>5.3</b>	<b>6.5</b>	<b>473,797</b>	<b>13.9</b>	<b>241</b>	<b>509</b>
DNA Transposon Superfamily	5.0	6.2	418,583	12.3	230	550
CACTA superfamily (DTC)	4.7	5.9	375,421	11.0	217	578
hAT superfamily (DTA)	0.01	0.01	607	0.0	0	402
Mutator superfamily (DTM)	0.15	0.19	18,936	0.6	7	370
Tc1/Mariner superfamily (DTT)	0.02	0.03	8,199	0.2	1	134
PIF/Harbinger (DTH)	0.08	0.10	9,007	0.3	4	402
unclassified (DTX)	0.03	0.03	6,413	0.2	1	191
MITEs (DXX)	0.20	0.25	52,112	1.5	9	178
Helitron (DHH)	0.03	0.04	1,643	0.0	1	818
unclassified DNA transposon	0.01	0.01	1,459	0.0	1	350
<b>Unclassified Element (TXX)</b>	<b>0.32</b>	<b>0.40</b>	<b>53,302</b>	<b>1.6</b>	<b>15</b>	<b>274</b>
<i>Retro-TE/DNA-TE ratio</i>				6.1		
<i>Gypsy/Copia ratio</i>				1.3		

Extended Data Table 4 | Information on gene families associated with malting quality

## a α-amylases

Gene name	ID	Chr	Strand	Coordinates on pseudomolecule (start to stop codon)	BAC sequence contig	Historical nomenclature	Copy-specific PCR primer for promoter region <i>amy1_1</i>
<i>amy4_1</i>	HORVU2Hr1G071710 <sup>*</sup>	2H	plus	511,664,000 – 511,667,683	mA0231C11_C8	N/A	N/A
<i>amy4_2</i>	HORVU3Hr1G067620 <sup>*</sup>	3H	minus	513,498,473 – 513,485,531	eA0011L11_C1	N/A	N/A
<i>amy3</i>	HORVU5Hr1G068350 <sup>*</sup>	5H	plus	517,452,674 – 517,454,307	rA0171B14_C3	N/A	N/A
<i>amy1_1a</i>	HORVU6Hr1G078330 <sup>*</sup>	6H	minus	533,880,485 – 533,879,015	hA0060C06_C2	<i>amy6_4</i> <sup>**</sup>	CD54_fw1a
<i>amy1_1b</i>	HORVU6Hr1G078360 <sup>*</sup>	6H	plus	534,112,867 – 534,114,337	eA0332P17_C1	<i>amy6_4</i> <sup>**</sup>	CD55_fw1b
N/A <sup>†</sup>	N/A	6H	plus	534,258,381 – 534,259,057	hB0076E06_C1	<i>amy6_4</i> <sup>**</sup>	N/A
<i>amy1_1c</i>	HORVU8Hr1G078420 <sup>*</sup>	6H	minus	534,499,528 – 534,498,058	mA0178F18_C1	<i>amy6_4</i> <sup>**</sup>	CD56_fw1c
<i>amy1_2</i>	HORVU6Hr1G080790 <sup>*</sup>	6H	plus	542,857,506 – 542,858,990	eA0239J18_C1	<i>amy6</i> <sup>**</sup>	N/A
<i>amy2_1</i>	HORVU7Hr1G091150 <sup>*</sup>	7H	minus	556,169,683 – 556,167,920	hA0261M10_C2	<i>amy32b</i> <sup>**</sup>	N/A
<i>amy2_2</i>	HORVU7Hr1G091240 <sup>*</sup>	7H	minus	557,398,785 – 557,397,068	hA0332A16_C1	N/A	N/A
<i>amy2_3</i>	HORVU7Hr1G091250 <sup>*</sup>	7H	minus	557,428,810 – 557,427,021	hA0332A16_C1	N/A	N/A
N/A <sup>‡</sup>	N/A	Un	plus	184,040,968 – 184,042,438	hA0174I01_C3	<i>amy6_4</i> <sup>**</sup>	N/A
<i>amy1_1d</i>	HORVU0Hr1G032700 <sup>*</sup>	Un	plus	195,047,130 – 195,048,600	hB0054J14_C4	<i>amy6_4</i> <sup>**</sup>	CD55_fw1b
<i>amy1_1e</i>	HORVU0Hr1G032850	Un	minus	196,262,594 – 196,261,798	hB0068J02_C14	<i>amy6_4</i> <sup>**</sup>	N/A

<sup>\*</sup>1 considered in phylogenetic tree<sup>\*\*</sup>2 Khurshed, B., and J. Rogers. 1988. Barley alpha-amylase genes. Quantitative comparison of steady-state mRNA levels from individual members of the two different families expressed in aleurone cells. *Journal of Biological Chemistry*. ASBMB 263:18953–18960.<sup>†</sup>3 Rogers, J. C., and C. Milliman. 1984. Coordinate increase in major transcripts from the high pl alpha-amylase multigene family in barley aleurone cells stimulated with gibberellic acid. *Journal of Biological Chemistry*. ASBMB 259:12234–12240.<sup>‡</sup>4 This amy sequence is located in a region of the genome that has been masked and is hence not considered when referring to the total gene count of α-amylases in the reference assembly<sup>5</sup> This amy sequence is a redundant data base entry originating from a short overlap between overlapping BAC sequences and is hence not considered when referring to the total gene count of α-amylases in the reference assembly

## b SWEETs

Gene name	Chromosome	Barley gene ID	Gene identifier of rice ortholog	Transcript coordinates (bp)
<i>SWEET1a</i>	3H	HORVU3Hr1G091230.1	OsSWEET1a (LOC_Os01g56880)	634,920,942–634,924,009
<i>SWEET1b</i>	1H	HORVU1Hr1G065100.2	OsSWEET1b (LOC_Os05g35140)	465,736,768–465,739,685
<i>SWEET2a</i>	6H	HORVU6Hr1G029520.3	OsSWEET2a (LOC_Os01g36070)	120,201,097–120,203,923
<i>SWEET2b</i>	3H	HORVU3Hr1G065770.8	OsSWEET2b (LOC_Os01g50460)	501,045,803–501,048,362
<i>SWEET3</i>	1H	HORVU1Hr1G029920.4	OsSWEET3a (LOC_Os05g12320)	167,987,102–167,989,745
<i>SWEET3b</i>			OsSWEET3b (LOC_Os01g12130)	
<i>SWEET4</i>	6H	HORVU6Hr1G055960.1	OsSWEET4 (LOC_Os02g19820)	356,677,679–356,682,060
<i>SWEET5</i>	1H	HORVU1Hr1G079940.2	OsSWEET5 (LOC_Os05g51090)	524,164,619–524,166,874
<i>SWEET6a</i>	2H	HORVU2Hr1G006510.1	OsSWEET6a (LOC_Os01g42110)	13,613,171–13,614,579
<i>SWEET6b</i>	2H	HORVU2Hr1G006520.1	OsSWEET6b (LOC_Os01g42090)	13,644,166–13,646,353
<i>SWEET7a</i>	7H	HORVU7Hr1G117490.1	OsSWEET7a (LOC_Os09g08030)	645,251,293–645,253,295
<i>SWEET7b</i>	7H	HORVU7Hr1G067000.1	OsSWEET7b (LOC_Os09g08270)	346,595,507–346,597,601
<i>SWEET7c</i>	4H	HORVU4Hr1G070740.1	OsSWEET7c (LOC_Os12g07860)	577,425,380–577,427,479
<i>SWEET11a</i>	5H	HORVU5Hr1G076770.4	OsSWEET11 (LOC_Os08g42350)	551,931,226–551,932,561
<i>SWEET11b</i>	7H	HORVU7Hr1G054710.2		221,745,516–221,747,264
<i>SWEET12</i>	3H	HORVU3Hr1G013170.1	OsSWEET12 (LOC_Os03g22590)	28,461,697–28,464,387
<i>SWEET13a</i>	6H	HORVU6Hr1G089600.1	OsSWEET13 (LOC_Os12g29220)	570,135,624–570,137,778
<i>SWEET13b</i>	6H	HORVU6Hr1G089540.2		570,019,114–570,020,991
<i>SWEET14a</i>	1H	HORVU1Hr1G010210.2	OsSWEET14 (LOC_Os11g31190)	23,166,698–23,169,065
<i>SWEET14b</i>	6H	HORVU6Hr1G000440.3		1,053,692–1,055,923
<i>SWEET15a</i>	7H	HORVU7Hr1G030160.4	OsSWEET15 (LOC_Os02g30910)	58,906,614–58,909,144
<i>SWEET15b</i>	4H	HORVU4Hr1G053450.1		445,034,384–445,035,937
<i>SWEET15c</i>	4H	HORVU4Hr1G053440.1		444,740,701–444,750,029
<i>SWEET16</i>	Unassigned	HORVU0Hr1G010080.2	OsSWEET16 (LOC_Os03g22200)	57,404,637–57,408,253

## c VPEs

Gene name	Chromosome	Barley gene ID	Gene identifier of rice ortholog	Transcript coordinates (bp)
<i>VPE1</i>	6H	HORVU6Hr1G060990.1	OsVPE3 (LOC_Os02g43010)	407,203,000–407,209,087
<i>VPE2a</i>	2H	HORVU2Hr1G091880.1	OsVPE1 (LOC_Os04g45470)	649,971,828–649,977,151
<i>VPE2b</i>	2H	HORVU2Hr1G092090.1		650,899,859–650,900,692
<i>VPE2c</i>	2H	HORVU2Hr1G092080.6		651,050,549–651,054,349
<i>VPE2d</i>	2H	HORVU2Hr1G092080.15		651,056,023–651,060,215
<i>VPE3</i>	3H	HORVU3Hr1G048520.3	OsVPE4 (LOC_Os05g51570)	335,443,989–335,450,401
<i>VPE4</i>	5H	HORVU5Hr1G066250.3	OsVPE5 (LOC_Os06g01610)	505,672,635–505,675,164
<i>VPE5</i>	3H	HORVU3Hr1G115610.8	OsVPE2 (LOC_Os01g37910)	693,484,495–693,492,152

# SCIENTIFIC DATA



OPEN

## Data Descriptor: Construction of a map-based reference genome sequence for barley, *Hordeum vulgare* L.

Received: 26 August 2016

Accepted: 9 February 2017

Published: 27 April 2017

Sebastian Beier *et al.* #

Barley (*Hordeum vulgare* L.) is a cereal grass mainly used as animal fodder and raw material for the malting industry. The map-based reference genome sequence of barley cv. 'Morex' was constructed by the International Barley Genome Sequencing Consortium (IBSC) using hierarchical shotgun sequencing. Here, we report the experimental and computational procedures to (i) sequence and assemble more than 80,000 bacterial artificial chromosome (BAC) clones along the minimum tiling path of a genome-wide physical map, (ii) find and validate overlaps between adjacent BACs, (iii) construct 4,265 non-redundant sequence scaffolds representing clusters of overlapping BACs, and (iv) order and orient these BAC clusters along the seven barley chromosomes using positional information provided by dense genetic maps, an optical map and chromosome conformation capture sequencing (Hi-C). Integrative access to these sequence and mapping resources is provided by the barley genome explorer (BARLEX).

Design Type(s)	genome assembly
Measurement Type(s)	whole genome sequencing assay
Technology Type(s)	DNA sequencing
Factor Type(s)	library preparation
Sample Characteristic(s)	<i>Hordeum vulgare</i>

Correspondence and requests for materials should be addressed to M.M. (email: mascher@ipk-gatersleben.de). #A full list of authors and their affiliations appears at the end of the paper.

## Background & Summary

Barley (*Hordeum vulgare* L.) is a cereal grass of great agronomical importance. The goal of the International Barley Genome Sequencing Consortium (IBSC) is the construction of a map-based reference sequence assembly of barley cultivar 'Morex' by means of hierarchical shotgun sequencing<sup>1</sup>. Towards this aim, the barley genomics community has developed an array of genome-wide physical and genetic mapping resources. These include libraries of bacterial artificial chromosomes (BACs)<sup>2</sup>, a genome-wide physical map<sup>3</sup>, a draft whole genome shotgun (WGS) assembly<sup>4</sup> and an ultra-dense genetic map<sup>5</sup>. The last stage on the road towards the reference genome is the shotgun sequencing of BAC clones along a minimum tiling path of the genome defined by the physical map. The advances in high-throughput sequencing technology enabled this task to be completed in a much shorter timeframe than was required for the completion of, for instance, the human<sup>6</sup> and maize<sup>7</sup> genomes. In addition to the generation of BAC raw sequence data, we constructed (i) physical genome maps by single-molecule optical mapping in nanochannels<sup>8</sup> and by chromosome conformation capture sequencing (Hi-C)<sup>9,10</sup>, and (ii) a high-resolution genetic map of a large bi-parental mapping population through genotyping-by-sequencing<sup>11</sup>. We undertook the sequence assembly of individual BACs, the construction of larger sequence scaffolds by merging sequences from adjacent clones and the integration of these super-scaffolds with the various genome-wide mapping resources constructed in the present effort as well as those published previously<sup>3,5</sup>. The final outcome of this approach was the construction of 'pseudomolecules', i.e., contiguous sequence scaffolds representing the seven chromosomes of barley.

We have submitted the relevant raw data to public sequence data archives, made analysis results available under permanent digital object identifiers (DOIs) and entered the positional information used for pseudomolecule construction into a bespoke information management system, the BARLEX genome explorer<sup>12</sup>. Here, we give (i) a comprehensive overview of datasets used for assembling the barley genome and methods employed in their generation, (ii) a detailed description of wet-lab procedures for BAC sequencing and the bioinformatics workflow of the sequence assembly and data integration procedures together with an outline of (iii) their browsable presentation in an online database. These resources document the construction of the map-based reference sequence of the barley genome and will enable researchers to inspect the evidence used to assemble, order and orient sequence scaffolds and may guide the further improvement of the genome sequence with complementary data sets.

## Methods

The main steps for the construction of the map-based reference sequence of the barley genome were (i) shotgun and mate-pair sequencing of BAC clones, (ii) sequence assembly of individual BAC clones and (iii) the construction of a pseudomolecule sequences by merging the sequences of adjacent BACs into super-scaffolds and ordering these using various sources of positional information such as physical maps, optical map and chromosome conformation capture. A schematic overview of our experimental procedures is given in Fig. 1.

### BAC sequencing

**Identification and analysis of gene-containing BACs.** Isolation of gene-containing BACs, construction of a minimal tiling path (MTP), sequencing of MTP clones and the annotation of genes were essentially as described previously<sup>13</sup>.

**Shotgun and mate-pair sequencing of MTP-BACs.** Sequencing of MTP-BACs was conducted in four laboratories (Leibniz Institute on Aging—Fritz Lipmann Institute (FLI) Jena, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Beijing Genomics Institute (BGI) and Earlham Institute (EI) Norwich). Depending on the instrumentation and established protocols, customized approaches were taken to sequence the barley MTP BACs.

### Barley chromosomes 1H, 3H and 4H (IPK and FLI)

#### Shotgun sequencing of MTP BACs

During the initial phase, BACs mostly from chromosome 3H (4870 clones) and a small number of clones from other chromosomes (34 from 1H; 31 from 2H; 50 from 4H; 101 from 5H; 33 from 6H; 64 from 7H; 107 from '0H') were shotgun sequenced using the Roche/454 GS FLX device (Data Citation 1, Data Citation 2, Data Citation 3, Data Citation 4, Data Citation 5, Data Citation 6, Data Citation 7, Data Citation 8, Data Citation 9). BAC DNA was prepared using a modified alkaline lysis protocol<sup>14</sup>. Construction of barcoded 454 sequencing libraries and sequencing using the Roche platform were performed as described<sup>15,16</sup>. The remaining BAC clones from chromosomes 1H, 3H and 4H were shotgun sequenced employing Illumina instruments. BAC DNA isolation, library construction, sequencing-by-synthesis (paired-end, 2 × 100 cycles) using the Illumina HiSeq2000 device was performed as described<sup>17</sup> (Data Citation 10, Data Citation 11, Data Citation 12, Data Citation 13). Pools of up to 667 BACs were individually barcoded and sequenced on one HiSeq2000 lane.

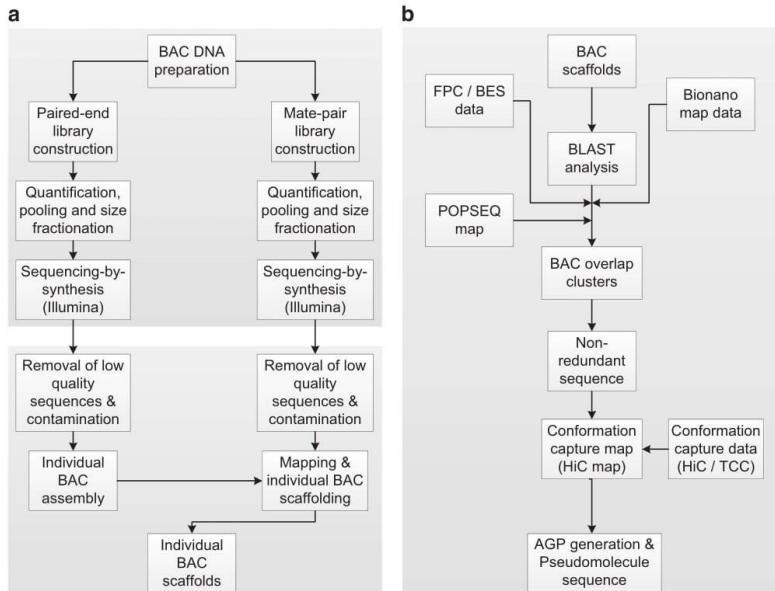
In addition, the Illumina GAIIX, HiSeq2500 and MiSeq machines were utilized to sequence pools of up to 384 clones per lane as described previously<sup>17</sup>.

*Mate-pair sequencing of MTP BACs*

For scaffolding of chromosomes 1H, 3H and 4H standard Illumina Nextera mate-pair libraries (span size: 8 kb) of BAC pools up to 384 BACs were constructed and sequenced using the Illumina HiSeq2000 (paired end, 2 × 100 cycles) and MiSeq (paired end, 2 × 250 cycles) as described<sup>17</sup> (Data Citation 14, Data Citation 15).

Barley chromosomes 5H, 6H and 7H (BGI)*Shotgun sequencing of MTP BACs*

Bacterial starter cultures were inoculated in 0.4 ml 2 × YT liquid medium<sup>18</sup> supplemented with chloramphenicol (17.5 µg ml<sup>-1</sup>) in 2 ml polypropylene 96-deep well-plates sealed with gas-permeable foil and incubated at 37 °C for 14 h in a shaking incubator (210 r.p.m.). For DNA isolation duplicates of cultures (1 ml 2 × YT liquid medium containing 17.5 µg ml<sup>-1</sup> chloramphenicol) were inoculated with 50 µl starter culture and incubated (37 °C, 14 h, 210 r.p.m.). BAC DNA was isolated using the alkaline lysis method essentially as described previously<sup>17</sup>. The DNA was dissolved (overnight, 4 °C) in 64 µl TE (pH 8.0) containing RNase A (30 µg ml<sup>-1</sup>) and stored at -20 °C. BAC plasmid DNA (0.5–2.0 µg in 60 µl) was randomly fragmented by focused-ultrasonicator (Covaris LE220 instrument: 21% duty factor, 500 PIP, 500 cycles per burst, 70 s treatment time) in 96-well plates (Axygen, PCR-96M2-HS-C) to an average size of 250–750 bp. The DNA fragments were purified using magnetic beads (GeneOn Purification kit, GO-PCRC-5000) according to the manufacturer's instructions. DNA was precipitated by adding 10 µl magnetic bead suspension and 75 µl Binding Buffer. The samples were mixed and incubated at room temperature for 5 min. Beads containing the DNA were reclaimed by using a magnet (96S Super Magnet Plate, ALPAQUA, A001322), and the clear supernatant was discarded. The beads were washed twice with 200 µl of 70% ethanol and dried completely. For the elution of DNA the beads were suspended in 42 µl Elution Buffer (EB, 10 mM Tris-Cl, pH 8.5) and incubated (5 min). The plate was placed on the magnet, and the supernatant (40 µl) was transferred into new 96-well plates. End-repair and A-Tailing were performed as described<sup>19</sup>. The reaction clean-ups were performed with GeneOn magnetic beads as described above. Barcode adaptors (1 µl, 20 µM) for the first index were ligated to the sticky ends of DNA fragments by using T4 DNA ligase<sup>19</sup>, incubated at 16 °C for at least 12 h. Each individual sample was provided with a different barcode of a set of 384 different indices (adapter and barcode sequences are available upon request). Equal volumes of the 384 individually barcoded adapter-ligated products were pooled. The pooled DNA was precipitated by adding 20 µl GeneOn magnetic beads and 650 µl Binding Buffer (GeneOn Purification Kit, GO-PCRC-5000)



**Figure 1. Assembly workflow.** (a) Assembly of individual BAC clones from paired-end and mate-pair read data. (b) Data integration procedures for pseudomolecule construction.

to 500  $\mu$ l pooled DNA. The suspension was mixed and incubated at room temperature for 5 min. The beads containing the DNA were reclaimed using a magnet, and the clear supernatant was discarded. The beads were washed twice with 500  $\mu$ l of 70% ethanol and dried completely. The DNA was eluted in 52  $\mu$ l EB. The sample was size-separated by using standard agarose gel electrophoresis (2% agarose gel, HyAgarose, 16250). DNA was revealed using ethidium bromide and excitation by visible blue light emitted from a Dark Reader blue light transilluminator (Clare Chemical Research) to select the target fragments (580–620 bp). The target region was extracted in 27  $\mu$ l EB using the QIAquick Gel Extraction kit (QIAGEN). The second index was introduced using the adapter-ligated products as template DNA (98 °C for 30 s, 10 cycles of: 98 °C for 10 s, 65 °C for 30 s and 72 °C for 30 s, final extension 72 °C for 5 min) (Enzymatics, CM0075) and PCR products (target region: 580–620 bp) were recovered by agarose gel electrophoresis (2% agarose gel, HyAgarose, 16250) as described above. Index primers were used for barcoding each 384 pooled BAC samples (index primer sequences are available upon request). The average size of the PCR products was determined by using an Agilent 2100 Bioanalyzer (Agilent DNA 1,000 Reagents). Typical average size of the libraries was between 574 to 674 bp. PCR products were quantified using real-time PCR and pooled for sequencing in equal proportion<sup>20</sup>. Paired-end sequencing (2  $\times$  100 cycles; first index: 11 cycles, second index: 8 cycles) was performed on the Illumina HiSeq2000 platform (Data Citation 16, Data Citation 17, Data Citation 18).

#### *Mate-pair sequencing of MTP BACs*

For the construction of mate-pair libraries (10 and 20 kb span size), 96 BACs corresponding to 6  $\mu$ g DNA were pooled into one tube. The DNA was fragmented to 10 or 20 kb by using the HydroShear DNA Shearing system from GeneMachines (10 kb: large assembly, speed code 12, cycles 12, volume 250  $\mu$ l; 20 kb: large assembly, speed code 13, cycles 20, volume 150  $\mu$ l). Following DNA fragmentation, the fragments were purified by using 0.6 volumes magnetic beads (Axygen, MAG-PCR-CL-250). The samples were mixed and incubated at room temperature for 10 min. Beads containing the DNA were reclaimed by using a magnet plate (96S Super Magnet Plate, ALPAQUA, A001322), and the clear supernatant was discarded. The beads were washed twice with 500  $\mu$ l of 70% ethanol and dried completely. For the elution of DNA the beads were resuspended in 80  $\mu$ l EB. End-repair and biotin-labeling were performed as described<sup>21</sup>. End-repaired DNA was purified using 0.6 volumes magnetic beads (Axygen, MAG-PCR-CL-250) as described for the purification of hydro-sheared DNA. The DNA was eluted in 79  $\mu$ l EB. 20 kb libraries (20–26 kb range) were size-selected using agarose gel (0.6%) electrophoresis. The ligation of the libraries, was performed by adding 1  $\mu$ l Barcode Adaptor (20  $\mu$ M, sequences are available upon request), 10  $\mu$ l T4 DNA ligase (Enzymatics, L603-HC) in a total volume of 100  $\mu$ l (20 °C, 15 min). 15 individually barcoded adaptor-ligated DNAs (10 kb) were pooled in equimolar manner and size-fractionated (9–11 kb) using agarose gel (0.6%) electrophoresis. DNA circularization and removal of non-circularized DNA was as described<sup>22</sup>. The DNA was isolated from the gel using the QIAquick Gel Extraction kit as described by the manufacturer (QIAGEN). Circular DNA was fragmented using the Covaris S2 device (10% duty cycle, 10 intensity, 1,000 bursts per second, 22 min (11 min) treatment time for 10 kb (20 kb) libraries in TC13 Covaris tubes), and biotinylated fragments derived from true mate-pair ligation events were purified using streptavidin-coupled Dynabeads (M-280, Invitrogen)<sup>19</sup>. Ends of the DNA fragments were repaired and provided with Illumina paired-end adapters as described for the construction of shotgun libraries. The bead-bound DNA was PCR-amplified using Phusion polymerase (NEB) (98 °C for 30 s, 18 cycles of: 98 °C for 10 s, 65 °C for 30 s, 72 °C for 30 s and a final extension: 72 °C for 5 min) using manufacturer's protocols (NEB). Size-selection was essentially performed as described for shotgun library construction. For the 10 kb (20 kb) mate-pair libraries, DNA in the size range between 270–420 bp (400–600 bp) was isolated and purified using the QIAquick Gel Extraction kit according to manufacturer's instructions (QIAGEN). The average size of the paired-end BAC libraries was determined electrophoretically using an Agilent 2100 Bioanalyzer (Agilent DNA 1,000 Reagents). Libraries were quantified using Real-Time PCR<sup>20</sup>. The mate-pair libraries were paired-end sequenced using the Illumina HiSeq2500 device (10 kb library: 150 cycles, 20 kb mate-pair library 50 cycles). Raw data are available as Data Citation 19, Data Citation 20, Data Citation 21).

#### Barley chromosomes 2H and 0H (EI)

##### *Shotgun sequencing of MTP BACs*

QRep 384 Pin Replicators (Molecular Devices, New Molton, UK) were used to inoculate clones from stock plates into 384 square deep well culture plates containing 140  $\mu$ l 2  $\times$  YT media supplemented with 12.5  $\mu$ g ml<sup>-1</sup> chloramphenicol<sup>18</sup>. The culture plates were sealed with a gas permeable seal and incubated for 22 h at 37 °C in a shaking incubator (200 r.p.m.). Cells were harvested by centrifugation (20 min, 3,220 g, 4 °C), the supernatant was discarded. BAC DNA was prepared using a modified alkaline lysis protocol (Beckman Coulter, High Wycombe, UK). Cell pellets were resuspended in 8  $\mu$ l of Resuspension Buffer (RE1) using a Microplate Shaker TiMix 5 control (Edmund-Buehler, Hechingen, Germany) (10 min, 1,400 r.p.m.). Cells were lysed by adding 8  $\mu$ l of the lysis solution (L2). After shaking (5 min, 500 r.p.m.) 8  $\mu$ l of cold Neutralisation Buffer (N3) were added. The plate was shaken (10 min, 500 r.p.m.) followed by a centrifugation (20 min, 3,220 g, 4 °C). The clear supernatant (14.33  $\mu$ l) was transferred

to a 384 well PCR plate, which contained 1  $\mu$ l of CosMc beads per well. The plate was mixed briefly (500 r.p.m.), 10  $\mu$ l of isopropanol was added and the suspension was mixed briefly again (500 r.p.m.). The plate was incubated at room temperature for 15 min to allow precipitation of the DNA onto the beads. The plate containing the DNA precipitate was moved onto a 96 pin 384 well plate compatible magnet (Alpaqua, Beverly, MA, USA) and left for 5 min for the beads to pellet. The supernatant was discarded and the beads were washed three times with 20  $\mu$ l 70% ethanol while placed in the magnet and air dried (room temperature, 5 min). The DNA was eluted from the beads in 20  $\mu$ l of 10 mM Tris HCl (pH 8.0) and transferred to fresh 384 well PCR plate. To remove contaminating host *E. coli* gDNA samples were treated with Epicentre Plasmid Safe ATP dependent DNase (Cambio, Cambridge, UK), which digests the fragmented *E. coli* and nicked BAC DNA but leaves supercoiled BAC DNA intact. To 20  $\mu$ l of DNA 2.5  $\mu$ l of 10x Reaction buffer, 1  $\mu$ l 25 mM ATP, 0.1  $\mu$ l ATP dependent DNase (10  $\mu$   $\mu$ l $^{-1}$ ) and 1.4  $\mu$ l water was added, and the samples were incubated at 37 °C (8 h) followed by 70 °C (20 min) to inactivate the DNase. Sequencing libraries (single index) from the initial sixteen 384 well plates of BACs (2H chromosome) were constructed in 384 well PCR plates (Fortitude, Wotton, UK) using the Epicentre Nextera Kit (Epicentre, Madison, WI, USA) and Robust 2G Taq polymerase (Kapa Biosciences, London, UK). The 384 adapter oligos with 9 bp barcodes each with a hamming distance of 4 (adapter sequences are available upon request) were designed using standard guidelines<sup>22</sup>. Briefly, 1  $\mu$ l of BAC DNA, 1  $\mu$ l Nextera HMW 5  $\times$  Reaction Buffer, 1  $\mu$ l of Nextera Enzyme (diluted 50-fold in 50% glycerol, 0.5  $\times$  TE pH 8.0) and 2  $\mu$ l of water were combined and incubated (5 min, 55 °C) as described<sup>23</sup>. For the denaturation of the Tn5 polymerase, 15  $\mu$ l PB Buffer (Qiagen, Manchester, UK) and for the reaction clean-up, 20  $\mu$ l AMPure XP (Beckman, High Wycombe, UK) beads were added using a Caliper Sciclone Robot (Perkin Elmer, Coventry, UK). Following an incubation (5 min, room temperature), the precipitated fragmented DNA was purified using a 96 well ring Magnet (Alpaqua, Beverly, MA, USA). The beads were washed twice with 20  $\mu$ l 70% ethanol while placed in the magnet before being air dried for 5 min. The fragmented DNA was eluted in 5  $\mu$ l 10 mM Tris HCl, pH 8.0 and transferred to a fresh 384 well PCR plate. To 5  $\mu$ l purified, fragmented DNA 2  $\mu$ l of 5  $\times$  2G B Reaction buffer, 0.2  $\mu$ l of 10 mM dNTPs, 0.1  $\mu$ l of Robust 2G Taq polymerase, 0.2  $\mu$ l of 50  $\times$  Nextera Primer Cocktail and 2.5  $\mu$ l 0.2  $\mu$ M barcoded P2 adapter primer were added in a total reaction volume of 10  $\mu$ l and amplified according to the following thermal cycling profile: 72 °C for 3 min, 95 °C for 1 min, followed by 21 cycles of 95 °C for 10 s, 65 °C for 20 s and 72 °C for 3 min. Post amplification the DNA concentration was determined using the Quant-It Picogreen dsDNA assay (Thermo Fisher, Cambridge, UK). Library DNA concentrations typically ranged from 4 to 40 ng  $\mu$ l $^{-1}$  (average of 16 ng  $\mu$ l $^{-1}$ ). For each sample from a 384 well plate a 5  $\mu$ l aliquot was pooled and split into two 2 ml Lo bind Eppendorf tubes (950  $\mu$ l each). To each aliquot 950  $\mu$ l of AMPure XP (Beckman, High Wycombe, UK) beads was added. Samples were mixed, incubated (5 min, room temperature) and placed on a magnet particle concentrator (MPC) until the beads were collected. The supernatant was discarded. The beads were washed twice with 20  $\mu$ l 70% ethanol while placed in the MPC and air dried (5 min). The pooled library was eluted from the beads in 17  $\mu$ l of 10 mM Tris HCl pH 8.0. The two 17  $\mu$ l aliquots of the library were combined and the DNA concentration was determined using the Qbit device with the Quant-It DNA HS Assay (Invitrogen). Typical DNA concentrations were above 100 ng  $\mu$ l $^{-1}$ . The DNA size selection was performed using the Blue Pippin (Sage Science, Beverly, MA, USA). About 3  $\mu$ g of the library in 30  $\mu$ l of 10 mM Tris HCl pH 8.0 and 10  $\mu$ l of the R2 ladder were separated (tight selection protocol, 650 bp) using a 1.5% agarose cassette according to the manufacturer's instructions (Sage Science, Beverly, MA, USA), thereby yielding an average insert size of about 485 bp. Size selected samples were collected in 40  $\mu$ l of TRIS-TAPS buffer, pH 8.0 (Sage Science, Beverly, MA, USA). The average size of the library was determined using a High Sensitivity Chip and an Agilent 2100 Electrophoresis Bioanalyzer (Agilent). The DNA concentration was measured using the Qbit device and the Quant-It DNA HS Assay (Invitrogen). Size selected libraries were quantified using the Kappa Biosciences Illumina library qPCR quantification kit (Kapa Biosciences) on a Step One qPCR machine (ThermoFisher) according to the manufacturer's instructions and compared against a known concentration of a PhiX control library. Several libraries were pooled for sequencing in an equimolar manner, and the final pool was re-quantified for sequencing relative to a standard library of a known concentration using the Kapa Biosciences Illumina library qPCR quantification kit. Sequencing-by-synthesis for 6,144 BACs from chromosome 2H was performed using an Illumina HiSeq2000 device (2  $\times$  100 cycles paired-end, single indexing read, 384 BACs/lane) according to manufacturer's instructions, thereby yielding at least 32 Gb/lane and an average sequence coverage of at least 500-fold per BAC. The remaining BAC clones from 2H (384 BACs/lane) and 0H (2304 BACs/lane) were sequenced with a HiSeq2500 machine (2  $\times$  150 cycles paired-end, dual indexing, rapid mode, yield: at least 30 Gb/lane) using a slightly adapted protocol with an additional normalization step prior to sample pooling. Briefly, a custom panel of 48 P5 and 48 P7 adapter oligos with 9 bp barcodes (with  $\geq$ 4 hamming distance) was designed to individually label up to 2,304 (48  $\times$  48) libraries by dual indexing. A mixture of 2  $\mu$ l of BAC DNA, 0.5  $\mu$ l Nextera 10  $\times$  Reaction Buffer, 0.1  $\mu$ l Nextera Enzyme and 2.4  $\mu$ l water was incubated (5 min, 55 °C). Tn5 denaturation, reaction clean-up, washing, elution and transfer to a fresh 384 well plate were as described for the single-indexing libraries. 5  $\mu$ l purified, fragmented DNA, 2  $\mu$ l of 5  $\times$  Kapa Robust 2G B Reaction buffer, 0.2  $\mu$ l of 10 mM dNTPs, 0.05  $\mu$ l of Kapa Robust 2G Taq polymerase, 1  $\mu$ l 2  $\mu$ M P5 primer, 1  $\mu$ l 2  $\mu$ M P7 primer were combined (reaction volume of 10  $\mu$ l) and amplified according to following thermal cycling profile: 72 °C for 3 min, 95 °C for 1 min, followed by

16 cycles of 95 °C for 10 s, 65 °C for 20 s and 72 °C for 3 min. The size profile and quantity was determined as described for single-indexing libraries. Amplified libraries were normalised using MagQuant bead technology (GC Biotech, Netherlands) on a Caliper Zephyr Robot (Perkin Elmer), essentially as described by the manufacturer. Normalised libraries were eluted in 10  $\mu$ l of 10 mM Tris HCl pH 8.0 and transferred to a fresh 384 well PCR plate. 5  $\mu$ l of 384 normalized samples were pooled (total volume 1,920  $\mu$ l). Purification using AMPure XP beads, washing, elution, size-selection (Blue Pippin) and quality checks prior to sequencing were essentially as described for single indexing libraries. Sequencing-by-synthesis of pooled libraries (2,304 BACs) was performed using an Illumina HiSeq2500 device (rapid run mode, 2  $\times$  150 cycles paired-end, dual indexing reads) according to manufacturer's instructions. At least 40 Gbp/lane, and an average sequence coverage of >100-fold per BAC were obtained (Data Citation 22, Data Citation 23, Data Citation 24, Data Citation 25).

#### *Mate-pair sequencing of MTP BACs*

BAC clones were inoculated as described for the preparation of shotgun libraries. The bacterial cultures were grown for 6 h at 37 °C in a shaking incubator at 200 r.p.m., and 384 clones were pooled. The pool was used to inoculate 250 ml 2  $\times$  YT media supplemented with chloramphenicol (12.5  $\mu$ g ml $^{-1}$ ). The cultures were incubated (18 h, 37 °C, 200 r.p.m.). Cells were harvested by centrifugation (3,220 g, 20 min, 4 °C), and the supernatant was discarded. Alkali lysis and DNA isolation steps were performed using the Large Construct kit (Qiagen, UK) essentially following the manufacturer's instructions. The DNA was resuspended in 4.75 ml Buffer Ex, 100  $\mu$ l 100 mM ATP (Fisher Scientific, UK) were added and contaminating *E. coli* DNA was removed using 150  $\mu$ l ATP dependent Exonuclease (Qiagen). During the incubation (1 h, 37 °C) a Qiagen Tip-100 column (Qiagen) was equilibrated in Buffer QBT (Qiagen). 5 ml of Buffer QS were added to the DNA, and the sample was applied to the equilibrated column. The column was washed twice with 10 ml of Buffer QC (Qiagen). The DNA was eluted with 7.5 ml of pre-warmed (65 °C) Buffer QF (Qiagen). The DNA was precipitated by adding 0.7  $\times$  volume of room temperature isopropanol and centrifugation (20 min, 3,220 g, 4 °C). The pellet was washed twice with 70% ethanol, air dried and dissolved in 200  $\mu$ l TE buffer according to manufacturer's guidelines. The DNA concentration was measured using a Qubit Fluorometer (Thermo Fisher, Cambridge, UK) and adjusted with water to 13 ng  $\mu$ l $^{-1}$ . For tagmentation 200  $\mu$ l diluted DNA were equilibrated (6 min, 55 °C) and subsequently provided with 52  $\mu$ l 5  $\times$  Tagment Buffer Mate-Pair and 8  $\mu$ l Mate-Pair Tagmentation Enzyme (Illumina, San Diego, USA). After the incubation (30 min, 55 °C), 65  $\mu$ l Neutralize Tagment Buffer (Illumina, San Diego, USA) were added, and the reaction was incubated (5 min, room temperature). One volume CleanPCR beads (GC Biotech, Alphen aan den Rijn, The Netherlands) was added, and the DNA was purified using magnetic separation. The DNA was eluted in 170  $\mu$ l of nuclease-free water, quantified using a Qubit fluorometer (DNA HS assay, Invitrogen) and analysed using the Agilent Bioanalyser (DNA 1,200 chip, Agilent, Stockport, UK). Strand displacement was performed by combining 105.3  $\mu$ l of fragmented DNA, 13  $\mu$ l 10x Strand Displacement Buffer (Illumina), 5.2  $\mu$ l dNTPs (Illumina), 6.5  $\mu$ l Strand Displacement Polymerase (Illumina) and incubation (30 min, room temperature). CleanPCR beads (0.75 volume) were added and the DNA was purified using a magnet. The DNA was eluted in 30  $\mu$ l nuclease-free water. The concentration was measured (Qubit, DNA HS assay, Invitrogen), and a 1.6 diluted sample was analysed using the Agilent Bioanalyser (DNA 1,200 chip, Agilent, Stockport, UK). Size selection was performed using a Pippin Blue (Sage Science, Beverly, MA, USA). 30  $\mu$ l DNA were provided with 10  $\mu$ l loading buffer and separated on a 0.75% agarose cassette (size selection centered at 7 kb and collection between 6–8 kb) according to the manufacturer's instructions (Sage Science, Beverly, MA, USA). Size selected samples were collected in 40  $\mu$ l of TRIS- TAPS buffer (pH 8.0) (Sage Science, Beverly, MA, USA), and analysed using the Agilent Bioanalyser (high sensitivity chip, Agilent, Stockport, UK) to determine the final library size. The DNA concentration was measured using the Qubit device and the Quant-It DNA HS Assay (Invitrogen). Circularisation was performed by combining 40  $\mu$ l size selected DNA, 12.5  $\mu$ l 10  $\times$  circularisation buffer (Illumina), 3  $\mu$ l Circularisation Enzyme (Illumina) and 75  $\mu$ l nuclease-free water. The reaction was incubated at 30 °C overnight. Linear DNA was digested by adding 3.75  $\mu$ l Exonuclease (Illumina) and incubation (30 min, 37 °C). The enzyme was inactivated by heat (30 min, 70 °C) and the addition of 5  $\mu$ l stop ligation (Illumina). Circularised DNA (130  $\mu$ l) was sheared in a Covaris MicroTube AFA Fiber (Pre-slit, Snap-cap, 6  $\times$  16 mm; 2 cycles of 37 s, 10% duty cycle, 200 cycles per burst, 4 intensity, 4 °C) using the Covaris S2 device (Covaris, Massachusetts, USA). M280 Dynabeads (Thermo Fisher) were prepared as described (Illumina). 130  $\mu$ l washed M280 beads were added to the fragmented DNA, mixed and placed on a lab rotator (20 min, room temperature). Library molecules were affinity purified and washed as described (Illumina). The beads were resuspended in a mixture of 85  $\mu$ l nuclease free water, 10  $\mu$ l 10x End Repair Reaction Buffer (Illumina) and 5  $\mu$ l end repair enzyme mix (Illumina) and incubated (30 min, 30 °C). End repaired library molecules bound to M280 beads were washed as described (Illumina). A-Tailing and adapter ligation were performed according to manufacturer's instructions (Illumina). For PCR amplification, the beads were resuspended in a reaction mixture (20  $\mu$ l nuclease-free water, 25  $\mu$ l 2x Kappa HiFi (Kappa Biosystems, London, UK), 5  $\mu$ l Illumina Primer Cocktail) and amplified (98 °C for 3 min, 12 cycles of 98 °C for 10 s, 60 °C for 30 s, 72 °C for 30 s followed by 72 °C for 5 min and storage of the sample at 4 °C). Beads were removed by magnetic separation and 45  $\mu$ l of the products were transferred to a 2 ml DNA Lobind Eppendorf tube. The DNA was precipitated by addition

of 31.5 µl CleanPCR beads (GC Biotech, Alphen aan den Rijn, The Netherlands). The beads were washed twice with 100 µl 70% ethanol, and the final library was eluted in 20 µl resuspension buffer (GC biotech). The DNA concentration was determined (Qubit, DNA HS assay, Invitrogen), followed by analysis using the Agilent Bioanalyzer (High sensitivity chip, Agilent, Stockport, UK). Up to 12 mate-pair libraries were pooled in an equimolar manner and measured using the Kappa qPCR Illumina quantification kit. Sequencing-by-synthesis of pooled mate-pair libraries was performed using an Illumina HiSeq2500 device (rapid run mode, 2 × 150 cycles paired-end, single indexing reads) according to manufacturer's instructions (Data Citation 26, Data Citation 27).

#### Sequence assembly of individual BACs

**Assembly of gene-containing BACs (UCR/JGI).** A total of 15,661 gene-bearing BACs were paired-end sequenced (2 × 100 cycles) using the Illumina HiSeq2000 platform (Illumina, Inc., San Diego, CA, USA) applying a combinatorial pooling design<sup>24</sup>, as described in Munoz-Amatriain *et al.*<sup>13</sup>. Reads were quality trimmed, deconvoluted, and then assembled BAC-by-BAC using Velvet version 1.2.09 (ref. 25) with the parameter k set to 45. Sequences of an additional 50 randomly chosen BACs included in Munoz-Amatriain *et al.*<sup>13</sup> were derived using the Sanger method by Jane Grimwood (US Department of Energy Joint Genome Institute) and Jeremy Schmutz (HudsonAlpha Institute for Biotechnology), including shatter and transposon sequencing. The assignment of BACs to chromosome arms/peri-centromeric regions was performed using CLARK<sup>26</sup>, an accurate *k*-mer-based classification method that is much faster than BLASTN or MegaBLAST. CLARK makes assignments by using a prebuilt database of *k*-mers that are specific to each chromosome arm/peri-centromeric region.

**Assembly of MTP BACs from barley chromosomes 1H, 3H, 4H, 6H and 7H (FLI and IPK).** A total of 10,148 BACs mainly originating from barley chromosome 3H were sequenced on the Roche 454 system. Reads were deconvoluted and assigned to individual BACs<sup>16</sup>. Reads were quality trimmed according to the manufacturer's recommendations. Reads were screened for *E. coli* and vector sequences with MegaBLAST<sup>27</sup>. Assemblies were then constructed from the clean reads using the MIRA software<sup>28</sup> as described in Steuernagel, *et al.*<sup>16</sup> and Taudien, *et al.*<sup>29</sup>.

A total of 41,004 BACs were sequenced on Illumina machines (mainly HiSeq2000) in pools of up to 672 individually barcoded BAC clones. Paired-end reads were quality trimmed with the CLC toolkit and screened for *E. coli* and vector sequences with MegaBLAST. Assemblies were obtained by running CLC Assembly Cell Version 4.0.6 beta with default parameters. Contigs derived with low read coverage as well as contigs smaller than 500 bp were removed using the criteria described in Beier, *et al.*<sup>17</sup>.

The resultant contigs were then compared to NCBI's nucleotide database using MegaBLAST to check for possible contamination. Contigs with non-plant hits were either completely removed or trimmed.

**Scaffolding of MTP BACs from barley chromosomes 1H, 3H, 4H, 6H and 7H (FLI and IPK).** Scaffolding was performed as described in Beier *et al.*<sup>17</sup>. Briefly, mate-pair reads were mapped against the concatenated assemblies of up to 384 BACs using BWA mem version 0.7.4 (ref. 30) with default parameters. Only read pairs mapping uniquely (minimal mapping quality of Q40) to different contigs of the same BAC assembly were retained. These reads were used to scaffold individual BACs using SSPACE version 3.0 Standard<sup>31</sup>.

If multiple mate-pair libraries were present (MiSeq mate-pair reads as well as HiSeq2000 mate-pair reads) an iterative scaffolding procedure<sup>17</sup> was used.

**Assembly of MTP BACs from barley chromosome 5H (BGI).** Obtained raw sequence reads from 5H MTP BACs were filtered to generate high-quality reads by the following criteria: (1) reads containing more than 2% of Ns or with poly-A structures were removed; (2) reads with ≥ 40% low quality bases for short insert size libraries (60% for large insert size libraries) were excluded; (3) reads containing adapters were removed; (4) PCR duplicates were detected and excluded; (5) removal of reads contaminated by *E. coli*, vector sequences or phage sequences. High-quality reads were then used for assembly.

BACs were assembled using SOAPdenovo version 2.01 (ref. 32) multiple times using different k and m values (main parameter in SOAPdenovo assembly). In total each BAC was assembled 45 times (k from 33 to 66, only odd numbers and m from 1 to 3). The N50 was examined for each assembly and the assembly with the largest N50 was retained as the final assembly result for each BAC.

**Scaffolding of MTP BACs from barley chromosomes 5H (BGI).** Assemblies from paired-end sequences were used as reference for mapping 2, 5 and 10 kb mate-pair reads obtained from barley genomic WGS data with SOAPaligner/soap2 version 2.21 with parameters -p 6 -v 3 -R. Mate-pair read pairs mapped in this fashion were used in conjunction with the corresponding paired-end read pairs to re-assemble each BAC using SOAPdenovo 2.01 as described above.

**Assembly of MTP BACs from barley chromosomes 2H and '0H' (EI).** Minimal tiling path BACs from (i) barley chromosomes 2H or from (ii) fingerprinted contigs not assigned to chromosomes (termed '0H') were sequenced. After demultiplexing, sample quality control (QC) information was generated using FastQC<sup>33</sup>. Contamination screening was carried out using Kontaminant<sup>34</sup>. Reads were screened using a *k*-mer size of 21 against a range of potential contaminants (*Phi X*, *E. coli*, *Enterobacter cloacae*

genomic DNA and BAC vector) and contaminated reads or reads with quality values  $< 30$  were removed.

ABySS assembler (v1.5.1)<sup>35</sup> was used to assemble the filtered paired-end reads of each BAC individually (k-71, l-91 b-0). Paired-end contigs were compared to NCBI's NR database using BLAST to check for hits to non-plant organisms using e-value 1e-4 as threshold. The obtained hits were compared to NCBI taxonomy using 'fastacmd' to obtain common names used to check for any non-plant hits.

**Scaffolding of MTP BACs from barley chromosomes 2H and '0H' (EI).** Illumina Nextera mate-pair libraries were created from pools of 384 BACs. After quality checking the reads using PAP<sup>34</sup>, the reads were merged using FLASH (version 1.2.9)<sup>36</sup>. Nextclip (v0.8)<sup>37</sup> was run on the flashed reads to trim the junction adapters. A k-mer-based approach was used to assign mate-pair reads to individual BACs with KAT (v1.0.4) (<https://github.com/TGAC/KAT>). Scaffolding and gap closing were performed on each BAC individually using an in-house shell script (available from GitHub: <https://github.com/DhSaTGAC/BAC-assembly-pipeline.git>). SOAPdenovo scaffolder version 2.01 (ref. 38) was applied to scaffold the ABySS paired-end contigs using the k-mer classified mate-pair reads with parameters  $k = 41$ , -G 30, -F, -w and -L 100. The resulting scaffolds were then edited to replace long stretches ( $> 20$ ) of C/G with 'N' characters as SOAP is known to substitute 'N's within paired-end contigs to C/G. The scaffolds were then passed through GapCloser (v1.12-r6), a SOAP2 module, to fill in long stretches of 'N's produced during the scaffolding steps. Contigs and scaffolds shorter than 500 bp were removed to produce the final assembly per BAC.

**Splash contamination checks of MTP BACs from barley chromosomes 2H and '0H' (EI).** The raw reads within each plate were aligned to one side of the vector sequence adjacent to the restriction enzyme cut site using exonerate<sup>39</sup>. Substrings of size 20 bp were extracted from aligning reads containing the BAC sequence adjacent to the vector sequence. Flanking sequences from each BAC were clustered based on a Hamming distance  $< 3$  and consensus sequences generated to account for sequencing errors. These were compared with neighboring wells to check for potential contamination caused by splash during lab processing steps. Where contamination between neighboring wells was indicated, the assembled contigs from each BAC in question were aligned in a pairwise fashion using exonerate and the total percentage of similar sequence ( $\geq 99\%$  identity) was computed. In cases where neighboring BACs shared more than 10% similar sequence, both BACs were resequenced.

#### Pseudomolecule construction

**Initial contamination removal.** Sequence assemblies of 66,586 MTP clones, 5,468 non-MTP BACs and 15,044 gene-bearing clones<sup>13</sup> (total number of unique BACs: 87,098) were combined into a single FASTA file (Data Citation 28, Data Citation 29, Data Citation 30). If a clone had two or more independent sequence assemblies, we selected the one with the largest N50 value for further analyses. BAC assemblies were aligned to a custom library of potential contaminants (Data Citation 31) including phages, bacterial and vector sequences using megablast<sup>27</sup>. Regions aligning to contaminants (criteria: (alignment length  $\geq 500$  bp **AND** identity  $\geq 80\%$ ) **OR** (identity  $\geq 90\%$ )) were removed from the assembly using UNIX scripts and BEDTools<sup>40</sup>. Sequences shorter than 500 bp or consisting of less than 500 proper nucleotides (ACGT characters) after contamination removal were discarded. This step removed 55.5 Mb (0.5%) of the assembled BAC sequence.

**Sequence alignment of BACs sequences and overlap detection.** After contamination removal, a set of 87,075 BAC assemblies (Table 1, Data Citation 32) was aligned against itself using megablast<sup>27</sup> with a word size of 44, retaining only alignments with identity  $\geq 99\%$  and alignment length  $\geq 500$  bp. Two sets of overlaps (stringent and permissive) between BACs were defined from the BLAST results of all BACs against each other. Pairs of BACs were considered as potentially overlapping under stringent criteria if there was at least one high-scoring pair (HSP) with alignment length  $\geq 5$  kb and identity  $\geq 99.8\%$ . Under permissive criteria, we required at least one HSP with alignment length  $\geq 2$  kb and identity  $\geq 99.5\%$ . For all pairs of potentially overlapping BACs (under either set of criteria), the size of their overlapping regions was determined using UNIX scripts and BEDTools<sup>40</sup> as the extent of non-redundant regions in the BAC sequences (i.e., contigs or scaffolds) contained in HSPs  $\geq 500$  bp and identity  $\geq 99.5\%$  between BAC sequences having at least one HSP with alignment length  $\geq 5$  kb and identity  $\geq 99.8\%$  (stringent criteria) or alignment length  $\geq 2$  kb and identity  $\geq 99.5\%$  (permissive criteria). HSPs less than 200 bp apart were combined into one with BEDTools (command 'merge'). BAC overlap information was imported into the R statistical environment<sup>41</sup> for use in genetic anchoring and merging sequence assemblies of adjacent BAC clones (see section 'Construction of the BAC overlap graph').

**Alignment of BACs to the BioNano map of barley cv. Morex.** An optical map of the genome of barley cv. Morex was generated using the Irys platform of BioNano Genomics using Nt.BspQI as the nicking enzyme. Further details of the optical map procedure are described in Mascher *et al.*<sup>42</sup> An *in silico* BspQI digest was performed with the Knickers software (<http://www.bionanogenomics.com>) using default parameters. Restriction maps of BAC sequences were aligned to the BioNano map of barley cv. Morex<sup>42</sup> (Data Citation 33) with IrysView software<sup>43</sup> (<http://www.bionanogenomics.com>) using the

MTP chromosome	no. of BACs in MTP	no. of sequenced BACs	no. of anchored BACs*	average no. of sequences	average N50 (kb)
1H	6,993	6,983 (99.9%)	6,410 (91.8%)	7.6	81.2
2H	9,061	8,969 (99.0%)	8,195 (91.4%)	9.9	104.5
3H	8,841	8,807 (99.6%)	8,303 (94.3%)	7.7	87.5
4H	8,314	8,306 (99.9%)	7,783 (93.7%)	6.7	91.2
5H	8,426	8,358 (99.2%)	7,573 (90.6%)	9.7	72.2
6H	8,305	7,886 (95.0%)	6,476 (82.1%)	7.4	70.7
7H	8,576	7,970 (92.9%)	6,842 (85.8%)	8.5	65.5
‘0H’†	8,256	8,031 (97.3%)	6,714 (83.6%)	7.6	83.6
Non-MTP	—	21,765	20,397 (93.7%)	14.5	33.7
Total	66,772	87,075	78,693 (90.4%)	9.8	70.3

**Table 1. BAC assembly and anchoring statistics.** \*Number and percentage of BAC clones that have been assigned genetic positions in the POPSEQ map. †BAC clones in physical contigs that had not been assigned to chromosomes.

command line tool RefAligner (version 3827) with the following parameters ‘-M 2 -T 1e-4 -extend 1 -biaswt 0’ to report all alignments with a confidence score  $\geq 4$ .

**Construction of the updated POPSEQ map of the Morex x Barke mapping population.** An ultra-dense linkage map had been constructed previously<sup>5</sup> by shallow whole-genome shotgun sequencing of 90 recombinant inbred lines (RILs) derived from a cross between the barley cultivars Morex and Barke. We wished to increase the resolution of this map by reducing the average fraction of missing data per SNP marker. Towards this aim, we sequenced the existing Illumina paired-end libraries of 87 RILs to higher coverage (2–3x) and combined them (Data Citation 34) with the existing read data set<sup>5</sup> (ENA accession: ERP002184). Map construction followed the procedures described in Chapman *et al.*<sup>44</sup>. Reads were aligned to the whole-genome shotgun assembly of barley cv. Morex<sup>4</sup> (NCBI accession: CAJW01) with BWA mem version 0.7.5a (ref. 45). Sorting, conversion to BAM format and removal of duplicate reads was done with PicardTools version 1.100 (<http://broadinstitute.github.io/picard/>). Variant detection and genotype calling were performed with SAMTools version 0.1.19 (commands ‘samtools mpileup -BD’ and ‘bcftools view -cvg’). The resultant VCF file was filtered using an AWK script (Supplementary Text S3 of Mascher *et al.* 2013 (ref. 46)). Heterozygous genotype calls were set to missing if their read depth was 0 or their genotype quality below 3. Heterozygous genotype calls were set to missing if their read depth was below 3 or their genotype quality below 5. Variants with (i) a quality scores below 40, (ii) more than 10% heterozygous genotype calls, (iii) more than 90% missing data after genotype call filtering, or (iv) a minor allele frequency below 5% were discarded. SNP information was aggregated at the contig level to derive consensus genotypes as described in the section ‘Framework map construction’ in the Methods section of Chapman *et al.*<sup>44</sup>. For map construction with MSTMap<sup>47</sup>, the population type ‘RIL8’ was used. Additional contigs were inserted into the framework map as described in Chapman *et al.*<sup>44</sup> (section ‘Anchoring scaffolds onto the framework map’) using previously published read data<sup>5</sup>. Variant calling and map construction were done for the Oregon Wolfe Barley (OWB) doubled haploid population using the same procedures with the following two changes: (i) heterozygous genotype calls were excluded and (ii) the population type ‘DH’ was used for map construction with MSTMap<sup>47</sup>. Map positions in the OWB map were interpolated into the Morex x Barke map using loess regression in R<sup>41</sup>. A consensus position was derived as follows: if map positions disagreed by more than 5 cM in both maps, a contig was considered unanchored; otherwise, the Morex x Barke position was preferred if available. The final map assigned genetic positions to 791,176 WGS contigs (Table 2, Data Citation 35), compared to 723,499 anchored contigs in the original POPSEQ map<sup>5</sup>.

**Genetic anchoring of single BAC clones.** The genetic positions of Morex WGS contigs in the updated POPSEQ map were lifted to BAC sequences via sequence alignment. The set of all contigs of the whole-genome shotgun assembly of barley cv. Morex<sup>4</sup> (NCBI accession: CAJW01) was aligned to all BAC assemblies with megablast<sup>27</sup> using a word size of 44 and retaining only alignments with identity  $\geq 99.8\%$  and alignment length  $\geq 1,000$  bp. For each BAC clone, the genetic positions of WGS contigs aligning to its constituent sequences were tabulated and a genetic position of a clone was derived using a majority rule with functions of the R package ‘data.table’ (<https://cran.r-project.org/web/packages/data.table/index.html>). Ninety per cent of contigs assigned to a BAC had to originate to the major chromosome and the standard deviation of genetic positions had to be  $\leq 3$  cM. BACs without alignments to anchored WGS contigs were considered as unanchored; those not meeting the consistency criteria were flagged as ‘inconsistently anchored’. In the second step, unanchored clones were positioned by utilizing positional information from neighboring BACs. We considered as neighbors of a given clone B all those BACs that overlapped for at least 10% of their assembled lengths with clone B. The genetic position of an

Chromosome	No. of anchored WGS contigs	Length of anchored WGS contigs (Mb)
1H	74,184	123.7
2H	130,436	202.6
3H	119,131	187.6
4H	96,642	170.6
5H	117,314	177.8
6H	121,384	168.4
7H	132,085	190.2
Total	791,176	1220.9

**Table 2.** Summary statistics of the updated POPSEQ map of the Morex WGS assembly.

unanchored BAC B with an assembled length  $\leq 300$  kb were borrowed from its neighbors if all of them were anchored to same chromosome and the standard deviation of genetic coordinates was at most 3 cM. If these criteria were fulfilled, the genetic position of B was set to the arithmetic mean of the genetic coordinates of its neighbors. Genetic positions were determined for 78,693 (90.4%) BACs (Table 1, Data Citation 36).

**Construction of the BAC overlap graph.** We converted the overlap information between BACs in a graph structure using the R package 'igraph'<sup>48</sup>. Nodes represented BACs. An edge was drawn between two nodes (BACs) if the criteria regarding sequence overlap and consistency of positional information were fulfilled as detailed below. The edge weights were set to the cumulative length of intervals in which two adjacent BACs overlapped. We named the connected components of this graph 'clusters'. These clusters are analogous to physical contigs in that they represent overlaps between BACs. In contrast to physical contigs, overlaps between BACs in the cluster graph are not derived from restriction maps, but from sequence alignments.

The initial overlap graph was refined in subsequent steps by adding edges that were supported by (i) additional information about links between BACs derived from BAC end sequences, (ii) the genome-wide physical map of barley<sup>3</sup> or (iii) the BioNano map. After each refinement step, we checked for the existence of branches in the overlap graph. Such branches should not occur in a linear genome and may have arisen from spurious sequence alignments or incorrect positional information. We also determined genetic locations of clusters by aggregating the positional information of their constituent BACs using a majority rule, requiring all anchored BACs to come from the same chromosome and the standard deviation of their genetic coordinates to be  $\leq 5$  cM. Clusters not meeting these criteria were considered inconsistently anchored. Edges giving rise to branches or to inconsistent genetic positions were detected and removed. To detect branches, we calculated a minimum spanning tree (MST) of each cluster using Prim's algorithm<sup>49</sup> as implemented in the igraph<sup>48</sup> function 'minimum.spanning.tree()'. A geodesic of the MST of maximal length was determined with the igraph function 'get.diameter()' and set as the linear (i.e., branchless) backbone of the cluster. In the MST, each BAC B was either part of the diameter or attached to a single BAC of the backbone, i.e., there existed a path from B to one and only one BAC of the backbone. The length of this path to a member of the backbone was defined as its rank. Groups of BACs attached the same backbone BAC were considered as a 'BAC bin' of the cluster. Branches were defined as groups of nodes with rank  $> 1$ . A cluster was said to be branched if it contained branches, i.e., had a non-linear structure. Note that due to redundancies in the BACs selected for sequencing, we expect BACs with rank equal to 1. After each insertion or removal of edges or nodes, connected components, MST backbones and genetic positions of clusters were re-calculated, and branches and inconsistencies with genetic data removed if necessary. The summary statistics of the overlap graph after each step are given in Table 3. The final clustering results summarized in Table 4 are available as Data Citation 36).

#### Step 1: Initial overlap graph from links within FP contigs

In the initial overlap graph, an edge between two BACs was drawn if both BACs were (i) on the same fingerprinted (FP) contig, (ii) the overlapping regions between them accounted for  $\geq 5\%$  of the length of either BAC and (iiiA) there were genetically anchored to the same chromosome within 3 cM of each other or (iiiB) one or both clone were unanchored. To determine overlap lengths, we used the permissive set of overlaps. BACs that were inconsistently anchored or whose assembled length was  $> 300$  kb were excluded from the graph. The initial graph had both branched and inconsistently anchored clusters. To remove inconsistencies in genetic positions, all edges involving unanchored clones were deleted in clusters showing inconsistent genetic positions. To remove branches in the initial graph, we first removed nodes representing non-MTP clones that were part of branches. This step was iterated twice. In the next steps, BACs in branches and originating from the set of gene-bearing BACs<sup>13</sup> were excluded. These BACs were sequenced using combinatorial pooling strategy and errors during demultiplexing may have given rise to chimeric assemblies. After these steps, nine clusters with branches remained in the graph. BACs in

Step	Datasets*	Clusters	BACs in clusters	Singleton BACs	Excluded BACs	Cluster N50 <sup>†</sup>	Average cluster size <sup>‡</sup>
1	BAC, FPC	9,637	71,828	13,211	2,036	21	12.9
2	BAC	4,890	79,871	4,002	3,202	60	38.3
3	BAC, OM	4,843	79,884	3,989	3,202	61	38.8
4	FPC, BES, OM	4,653	79,884	3,989	3,202	65	41.2
5	FPC, BES	4,562	79,908	3,965	3,202	66	41.7
6	BAC, OM	4,486	79,918	3,955	3,202	66	42.4
7	FPC, BAC	4,485	79,919	3,954	3,202	66	42.4
8	FPC, OM	4,390	79,919	3,954	3,202	66	43.0
9	exBAC	4,382	80,010	3,938	3,127	66	43.1
10	BAC, OM	4,323	80,010	3,938	3,127	67	43.8
11	FPC, OM	4,259	80,010	3,938	3,127	69	45.2
12	BES, FPC	4,251	80,010	3,938	3,127	69	45.2

**Table 3. Cluster summary statistics after each step of the BAC overlap graph construction.** \*Datasets used in each step (BAC, BAC sequence overlap; FPC, physical map; OM, optical map; BES, BAC end sequences; exBAC, previously excluded BAC assemblies. Consistency with the POPSEQ genetic map was checked in each step. <sup>†</sup>An N50 value  $N$  indicates that half of all clusters contain at least  $N$  BACs. <sup>‡</sup>Arithmetic mean of the number of BACs per cluster.

	1H	2H	3H	4H	5H	6H	7H	Un
Number of clusters	389	605	324	415	549	768	943	242
Number of singletons	65	214	74	78	173	167	162	1190
Assembly length (Mb)	562.8	785.5	704	655.5	687.8	600.2	663.8	130.6
Length in clusters (Mb)	555.9	760.3	695.8	648.4	668.2	581.1	646	28.9
Length in singletons (Mb)	6.9	25.1	8.3	7.1	19.5	19.1	17.7	101.7
N50 (Mb)	2.5	2.1	3.6	2.5	2.0	1.1	1	0.1

**Table 4. Final cluster statistics.**

these branches were removed from the graph. After these steps, the graph was unbranched and showed no inconsistencies with the genetic map. The graph consisted of 9,637 clusters and 13,211 singletons (Table 3).

#### Step 2: Adding links between FP contigs

Next, we added edges between BACs on different FP contigs. An edge between two BACs was drawn if (i) the overlapping regions between them accounted for  $\geq 10\%$  of the length of either BAC and (iii) they were genetically anchored to the same chromosome within 3 cM of each other. Stringent overlap criteria were used in this step. This graph had branches, which were removed in subsequent steps. First, clones shorter than 50 kb or having an N50  $< 10$  kb were excluded. Then, nodes representing non-MTP clones that were part of branches were deleted. This step was repeated once. Then, edges where both clones were part of branches and in different FPCs were removed, followed by another removal of non-MTP clones. In the next step, clones in branches that were longer than 250 kb were removed. These large assemblies may combine sequences of two unrelated BACs as a result of chimeric inserts or cross-contamination between neighboring well positions. Next, gene-bearing clones<sup>13</sup> in branches were deleted. Finally, all remaining clones in branches were discarded. The resultant graph had no branches and all its clusters were consistently anchored to the genetic map. This step reduced the number of clusters from 9,637 to 4,980 and led to the exclusion of 1,166 putatively chimeric BAC assemblies giving rise to non-linear structures (Table 3).

#### Step 3: Adding links with permissive overlap criteria, but support by the BioNano map

In the next steps, we tried to find additional links between BACs that would support the joining of neighboring clusters. This was motivated by our desire to have fewer, but large clusters (i.e., increase the contiguity of the overlap graph) to facilitate the construction of the Hi-C map (see below). Towards this aim, we added edges to the graph using less stringent overlap criteria, but requiring support from other datasets. If the inclusion of an edge gave rise to a branch or map inconsistencies, this edge was removed

again. We note that in some cases edges do not represent true sequence overlaps between BACs, but only evidence for close proximity of two BACs.

In the first step, we added edges between two BACs if (i) they were located at the ends of clusters, (ii) the overlapping regions between them accounted for  $\geq 10\%$  of the length of either BAC, (iii) they were genetically anchored to the same chromosome within 3 cM of each other and (iv) and the link was supported by the BioNano map. The BACs at the ends of clusters were determined from the MST traversals of clusters. Support by the BioNano map means the presence of a single contig of the BioNano map (an ‘optical genome map’ (OM) in BioNano’s nomenclature) that links to two clusters. To find such genome maps, we aggregated the alignment information between BAC sequences and OM at the level of clusters. In the alignment table between BioNano maps and BAC sequences, we only retained the best alignment of each BAC sequence contig. A cluster was considered aligned to a OM if the sum of the confidence scores (as reported by BioNano’s refaligner software) of its BAC sequences was at least 25. A OM was joining two clusters if (i) the distance in the OM between restriction map alignments pertaining to the two clusters was (i)  $\leq 300$  kb and (ii) the order and orientation of alignments to the OM were consistent with the order of BACs in the MSTs of the clusters, requiring a rank correlation above 0.5. Adding all edges meeting these criteria to the overlap graph did not result in branches or inconsistent map positions within clusters. The graph consisted of 4,843 clusters (Table 3).

#### Step 4: Adding links supported by FP contigs, BAC end sequences and the BioNano map

We added edges representing pairs of BAC end sequences linking BACs at ends of clusters on the conditions that (i) these links were supported by the BioNano map and (ii) the joined BACs originated from the same FPC contig. BAC end sequences of cv. Morex (EMBL ENA accessions: HF140858-HF362636, HE975059-HE977519, HF000001-HF140857, HE867107-HE939654, HE939655-HE956691 and HF362637-HF479769) were aligned to all BAC assemblies with megablast<sup>27</sup> using a word size of 28 and considering only hits with identity  $\geq 99.5\%$  and alignment length  $\geq 500$  bp. We identified pairs of BAC end sequences that aligned to BACs B1 and B2 from two different clusters C1 and C2. BACs B1 and B2 were required to be the end of their clusters and to belong to same FPC contig and were less than 200 kb apart from each other in the physical map (using the conversion factor 1 FPC consensus band = 1.24 kb<sup>3</sup>) map. Moreover, we required the clusters C1 and C2 to be connected by a BioNano contig under the criteria described in the section ‘Adding links with permissive overlap criteria, but support by the BioNano map’. If all these criteria were fulfilled, we added an edge between B1 and B2. This step did not introduce branches or inconsistently anchored clusters to the graph. The number of clusters decreased to 4,653 (Table 3).

#### Step 5: Adding links supported by FP contigs and BAC end sequences

In this step, we used BAC end sequences and FP information to find additional links as described in the previous step, but we did not require support by the BioNano map. This step introduced branches to the graph that were removed by pruning newly introduced edges between BACs in branches. The updated graph was composed of 4,562 clusters (Table 3).

#### Step 6: Using FP information and inconsistently anchored BACs to bridge gaps

In previous steps, we had excluded inconsistently anchored BAC assemblies from the overlap analysis. We speculated that many of these assemblies may contain BAC sequences from two unlinked genomic loci as a consequence of chimeric inserts or cross-contamination between neighboring wells during handling of BAC plates for MTP rearraying or sequencing. So if both BACs were fully assembled, one could use their sequences to link BAC clusters under the condition that further evidence corroborates the connection. We identified inconsistently anchored BACs (termed ‘link BACs’) that showed stringent sequence overlaps ( $\geq 10\%$  of the assembled length of either BAC) to two BACs B1 and B2 at the ends of different clusters. We required BACs B1 and B2 to originate from the same FP contig and to be anchored within 1 cM of each other in the POPSEQ genetic map. If these criteria were met, we added an edge between B1 and B2 in the overlap graph. We did not add the link BAC itself to avoid introducing contaminant sequences from other parts of the genome. This step did not introduce branches or inconsistencies with genetic data. The number of clusters decreased to 4,486 (Table 3).

#### Step 7: Using singletons BACs to bridge gaps in FP contigs

In this step, we tried to find single BACs that can close gaps within FP contigs. We identified pairs BACs B1 and B2 that were located on the same FP contigs, but different clusters, and searched for a third B3 that had stringent sequence overlap ( $\geq 10\%$  of the assembled length of either BAC) to both B1 and B2. We required that B3 was a singleton (i.e., a cluster of size 1) and was within 3 cM of both B1 and B2 and the POPSEQ genetic map. If these criteria, were fulfilled we added edges B3  $\leftarrow$  B1 and

B3 < -> B2. No branches or inconsistencies with the POPSEQ map were introduced in this step. This step resulted in the merging of two adjacent clusters and the incorporation of one singleton (Table 3).

Step 8: Using FP information and BioNano data

We searched for links between two BAC clusters that were part of the same FP contig and that were supported by alignments to a single BioNano contig. We searched the BioNano map for links between clusters as described in the section 'Adding links with permissive overlap criteria, but support by the BioNano map'. We required the alignments of connected clusters to be no farther apart than 300 kb and that the corresponding BACs came from the same FP contig and were located within 300 kb in the FP map. Moreover, the order and orientation in the FP contig and the BioNano map were required to be consistent with each other. If these criteria were fulfilled, we added an edge between the BACs at the abutting end of the two connected clusters. This step introduced inconsistencies to the POPSEQ map that were removed by deleting all newly inserted edges in the affected clusters. This step reduced the number of clusters from 4,485 to 4,390 (Table 3).

Step 9: Adding BACs previously considered as inconsistently anchored

We searched for BACs who (i) were flagged as inconsistently anchored because of the standard deviation of the genetic coordinates of the Morex WGS aligned to them was larger than 3 cM, (ii) had stringent overlaps to non-singleton BACs. We required that all Morex WGS contigs aligning to these BACs originated from the same chromosome. We added these BACs and edges leading to them to the overlap graph. This step introduced branches to the overlap graph, which were removed by deleting the newly added BACs in branched clusters. This step resulted in the incorporation of 75 additional BACs into the overlap graph (Table 3).

Step 10: Using BAC overlap information and BioNano data

In this step, we used BAC sequence overlap information and BioNano map data to add edges to the overlap graph. We found potential connections between clusters as detailed in the section 'Adding links with permissive overlap criteria, but support by the BioNano map'. If the two BACs B1 and B2 at the adjoining ends of the two linked clusters were within 3 cM of each other and the overlapping regions was ( $\geq 10\%$  of the assembled length of either BAC), we added an edge between B1 and B2. This step did not introduce branches or inconsistencies with the genetic map. The updated graph consisted of 4,323 clusters (Table 3).

Step 11: Using FP information to bridge gaps

In this step, we aimed to use the BioNano map to close gaps between two BACs B1 and B2 that are near to each other in the physical map and were expected to overlap with a common BAC B3 between them (layout: B1 -> B3 -> B2) based on fingerprinting results, but their sequence assemblies failed to do so, resulting in a short gap between B1 and B2. Towards this purpose, we identified pairs of BACs B1 and B2 that (i) were on the same chromosome less than 3 cM part and (ii) located at the ends of two different overlap clusters and (iii) came from the same FP contigs, (iv) were separated by less than 300 kb in the FPC map with a single BAC B3 between them in the FPC map. Such cases may occur if both B1 and B2 were expected to overlap with B3 according to FPC information, but either the overlapping regions could not be detected in the alignment of the sequence assemblies because of low assembly quality or because of BAC mix-ups during fingerprinting, re-arranging of MTP clones or sequencing library preparation, so that B1 and B2 were separated by a gap in the overlap graph. We added an edge between B1 and B2 if the following conditions were fulfilled: (i) the two clusters of B1 and B2 could be aligned to the same contig of the BioNano map, (ii) the aligned regions were less than 300 kb apart in the BioNano map and (iii) the orientation of the BioNano contigs and the overlap clusters were consistent. This step did not introduce branches or inconsistencies with genetic data. This step decreased the number of clusters from 4,323 to 4,259 (Table 3).

Step 12: Adding links supported by BAC end sequences and the BioNano map

We identified BACs link supported by BAC end sequences and the BioNano map as described in Step 4, but did not require the connected BACs to come from the same FP contig. Added links meeting the criteria to the overlap graph did not create branches or inconsistencies. The final graph consisted of 80,010 BACs in 4,251 clusters and 3,938 singleton BACs (Table 3).

**Construction of non-redundant sequences of BAC overlap clusters**

A non-redundant sequence was constructed for each BAC cluster by detecting and removing sequence overlaps between neighboring BACs using an iterative procedure. In the initial step, the complete sequence of the largest sequence scaffold among the assemblies of all BACs in a cluster was added to the

set of visited BAC sequence scaffolds, all other sequence scaffolds were part of the set of unvisited BAC sequence scaffolds. The set of unvisited sequence scaffolds was then aligned to the visited sequence scaffolds with megablast<sup>27</sup> with a word size of 44, accepting only high-scoring pairs with an alignment length  $\geq 500$  bp and an alignment identity  $\geq 99.5$  bp. Alignments between two sequence scaffolds from BACs B1 and B2 were only allowed if B1 and B2 were separated in the minimum spanning tree of the cluster by no more than 10 BACs. Regions contained in alignments to visited scaffolds satisfying these criteria were subtracted from the unvisited sequence scaffolds using BEDTools<sup>40</sup>. Sequence scaffolds that were composed of less than 500 proper nucleotides (ACGT characters) after subtraction were discarded. The largest sequence scaffold among the unvisited scaffolds was moved from the set of unvisited to the set of visited scaffolds. These steps of alignment, redundancy removal and selection of the largest unvisited scaffold were repeated until no unvisited scaffolds remained. Finally, stretches of N characters at the ends of non-redundant fragments of sequence scaffolds were trimmed with an AWK script. After these procedures had been carried out for all BAC clusters, the resultant non-redundant sequences were written into a single FASTA file (Data Citation 37).

#### Construction of a high-resolution GBS map of the Morex x Barke population

At this stage, we constructed a high-resolution linkage map from GBS data using the non-redundant sequence as a reference for read alignment. This map was used to derive orientations of BAC overlap clusters in the Hi-C map (see 'Orienting clusters by Hi-C and GBS') and to validate the order of clusters in the Hi-C map (see 'Technical Validation'). GBS libraries of 2,398 recombinant inbred lines of the Morex x Barke lines were constructed using published protocols<sup>46,50</sup> and subjected to Illumina or IonTorrent sequencing (Data Citation 38). Adapters were trimmed from GBS reads with cutadapt<sup>51</sup> version 1.8.1. Reads shorter than 30 bp after trimming were discarded. Trimmed reads were mapped to the non-redundant sequence of BAC clusters with BWA<sup>45</sup> mem version 0.7.12. The resultant alignment files were converted to BAM format with SAMtools<sup>52</sup> (version 0.1.19), sorted with Novosort (Novocraft Technologies Sdn Bhd, Malaysia, <http://www.novocraft.com/>) and merged into a single BAM files with Picard (version 1.128, <http://broadinstitute.github.io/picard/>). Multi-sample SNP calling was performed with FreeBayes<sup>53</sup> using the parameters '-i -X -u -n 2 -S 5 -e 2 -m 20 -q 20 --min-coverage 500 -G 200 -F 1 -w --genotype-qualities --report-genotype-likelihood-max'. The resulting VCF file was filtered with an AWK script (Text S3 of Mascher *et al.*<sup>46</sup>). Only bi-allelic SNP with a quality score  $\geq 40$  were considered. Homozygous genotype calls were set to missing if their read depth was below 2 or their quality score below 20. Heterozygous genotype calls were ignored. Variants with more than 50% missing data or a minor allele frequency below 30% were discarded. The filtered SNP-by-individual matrix was imported into the R statistical environment<sup>41</sup> for further processing. After removing samples with less than 6,000 successful genotype calls, the final marker-by-individual matrix was constructed by discarding SNPs with more than 10% missing data. Genetic map construction was done with MSTMap<sup>47</sup> with a *P*-value cut off of  $1 \times 10^{-60}$  using the population type 'RIL8'. The final map included genotypic data from 1,613 individuals at 2,637 variant positions (Table 5, Data Citation 39).

#### Hi-C map construction

Hi-C map construction comprised the steps (i) data alignment to the non-redundant sequence, (ii) ordering and (iii) orienting BAC clusters using Hi-C link information.

**Alignment of Hi-C data to restriction fragments.** A BED file representing all intact HindIII restriction fragments  $\geq 100$  bp within in the non-redundant sequence was constructed using a custom AWK script. Whole genome shotgun reads<sup>4</sup> of barley cv. Morex corresponding to  $\sim 14x$  whole genome coverage were aligned to non-redundant sequence with BWA mem 0.7.12 (ref. 45), converted to BAM format with SAMtools<sup>52</sup>. Duplicate removal and sorting were done with Novosort. The coverage of the non-redundant sequence with WGS reads was calculated with SAMtools<sup>52</sup> using the command 'depth -Q 20 -q 10' and written into a BED file. This file was used to calculate the average coverage of each HindIII

Chromosome	No. of SNPs	No. of bins	Map length (cM)
1H	346	195	133.3
2H	383	231	153.2
3H	385	231	154.9
4H	237	135	115.5
5H	474	265	173.3
6H	362	188	122.7
7H	450	253	143.9
total	2,637	1,498	996.8

**Table 5. Summary statistics of the GBS map.**

fragment using the BEDTools<sup>40</sup> command ‘map’. Fragments with an average coverage below 7 or above 21 were discarded.

Paired-end reads<sup>9</sup> (Data Citation 40) obtained using the Hi-C and TCC protocols<sup>9,54</sup> as described in ref. 42 were trimmed using cutadapt<sup>51</sup> version 1.8.1 using as the adapter sequence the ‘extended’ NheI restriction site (AAGCTAGCTT) created by ligating two blunted HindIII fragments<sup>9</sup>. Trimmed read pairs were mapped as single ends to the non-redundant sequence using BWA mem version 0.7.12 (ref. 45) with parameters ‘-M -P -S’ and then converted to BAM format with SAMtools<sup>52</sup>. After duplicate removal with Novosort (Novocraft Technologies Sdn Bhd, Malaysia, <http://www.novocraft.com/>), BAM files were sorted by read name to group the two mates of a pair together. Hi-C mapping information was then converted from BAM to BED format and assigned to HindIII restriction fragments with BEDTools<sup>40</sup> using the command ‘pairtobed -bedpe -type both’ requiring both mates of a pair to have mapping quality  $\geq 10$ . A custom AWK script was used to calculate the size of sequence fragments that read pairs originated from based on the distance of mapped ends to the next HindIII restriction site. After discarding fragments with size  $\geq 500$  bp, read pairs linking two different clusters (Hi-C links) were tabulated using standard UNIX tools (AWK, sort, uniq) and the link counts for each cluster pair were imported into R<sup>41</sup>.

**Ordering scaffolds by Hi-C.** Clusters whose non-redundant sequence was less than 30 kb or which had less than 20 restriction fragments were not used for making the Hi-C map. Scaffold ordering with Hi-C data was done using a custom R implementation of the algorithm outlined in Burton *et al.*<sup>10</sup>. First, the Hi-C link information was entered into graph structure using the R package ‘igraph’ (<http://igraph.org/r/>). The graph was composed of nodes representing the clusters and of edges representing Hi-C links anchored genetically to the same chromosome within 15 cM of each other. Only links between clusters anchored genetically to the same chromosome within 15 cM of each other were considered. For each of the seven largest connected components (corresponding to the seven chromosomes of barley), a minimum spanning tree was calculated with Prim’s algorithm<sup>49</sup> as implemented in igraph. This resulted in a backbone map into which further nodes (clusters) were inserted so as to minimize the additional weight incurred by each node insertion. Subsequently, the 2-opt heuristics and single node relocation as used in the MSTMap algorithm for genetic mapping<sup>47</sup> were applied to incorporate local perturbations that reduce the weight sum of the initial solution. The resultant paths of each connected component (chromosome) were oriented from short to long arm by comparison to the POPSEQ genetic map.

**Orienting clusters by Hi-C and GBS.** To orient clusters relative to the telomeres of the long and short chromosome arm, clusters were divided into bins of 300 kb size that were ordered by Hi-C as described above. If a cluster comprises several bins, the scaffold orientation can be inferred from the order of its constituent bins in the global Hi-C map of all 300 kb bins, which is oriented on a chromosome scale (from short to long arm) by comparison to the genetic map as described above. Local inversions may arise in the Hi-C map of the bins because of the reduced accuracy of Hi-C mapping when smaller intervals are used to aggregate Hi-C link information. To correct inverted orientations in the bin map, we checked how the relative order of a cluster C and its two adjacent clusters was correlated with that of their constituent bins. If the correlation coefficient was negative, the orientation of cluster C was reversed. If no HiC orientation could be determined, but orienting clusters was possible using GBS marker information, this information was used instead. The orders and orientation of sequence clusters are given in Data Citation 41.

#### Construction of pseudomolecule sequences

We constructed a FASTA file containing a single entry for each barley chromosome (a ‘pseudomolecule’) and an additional entry combining all sequence not anchored to chromosomes. Prior to the construction of pseudomolecules, we (i) identified genes incomplete or missing in the non-redundant sequence, but represented by (a) BAC sequence that had been excluded from the construction of the non-redundant sequence, or by (b) Morex WGS contigs<sup>4</sup>; and (ii) performed a final scan for contaminant sequences.

**Identification of additional gene-bearing sequences.** The sets of (i) barley high-confidence (HC) genes annotated on the WGS assembly of cv. Morex<sup>4</sup> and (ii) barley full-length cDNA (fl-cDNA) sequences<sup>55</sup> were aligned with GMAP<sup>56</sup> version 2014-12-21 to (a) the set of all BAC assemblies, (b) Morex WGS contigs<sup>4</sup> and (c) the non-redundant sequence.

First, we identified genes (as represented by the HC genes or fl-cDNAs) whose best alignment to the set of assembled sequences of all BACs in clusters (as opposed to BACs excluded from the overlap analysis) represented at least 5% more of their coding sequence than their best alignment to the non-redundant sequence. Such cases arise if during the iterative construction of the non-redundant sequence, a sequence contig (or scaffold) C1 that breaks within a gene G is chosen before a contig C2 that contains a larger part of G than C1, but the total length of C1 is larger than that of C2. To amend such situations, we added contigs of type C2 to the non-redundant sequence and removed contigs of the non-redundant sequence that had previously represented the sequence now covered by C2. Towards this purpose, we aligned the sequence of each C2-type contig C to the non-redundant sequence of its BAC cluster of origin with megablast<sup>27</sup> using a word size of 44 and considering only high-scoring pairs with an alignment

length  $\geq 500$  bp and an alignment identity  $\geq 99.5\%$ . Regions of the old non-redundant sequence covered by C (as determined by commands of BEDTools<sup>40</sup> suite) were removed and contig C was added instead. This procedure was performed for each C2-type contig.

Next, we queried the GMAP alignments for genes that had no alignments to the non-redundant sequence, but were represented either in (a) the Morex WGS contigs or in (b) sequences of BACs excluded from the overlap analysis. We considered sequence of type (a) and (b) as 'additional gene-bearing sequences'. We aligned these additional gene-bearing sequences to the non-redundant sequence with megablast<sup>27</sup> using a word size of 44 and considering only high-scoring pairs with an alignment length  $\geq 500$  bp and an alignment identity  $\geq 99.5\%$ . Regions covered by the non-redundant sequence under these alignment criteria were subtracted from the additional gene-bearing sequences and sequence fragments with a length  $\geq 500$  bp were added to the non-redundant sequence.

**Final contamination removal.** We identified regions in the non-redundant sequence that were not covered by whole-genome shotgun reads of cv. Morex. Alignment of WGS reads and read depth calculation were done as described in the section 'Alignment of Hi-C data to restriction fragments'. Regions of the non-redundant sequence not covered by Morex WGS reads and with a length  $\geq 500$  bp were extracted using UNIX command line tools and BEDTools<sup>40</sup> (command 'getfasta'). The extracted sequences were aligned to the NCBI NT database with megablast<sup>27</sup> using a word size of 44 and requiring the high-scoring pairs to have a length of at least 100 bp and an alignment identity  $\geq 80\%$ . We retained only hits whose description in the NCBI NT database did not match the following regular expression (R syntax) representing a list of common and taxonomic names of plant species:

'Hordeum|Triticum|Populus|Aegilops|Avena|Ahnus|A|\squarrosa|Morus|Nelumbo|Brassica|Cucumis|Citrus|Camelina|Fragaria|Lotus|Tarenaya|Spartina|Eucommia|Sorghum|Corylus|Theobroma|Phaseolus|Barley|Trifolium|Elymus|Brachypodium|Beta|vulgaris|Ricinus|Licanium|Phoenix|H|\vulgare|Pyrus|Malus|Prunus|Saccharum|Hypericum|Wheat|Oryza|Zahlhorstia|Secale|Vitis|Quercus'

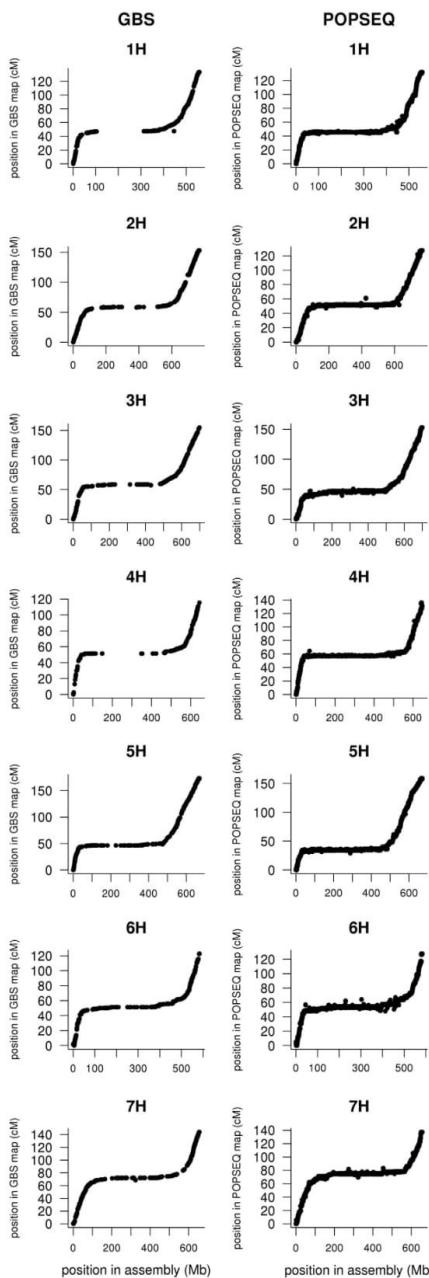
Regions overlapping the BLAST hits passing these filters were cut from the non-redundant sequence with BEDTools<sup>40</sup> (command 'subtract'). Sequences shorter than 500 bp after the removal of contaminant sequences were discarded. This step removed 5 Mb (0.1%) of the assembled sequence.

**Construction of pseudomolecule sequences for chromosome 1H–7H and chrUn.** We constructed pseudomolecules of the seven barley chromosomes by placing the sequence fragments of single BAC assemblies that constitute the non-redundant sequence according to the Hi-C map positions of the BAC overlap clusters these fragments belong to. Sequences not anchored by Hi-C were placed on chrUn ('chromosome unassigned'). The order of clusters was taken from the Hi-C map. BACs within the same cluster were ordered according to the minimum spanning tree of the BAC overlap graph of the cluster and oriented relative to the telomeres using the Hi-C orientation of the cluster if available. The relative order of sequence fragments originating from the same BAC bin (see section 'Construction of the BAC overlap graph') could not be determined so that the placement of sequences within a BAC bin (average size: 70 kb) is arbitrary. ChrUn is composed of (i) sequence fragments originating from BAC overlap clusters not placed in the Hi-C map, or (ii) gene-bearing fragments of BAC sequences and Morex WGS contigs selected in addition to the non-redundant sequence (see section Identification of additional gene-bearing sequences). A gap of 100 N characters was inserted between adjacent sequence fragments. Pseudomolecules of all chromosomes and chrUn were combined into a single FASTA file (Data Citation 42). To accommodate limitations of the Sequence/Alignment Map format (see Usage Notes) split pseudomolecules with a size below 512 Mb were constructed by breaking pseudomolecules arbitrarily at breaks between sequence contigs (Data Citation 43, Data Citation 44). A BED file indicating the placement of BAC sequence fragments, Morex WGS contigs and intercalating gaps in the (split) pseudomolecules is available for download (Data Citation 45, Data Citation 46).

A tabular summary of the positional information incorporated into pseudomolecules is given in Data Citation 41.

#### Masking of residual redundancy

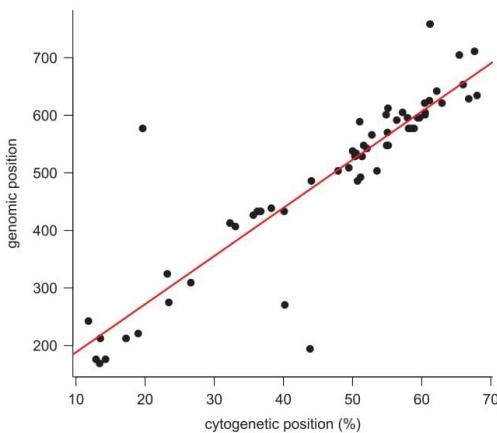
Residual redundancy arising from undetected overlaps between adjacent BACs was detected and masked by aligning the pseudomolecules sequence to itself with megablast<sup>27</sup>. Genomic intervals contained in BLAST hits with a length  $\geq 5$  kb and an identity  $\geq 99.8\%$  were considered as potentially redundant (PR) regions. PR regions were classified to decide which sequence of a redundant pair to mask: (i) PR regions assigned to chromosomal pseudomolecules (as opposed to chrUn), but having BLAST hits only to other chromosomes were considered as originating from chimeric BAC assemblies incorporating unrelated sequences from different chromosomes and masked with Ns; (ii) an analogous procedure was used to find intrachromosomal chimeras based on Hi-C map information; (iii) PR regions on chrUn that had alignments to regions on chromosomal pseudomolecules were masked, (iv) for other PR regions one sequence of a redundant pair was chosen arbitrarily. Positions of masked regions on the (split) pseudomolecules were written into a BED file (Data Citation 47, Data Citation 48). Masking was done with BEDTools<sup>40</sup> (command 'mask') overwriting nucleotides in redundant intervals with N characters. Masked versions of the (split) pseudomolecules are provided as Data Citation 49, Data Citation 50.



**Figure 2. Collinearity between the Hi-C map and two genetic maps.** The positions of genetic markers (x-axis) are plotted against their genetic positions (y-axis) in a GBS map (top row) and a POPSEQ map (bottom row) of the Morex x Barke recombinant inbred lines.

#### POPSEQ genetic map based on pseudomolecule sequence

After the construction of the map-based reference sequence, we constructed an updated high-resolution genetic map of the Morex x Barke population to validate the order of genetic map in the reference



**Figure 3. Collinearity between the Hi-C map and a cytogenetic map of chromosome 3H.** Dots mark the positions of probes in the cytogenetic map (x-axis) and the Hi-C-derived pseudomolecule (y-axis). A linear regression line (red) was fitted with the R function `lm()`. Note that cytogenetic data is not available for distal regions because probes were designed only for non-recombining peri-centromeric regions<sup>61</sup>.

sequence. Raw reads (see section ‘Construction of the updated POPSEQ map of the Morex x Barke mapping population’) were aligned to the barley pseudomolecules with BWA mem (version 0.7.12)<sup>45</sup>. Checking mated mapped paired reads, sorting, conversion to BAM format and marking of duplicate read pairs were done with PicardTools version 2.300 (<http://broadinstitute.github.io/picard/>). Variant detection and genotype calling were performed using GATK Toolkit version 3.3.0 (command ‘HaplotypeCaller’)<sup>57</sup>. A total of five RILs with >3% heterozygous variants were removed. A variant position was removed if more than 10% of all samples were called heterozygous, there were more than 80% missing data, or the minor allele frequency (in the non-missing data) was smaller than 5%. SNP information was aggregated at the contig level to derive consensus genotype blocks with false discovery rate calculated based on the quality of each variant call in the block. High-confidence genotype blocks were obtained based on a Bonferroni correction threshold. Given the fact that the length of crossover tracts is significantly larger than that of non-crossover tracts and non-crossover tracts would enlarge the genetic distance artificially, we only retained high-confidence genotype blocks with more than 1 Mb tract length, which are likely to be derived from crossovers. Representative non-redundant genomic variants of high-confidence genotype blocks were extracted and used for the construction of a high-resolution map through MSTMap<sup>47</sup>. We further anchored all remaining markers to the genetic map by the C program ‘canchor’<sup>5</sup>. The final POPSEQ map consisted of 9,012,742 SNP variants defined on the pseudomolecule sequence Data citation 51).

#### Representation of full-length cDNAs

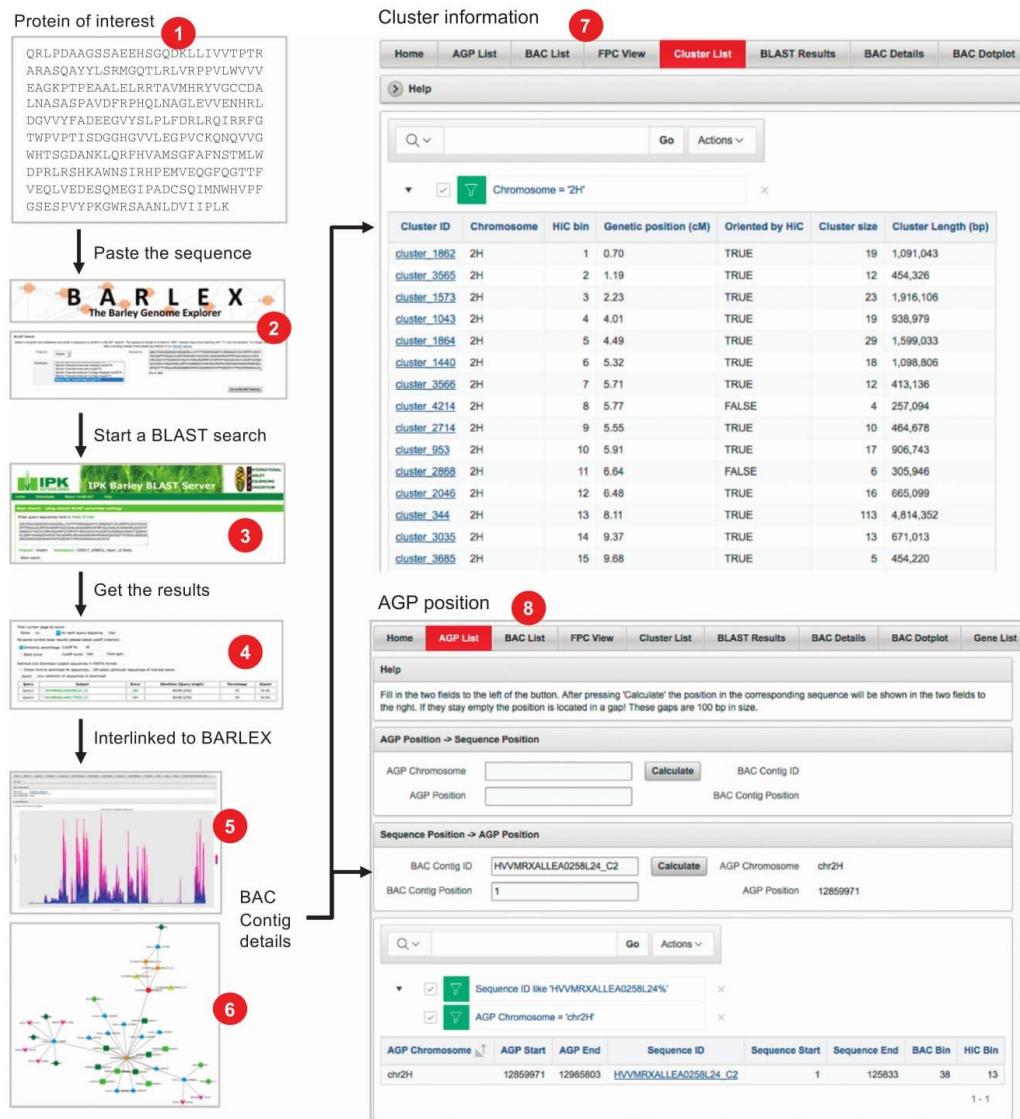
The representation of gene models in the whole-genome genome assembly of barley cv. Morex<sup>4</sup> and in the pseudomolecules was compared by aligning a set of 22,651 publicly available full-length cDNAs<sup>55</sup> to the assemblies using the GMAP splice aligner software<sup>56</sup>. The GMAP alignment output was then filtered. If a full-length cDNA had multiple hits, only the hit with the highest % identity was considered. Hits were further filtered by identity ( $\geq 98\%$ ) and coverage ( $\geq 95\%$ ). This resulted in a set of hits representing genes recovered intact on a single genomic contig/chromosome.

#### Code availability

R and shell source code for the construction of the BAC overlap graph and the Hi-C map is provided as Data Citation 52. Code can be re-used under the terms of the MIT license.

#### Data Records

BAC sequence raw data was submitted to the European Nucleotide Archive (ENA) (Data Citation 1, Data Citation 2, Data Citation 3, Data Citation 4, Data Citation 5, Data Citation 6, Data Citation 7, Data Citation 8, Data Citation 9, Data Citation 10, Data Citation 11, Data Citation 12, Data Citation 13, Data Citation 14, Data Citation 15, Data Citation 16, Data Citation 17, Data Citation 18, Data Citation 19, Data Citation 20, Data Citation 21, Data Citation 22, Data Citation 23, Data Citation 24, Data Citation 25, Data Citation 26, Data Citation 27). BAC assemblies were submitted to ENA or NCBI (Data Citation 28, Data Citation 29). Raw data for POPSEQ (Data Citation 35), GBS (Data Citation 38) and Hi-C mapping (Data Citation 40) were submitted to ENA. Processed datasets are accessible as



**Figure 4.** Accessing sequence and positional information with the barley genome explorer (BARLEX). The barley pseudomolecule data was imported into BARLEX, where it is directly linked to the IPK Barley BLAST server. Users can paste a nucleotide or amino acid sequence (1) into the BARLEX input query form and select reference database such as pseudomolecules sequence, the set of all BAC assemblies or annotated genes (2). The sequence is then transferred to the IPK barley BLAST Server (3). The web page with the BLAST results (4) contains references to BARLEX information pages for different structural units (BAC sequence contigs, BAC, BAC cluster, chromosomal Hi-C map). For example, the pages of BAC sequence contigs visualize the repeat content based on genome-wide  $k$ -mer histograms (5) and are linked to a graph-based visualization (6) of the entire BAC assembly. Summary statistics and positional information of BAC clusters are presented in tables that can be searched, sorted and subsetted using user-defined criteria (7). Users can convert pseudomolecule coordinates (AGP positions) to intervals in the underlying BAC sequence assemblies (8).

Digital Object Identifiers (DOIs) in the Plant Genomics and Phenomics Research Data Repository<sup>58</sup> (Data Citation 30, Data Citation 31, Data Citation 32, Data Citation 33, Data Citation 34, Data Citation 36, Data Citation 37, Data Citation 39, Data Citation 41, Data Citation 42, Data Citation 43, Data Citation 44, Data Citation 45, Data Citation 46, Data Citation 47, Data Citation 48, Data Citation 49, Data Citation 50, Data Citation 51, Data Citation 52). DOIs were registered with e!DAL<sup>59</sup>.

### Technical Validation

#### Collinearity between genetic maps and pseudomolecules

To validate the order of scaffolds in the Hi-C map, we compared the order of genetic marker loci in the Hi-C-derived pseudomolecules to their positions in linkage maps. First, we used genotyping-by-sequencing (GBS)<sup>11,50</sup> to type single-nucleotide polymorphisms (SNPs) segregating in a bi-parental population comprising 2,398 recombinant inbred lines (RILs). A total of 2,637 SNPs were detected by aligning GBS reads and calling variants and genotypes using a previously published pipeline<sup>46</sup>. Second, we reanalysed WGS re-sequencing data of a subset of the same population (POPSEQ data) comprising 90 RILs. Construction of a framework linkage map and insertion of additional markers were performed essentially as described by Chapman *et al.*<sup>44</sup> A dot plot comparison of physical and genetic SNP positions revealed that marker orders were highly collinear between the pseudomolecules and both the GBS and POPSEQ map of the Morex x Barke population (Fig. 2).

#### Collinearity between a cytogenetic map and the pseudomolecule of chromosome 3H

We could not validate the order of BAC overlap clusters in the large peri-centromeric regions because of severely repressed recombination<sup>3,60</sup>. Therefore, we compared the order of probes mapped by fluorescence *in-situ* hybridization to chromosomal locations on chromosome 3H and their corresponding sequences in the pseudomolecule of 3H. Since probes were derived from BAC sequences associated with physical contigs, their position from the reference sequence could be determined from the BAC overlap graph. The comparison showed that the cytogenetic and Hi-C maps were highly collinear in peri-centromeric regions of chromosome 3H (Fig. 3).

#### Representation of full-length cDNAs

To assess the completeness of our assembly, we checked for the presence of high-confidence transcript sequences. The representation of gene models in the whole-genome shotgun assembly of barley cv. Morex<sup>4</sup> and in the map-based reference assembly was compared by aligning a set of 22,651 publicly available full-length cDNAs<sup>55</sup> of barley cv. 'Haruna Nijo'. After aligning and filtering, 18,062 (79.74%) intact full-length cDNAs were found in the pseudomolecules, whereas only 10,496 (46.33%) were recovered in the whole-genome assembly. This increase in the number of correctly represented full-length cDNAs vindicates the effort invested in the map-based assembly. Nevertheless, a significant proportion of genes remain fragmented even in the pseudomolecule assembly (20.26%), and presumably these largely represent difficult to assemble genes that contain e.g., microsatellites, long homopolymer stretches and other difficult features, and/or form part of complex gene families that are difficult to resolve. It is likely that only longer read technologies such as Pacific Biosciences (<http://www.pacb.com>) or Oxford Nanopore (<https://www.nanoporetech.com>) will be able to resolve these more difficult cases. Further results on gene space completeness based on an automated gene annotation of the pseudomolecules, and on the representation of repetitive elements are described elsewhere<sup>42</sup>.

#### Usage Notes

Positional information for BAC sequences, physical contigs and WGS contigs can be accessed via the barley genome explorer BARLEX (Fig. 4). BLAST searches against the barley pseudomolecules can also be carried out in BARLEX. We note that processing BAM files with short read alignments to the full pseudomolecules with commonly used tools such as SAMtools<sup>52</sup> or BEDTools<sup>40</sup> may not work as expected because of restrictions on the chromosome size (512 Mb) for indexing file in Sequence Alignment/Map (SAM) format<sup>52</sup>. To circumvent this issue, we have split the pseudomolecules into two part and provide (i) a FASTA file with split pseudomolecules (Data Citation 44) along with the intact sequences and (ii) a BEDfile to convert between full and split pseudomolecule coordinate (Data Citation 43). Alternatively, the CRAM format (<https://samtools.github.io/hts-specs/CRAMv3.pdf>) may be used instead of the BAM format. We note that the orientation of sequence contigs within individual BACs in the pseudomolecules is arbitrary, thus the order and orientation of sequences in the pseudomolecules is accurate only up to resolution of ~100 kb.

#### References

1. Schulte, D. *et al.* The international barley sequencing consortium--at the threshold of efficient access to the barley genome. *Plant physiology* **149**, 142–147 (2009).
2. Schulte, D. *et al.* BAC library resources for map-based cloning and physical map construction in barley (*Hordeum vulgare* L.). *BMC genomics* **12**, 247 (2011).
3. Ariyadasa, R. *et al.* A sequence-ready physical map of barley anchored genetically by two million single-nucleotide polymorphisms. *Plant physiology* **164**, 412–423 (2014).
4. International Barley Genome Sequencing Consortium. A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**, 711–716 (2012).

5. Mascher, M. *et al.* Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *The Plant Journal* **76**, 718–727 (2013).
6. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
7. Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
8. Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature biotechnology* **30**, 771–776 (2012).
9. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
10. Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature biotechnology* **31**, 1119–1125 (2013).
11. Poland, J. A., Brown, P. J., Sorrells, M. E. & Jannink, J.-L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE* **7**, e32253 (2012).
12. Colmsee, C. *et al.* BARLEX—the Barley Draft Genome Explorer. *Mol Plant* **8**, 964–966 (2015).
13. Munoz-Amatriain, M. *et al.* Sequencing of 15 622 gene-bearing BACs clarifies the gene-dense regions of the barley genome. *Plant Journal* **84**, 216–227 (2015).
14. Pasquarello, M. *et al.* The barley Frost resistance-H2 locus. *Functional & integrative genomics* **14**, 85–100 (2014).
15. Meyer, M., Stenzel, U. & Hofreiter, M. Parallel tagged sequencing on the 454 platform. *Nature protocols* **3**, 267–278 (2008).
16. Steuernagel, B. *et al.* De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley. *BMC genomics* **10**, 547 (2009).
17. Beier, S. *et al.* Multiplex sequencing of bacterial artificial chromosomes for assembling complex plant genomes. *Plant biotechnology journal* **14**, 1511–1522 (2016).
18. Sambrook, J. & Russell, D. W. *Molecular cloning: a laboratory manual*. 3rd edition (Cold Spring-Harbour Laboratory Press, 2001).
19. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome biology* **12**, R18 (2011).
20. Quail, M. A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nature methods* **5**, 1005–1010 (2008).
21. Asan *et al.* Paired-end sequencing of long-range DNA fragments for de novo assembly of large, complex Mammalian genomes by direct intra-molecule ligation. *PLoS ONE* **7**, e46211 (2012).
22. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* **2010**, pdb prot5448 (2010).
23. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome biology* **11**, R119 (2010).
24. Lonardi, S. *et al.* Combinatorial pooling enables selective sequencing of the barley gene space. *PLoS computational biology* **9**, e1003010 (2013).
25. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**, 821–829 (2008).
26. Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC genomics* **16**, 236 (2015).
27. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *Journal of computational biology: a journal of computational molecular cell biology* **7**, 203–214 (2000).
28. Chevreux, B., Wetter, T. & Suhai, S. In *German conference on bioinformatics* (1999); 45–56.
29. Taudien, S. *et al.* Sequencing of BAC pools by different next generation sequencing platforms and strategies. *BMC research notes* **4**, 411 (2011).
30. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
31. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
32. Brenchley, R. *et al.* Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**, 705–710 (2012).
33. Andrews, S. FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> (2010).
34. Leggett, R. M., Ramirez-Gonzalez, R. H., Clavijo, B. J., Waite, D. & Davey, R. P. Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. *Frontiers in genetics* **4**, 288 (2013).
35. Simpson, J. T. *et al.* ABYSS: a parallel assembler for short read sequence data. *Genome research* **19**, 1117–1123 (2009).
36. Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
37. Leggett, R. M., Clavijo, B. J., Cissold, L., Clark, M. D. & Caccamo, M. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics* **30**, 566–568 (2014).
38. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18 (2012).
39. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics* **6**, 31 (2005).
40. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
41. R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, 2015).
42. Mascher, M. *et al.* A chromosome conformation capture ordered sequence of the barley genome. *Nature* doi:10.1038/nature22043 (2017).
43. Cao, H. *et al.* Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *GigaScience* **3**, 1 (2014).
44. Chapman, J. A. *et al.* A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome biology* **16**, 26 (2015).
45. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://arxiv.org/pdf/1303.3997v2.pdf> (2013).
46. Mascher, M., Wu, S., Amand, P. S., Stein, N. & Poland, J. Application of genotyping-by-sequencing on semiconductor sequencing platforms: a comparison of genetic and reference-based marker ordering in barley. *PLoS ONE* **8**, e76925 (2013).
47. Wu, Y., Bhat, P. R., Close, T. J. & Lonardi, S. Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS genetics* **4**, e1000212 (2008).
48. Csardi, G. & Nepusz, T. The igraph software package for complex network research, *InterJournal, Complex Systems* 1695 (2006).
49. Prim, R. C. Shortest connection networks and some generalizations. *Bell system technical journal* **36**, 1389–1401 (1957).
50. Wendler, N. *et al.* Unlocking the secondary gene-pool of barley with next-generation sequencing. *Plant biotechnology journal* **12**, 1122–1131 (2014).

51. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10–12 (2011).
52. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
53. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at <https://arxiv.org/pdf/1207.3907v2.pdf> (2012).
54. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology* **30**, 90–98 (2012).
55. Matsumoto, T. *et al.* Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. *Plant physiology* **156**, 20–28 (2011).
56. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
57. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491–498 (2011).
58. Arend, D. *et al.* PGP repository: a plant phenomics and genomics data publication infrastructure. *Database* **2016**, baw033 (2016).
59. Arend, D. *et al.* eIDAL—a framework to store, share and publish research data. *BMC bioinformatics* **15**, 214 (2014).
60. Künzel, G., Korzun, L. & Meister, A. Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. *Genetics* **154**, 397–412 (2000).
61. Aliyeva-Schnorr, L. *et al.* Cytogenetic mapping with centromeric bacterial artificial chromosomes contigs shows that this recombination-poor region comprises more than half of barley chromosome 3H. *The Plant Journal* **84**, 385–394 (2015).

### Data Citations

1. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB9062 (2016).
2. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB9097 (2016).
3. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB9098 (2016).
4. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB9099 (2016).
5. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB9100 (2016).
6. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB9101 (2016).
7. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB9102 (2016).
8. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB9103 (2016).
9. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB9104 (2016).
10. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB8576 (2016).
11. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB8577 (2016).
12. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB8578 (2016).
13. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB8619 (2016).
14. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB8579 (2016).
15. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB8580 (2016).
16. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB89429 (2016).
17. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB9430 (2016).
18. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB89431 (2016).
19. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB10963 (2016).
20. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB11489 (2016).
21. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB12096 (2016).
22. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB11758 (2016).
23. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB9428 (2016).
24. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB11991 (2016).
25. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB89427 (2016).
26. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB11798 (2016).
27. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB11992 (2016).
28. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB13020 (2016).
29. Muñoz-Amatriain, M. *et al.* NCBI BioProject PRJNA198204 (2015).
30. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/21> (2016).
31. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/28> (2016).
32. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/12> (2016).
33. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/31> (2016).
34. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB13028 (2016).
35. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/33> (2016).
36. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/22> (2016).
37. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/30> (2016).
38. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB14130 (2016).
39. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/29> (2016).
40. International Barley Genome Sequencing Consortium. *European Nucleotide Archive* PRJEB14169 (2016).
41. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/20> (2016).
42. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/34> (2016).
43. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/27> (2016).
44. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/36> (2016).
45. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/23> (2016).
46. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/24> (2016).
47. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/25> (2016).
48. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/26> (2016).
49. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/35> (2016).
50. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/37> (2016).
51. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/17> (2016).
52. International Barley Genome Sequencing Consortium. IPK Gatersleben <http://dx.doi.org/10.5447/IPK/2016/19> (2016).

### Acknowledgements

This work was carried out under the auspices of the International Barley Genome Sequencing Consortium and supported from the following funding sources: German Ministry of Education and Research (BMBF) grant 0314000 'BARLEX' and 0315954 'TRITEX' to M.P., U.S. and N.S and 031A536

'de.NBI' to U.S. Leibniz Association grant ('Pakt f. Forschung und Innovation') 'sequencing barley chromosome 3H' to N.S. and U.S.; Scottish Government/UK Biotechnology and Biological Sciences Research Council (BBSRC) grant BB/100663X/1 to R.W, P.E.H., J.R.; BBSRC grants BB/I008357/1 to M. D.C., M.C. and BB/I008071/1 to P.K.; of Finland grant 266430 and a BioNano grant to A.H.S.; Carlsberg Foundation grant nr. 2012\_01\_0461 to the Carlsberg Research Laboratory; Grain Research and Development Corporation (GRDC) grant DAW00233 to C.L. and P.L.; Department of Agricultural and Food, Government of Western Australia grant 681 to C.L.; National Natural Science Foundation of China (NSFC) grant 31129005 to C.L. and G.Zhang; NSFC grant 31330055 to G.Zhang.; Czech Ministry of Education, Youth and Sports grant LO1204 to J.D.; National Science Foundation grant DBI 0321756 'Coupling EST and Bacterial Artificial Chromosome Resources to Access the Barley Genome' to T.J.C. and S.L.; United States Department of Agriculture (USDA), Agriculture and Food Research Initiative Plant Genome, Genetics and Breeding Program of USDA-CSREES-NIFA grant 2009-65300-05645 'Advancing the Barley Genome' and 2011-68002-30029 'TriticaceaeCAP' to T.J.C., S.L. and G.J.M.; United States National Science Foundation (NSF)-ABI grant DBI-1062301 to T.J.C. and S.L.; University of California grant CA-R-BPS-5306-H to T.J.C and S.L.; National Science Foundation grant DBI 0321756 'Algorithms for Genome Assembly of Ultra-deep Sequencing Data' to S.L. Next-generation sequencing and library construction was delivered via the BBSRC National Capability in Genomics (BB/J010375/1) at Earlham Institute (formerly The Genome Analysis Centre) by members of the Platforms and Pipelines group and BBSRC Institute Strategic Programme funding for Bioinformatics (BB/J004669/1) to M.D.C., S.A. and M.C. We gratefully acknowledge: (1) the excellent technical assistance by Susanne König, Manuela Knauf, Uli Beier, Anne Kusserow, Katrin Trnka, Ines Walde, Sandra Drieslein, Cynthia Voss; (2) Doreen Stengel, Anne Fiebig, Thomas Münch, Danuta Schüler and Daniel Arend and Matthias Lange for sequence raw data management and data submission to EMBL/ENA and registration of DOIs; (3) Dr Hélène Berges, Arnaud Belloc and Sonia Vautrin (CNRGV) for management and distribution of barley BAC libraries; (4) Andreas Graner and David Marshall for scientific discussions.

#### Author Contributions

*BAC sequencing and assembly (1H, 3H, 4H):* S.B., A.Himmelbach, S.T., M.F., M.G., M.M., U.S. (co-leader), M.P. (co-leader), N.S. (leader); *BAC sequencing and assembly (2H, unassigned):* D.S., D.H., S. A. (co-leader), M.D.C. (co-leader), M.C. (co-leader), R.W. (leader); *BAC sequencing and assembly (5H, 7H):* X.Z., R.A.B., Q.Z., C.T., J.K.M., B.C., G.Zhou, F.D., Y.H., S.Y., S.Cao, S.Wang, X.L., M.I.B., P.L., G.Zhang (co-leader), C.Li (leader); *BAC sequencing and assembly (6H):* S.B., S.Wang, C.Lin, H.L., U.S., M. H. (co-leader), I.B. (leader); *BAC sequencing (gene-bearing):* M.M.-A., R.O., S.Wanamaker, S.L. (co-leader), T.J.C. (leader); *Optical mapping:* A.Hastie, H.S., J.T., H.S., J.V., S.Chan, M.M., N.S., J.D., A.H.S. (leader); *Chromosome conformation capture:* A.Himmelbach, S.G., M.M. (co-leader), N.S. (leader); *Pseudomolecule construction:* M.M. (leader), S.B., C.C., D.B., T.S., P.K., N.S., U.S. (co-leader); *Validation:* L.L., M.B., L.A.-S., A.Houben, J.A.P., N.S., G.J.M., M.M. (leader). All authors read and commented on the manuscript.

#### Additional information

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Beier, S. *et al.* Construction of a map-based reference genome sequence for barley, *Hordeum vulgare* L. *Sci. Data* 4:170044 doi: 10.1038/sdata.2017.44 (2017).

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0>

Metadata associated with this Data Descriptor is available at <http://www.nature.com/sdata/> and is released under the CC0 waiver to maximize reuse.

© The Author(s) 2017

Sebastian Beier<sup>1,\*</sup>, Axel Himmelbach<sup>1,\*</sup>, Christian Colmsee<sup>1</sup>, Xiao-Qi Zhang<sup>2</sup>, Roberto A. Barrero<sup>3</sup>, Qisen Zhang<sup>4</sup>, Lin Li<sup>5</sup>, Micha Bayer<sup>6</sup>, Daniel Bolser<sup>7</sup>, Stefan Taudien<sup>8</sup>, Marco Groth<sup>8</sup>, Marius Felder<sup>8</sup>, Alex Hastie<sup>9</sup>, Hana Šimková<sup>10</sup>, Helena Staňková<sup>10</sup>, Jan Vrána<sup>10</sup>, Saki Chan<sup>9</sup>, María Muñoz-Amatriain<sup>11</sup>, Rachid Ounit<sup>12</sup>, Steve Wanamaker<sup>11</sup>, Thomas Schmutz<sup>1</sup>, Lala Aliyeva-Schnorr<sup>1</sup>, Stefano Grasso<sup>13</sup>, Jaakko Tanskanen<sup>14</sup>, Dharanya Sampath<sup>15</sup>, Darren Heavens<sup>15</sup>, Sujie Cao<sup>16</sup>, Brett Chapman<sup>3</sup>, Fei Dai<sup>17</sup>, Yong Han<sup>17</sup>, Hua Li<sup>16</sup>, Xuan Li<sup>16</sup>, Chongyun Lin<sup>16</sup>, John K. McCooke<sup>3</sup>, Cong Tan<sup>3</sup>, Songbo Wang<sup>16</sup>, Shuya Yin<sup>17</sup>, Gaofeng Zhou<sup>2</sup>, Jesse A. Poland<sup>18</sup>, Matthew I. Bellgard<sup>3</sup>, Andreas Houben<sup>1</sup>, Jaroslav Doležel<sup>10</sup>, Sarah Ayling<sup>15</sup>, Stefano Lonardi<sup>12</sup>, Peter Langridge<sup>19</sup>, Gary J. Muehlbauer<sup>5,20</sup>, Paul Kersey<sup>7</sup>, Matthew D. Clark<sup>15,21</sup>, Mario Caccamo<sup>15,22</sup>, Alan H. Schulman<sup>14</sup>, Matthias Platzer<sup>8</sup>, Timothy J. Close<sup>11</sup>, Mats Hansson<sup>23</sup>, Guoping Zhang<sup>17</sup>, Ilka Braumann<sup>24</sup>, Chengdao Li<sup>2,25,26</sup>, Robbie Waugh<sup>6,27</sup>, Uwe Scholz<sup>1</sup>, Nils Stein<sup>1,28</sup> & Martin Mascher<sup>1,29</sup>

<sup>1</sup>Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, 06466 Seeland, Germany. <sup>2</sup>School of Veterinary and Life Sciences, Murdoch University, Murdoch, Western Australia 6150, Australia. <sup>3</sup>Centre for Comparative Genomics, Murdoch University, Murdoch, Western Australia 6150, Australia. <sup>4</sup>Australian Export Grains Innovation Centre, South Perth, Western Australia 6151, Australia. <sup>5</sup>Department of Agronomy and Plant Genetics, University of Minnesota, St Paul, Minnesota 55108, USA. <sup>6</sup>The James Hutton Institute, Dundee DD2 5DA, UK. <sup>7</sup>European Molecular Biology Laboratory—The European Bioinformatics Institute, Hinxton CB10 1SD, UK. <sup>8</sup>Leibniz Institute on Aging—Fritz Lipmann Institute (FLI), 07745 Jena, Germany. <sup>9</sup>BioNano Genomics Inc., San Diego, California 92121, USA. <sup>10</sup>Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, 78371 Olomouc, Czech Republic. <sup>11</sup>Department of Botany & Plant Sciences, University of California, Riverside, Riverside, California 92521, USA. <sup>12</sup>Department of Computer Science and Engineering, University of California, Riverside, Riverside, California 92521, USA. <sup>13</sup>Department of Agricultural and Environmental Sciences, University of Udine, 33100 Udine, Italy. <sup>14</sup>Green Technology, Natural Resources Institute (Luke), Viikki Plant Science Centre, and Institute of Biotechnology, University of Helsinki, 00014 Helsinki, Finland. <sup>15</sup>Earlham Institute, Norwich NR4 7UH, UK. <sup>16</sup>BGI-Shenzhen, Shenzhen 518083, China. <sup>17</sup>College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, China. <sup>18</sup>Kansas State University, Wheat Genetics Resource Center, Department of Plant Pathology and Department of Agronomy, Manhattan, Kansas 66506, USA. <sup>19</sup>School of Agriculture, University of Adelaide, Urrbrae, South Australia 5064, Australia. <sup>20</sup>Department of Plant and Microbial Biology, University of Minnesota, St Paul, Minnesota 55108, USA. <sup>21</sup>School of Environmental Sciences, University of East Anglia, Norwich NR4 7UH, UK. <sup>22</sup>National Institute of Agricultural Botany, Cambridge CB3 0LE, UK. <sup>23</sup>Department of Biology, Lund University, 22362 Lund, Sweden. <sup>24</sup>Carlsberg Research Laboratory, 1799 Copenhagen, Denmark. <sup>25</sup>Department of Agriculture and Food, Government of Western Australia, South Perth, Western Australia 6150, Australia. <sup>26</sup>Hubei Collaborative Innovation Centre for Grain Industry, Yangtze University, Jingzhou, Hubei 434025, China. <sup>27</sup>School of Life Sciences, University of Dundee, Dundee DD2 5DA, UK. <sup>28</sup>School of Plant Biology, University of Western Australia, Crawley 6009, Australia. <sup>29</sup>German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103 Leipzig, Germany. \*These authors contributed equally to this work.

## A critical comparison of technologies for a plant genome sequencing project

Pirita Paajanen<sup>1\*</sup>, George Kettleborough<sup>1\*</sup>, Elena López-Girona<sup>2</sup>, Michael Giolai<sup>1</sup>, Darren Heavens<sup>1</sup>, David Baker<sup>1</sup>,  
Ashleigh Lister<sup>1</sup>, Gail Wilde<sup>2</sup>, Ingo Hein<sup>2</sup>, Iain Macaulay<sup>1</sup>, Glenn J. Bryan<sup>2</sup>, and Matthew D. Clark<sup>1+</sup>

\*contributed equally to this work

+corresponding author: matt.clark@earlham.ac.uk

<sup>1</sup>Earlham Institute, Norwich, UK; <sup>2</sup>The James Hutton Institute, Invergowrie, Dundee, UK.

A high quality genome sequence of your model organism is an essential starting point for many studies. Old clone based methods are slow and expensive, whereas faster, cheaper short read only assemblies can be incomplete and highly fragmented, which minimises their usefulness. The last few years have seen the introduction of many new technologies for genome assembly. These new technologies and new algorithms are typically benchmarked on microbial genomes or, if they scale appropriately, human. However, plant genomes can be much more repetitive and larger than human, and plant biology makes obtaining high quality DNA free from contaminants difficult. Reflecting their challenging nature we observe that plant genome assembly statistics are typically poorer than for vertebrates. Here we compare Illumina short read, PacBio long read, 10x Genomics linked reads, Dovetail Hi-C and BioNano Genomics optical maps, singly and combined, in producing high quality long range genome assemblies of the potato species *S. verrucosum*. We benchmark the assemblies for completeness and accuracy, as well as DNA, compute requirements and sequencing costs. We expect our results will be helpful to other genome projects, and that these datasets will be used in benchmarking by assembly algorithm developers.

[Supplemental material is available for this article.]

*Keywords:* assembly, long reads, short reads, optical mapping, Pacific Biosciences, PacBio, 10x Genomics.

Developments in high-throughput sequencing have revolutionised genetics and genomics, with lower costs leading to an explosion in genome sequencing project size [1]. This diversity of sequencing and assembly methods, coupled to the activities of many laboratories, are generating multiple assemblies. These need to be compared to ensure that optimal approaches have been used.

The existence of very high quality references [4, 14] has made the human genome popular for demonstrating new sequencing technologies and assembly algorithms. The human genome has now been sequenced and assembled using various technologies including Sanger, 454, IonTorrent, Illumina, Pacific Biosciences (PacBio), 10x Genomics and even nanopore sequencing technologies [25, 31, 6, 37, 46, 17]. Hybrid approaches have also been used which combine complementary technologies, for example PacBio and BioNano [33].

However, the human genome is not representative of all eukaryotic genomes; plant genomes in particular are typically more repetitive (including multi-kilobase long retrotransposon elements as well as even longer regions comprising of “nested” transposon insertions). Plant biology also poses challenges for the isolation of high quality high molecular weight DNA, due to strong cell walls, co-purifying polysaccharides, and secondary metabolites which inhibit enzymes or directly damage DNA [13]. Thus technologies that work well on vertebrate genomes may not work well for plants [18]. For these reasons slow and expensive clone based minimal tiling path sequencing approaches have persisted in plants [9, 30] long after faster, cheaper short read whole genome assemblies were first demonstrated for vertebrate genomes [26]. Plant genomes also vary hugely in size, from 61 Mbp (*Genlisea tuberosa*, a member of the bladderwort family [12]) to 150 Gbp (*Paris japonica*, a relative of lilies [32]), it is still nontrivial to design a *de novo* assembly project which involves an ensemble of technologies. Each platform comes with its own input requirements, computational requirements, quality of output and, of course, labour and

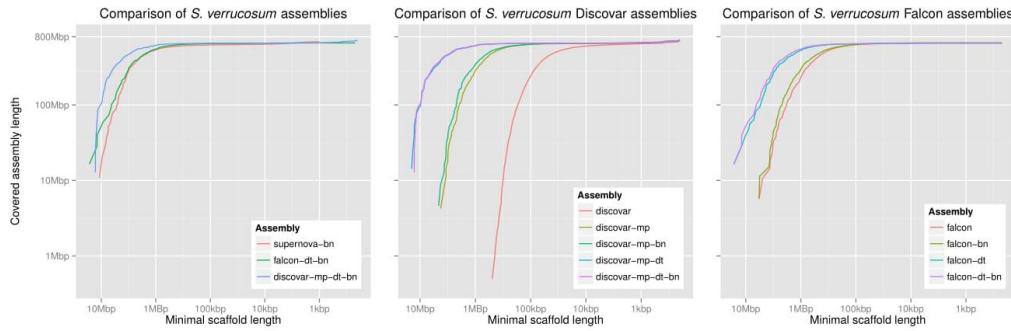
materials costs.

In this paper we compare several practical *de novo* assembly projects of a self-compatible, diploid Mexican wild potato species *Solanum verrucosum* using Illumina, PacBio, BioNano, 10x Genomics and Dovetail technologies. We see how plant biology poses some additional challenges for the isolation of high quality high molecular weight DNA. The genome size of about 722 Mbp is suitable for testing many different technologies whilst keeping the costs reasonable. Using the genome of *S. verrucosum* we are able to demonstrate that repeat content does limit the contiguity of the assembly by comparing the assembly to BAC sequences, and find out which technology can resolve large repeats. As its relative *S. tuberosum* has been assembled [34], we can use synteny to analyse long-range scaffolding accuracy. We find that the long-range scaffolding can cause chimeric scaffolds for some assemblies, but not others.

Our results can be used as guidance for further sequencing assembly projects and provide a basis for comparative genome studies, as each sequencing strategy and assembly method has its own biases.

## Results

The results of this study are presented in two parts. First we compare short read (Illumina) with long read (PacBio) based assemblies. In the second part we take the best performing Illumina and PacBio assemblies, and then add longer-range scaffolding data from newer technologies, namely *in vitro* Hi-C (Dovetail), optical mapping (BioNano Genomics), and read clouds (10x Genomics Chromium) technologies. Validating the assemblies for sequence and scaffolding accuracy we find strengths and weaknesses, and that methods differ hugely in their DNA, time, computational requirements and cost.



**Figure 1:** Comparison of contig/scaffold lengths and total assembly sizes of the various *S. verrucosum* assemblies.

Budget constraints do play a large part in the choice of technologies to be adopted for any genome project. Assembly and scaffolding methods are often effectively the choice of sequencing method, but the properties of the genome will also affect the results. Heterozygosity, in particular, complicates the assembly process and if individual haplotypes are desired this places limitations on which strategies can be used. The careful choice of organism where possible, such as a highly inbred plant or doubled haploid, can remove the problems caused by structural heterozygosity. This approach was adopted for the potato DM reference, whereby a completely homozygous “doubled monoploid” was used instead of a highly heterozygous potato genotype. The original heterozygous diploid RH genotype selected for sequencing proved difficult to assemble due to the extremely high level of haplotype diversity.

### Contig assembly and scaffolding

The first stage of an assembly is to piece together reads to form long contiguous sequences, or *contigs* for short. These contigs can be ordered and oriented using longer-range information such as jumping/mate pair libraries. Throughout this paper we will refer to different contig assemblies that have been scaffolded. We use a naming convention which shows all of the steps used to construct the assembly. Each assembly name contains the steps used in order, separated by a hyphen. For example, the *discoverar-mp-dt-bn* assembly is the *discoverar* contig assembly scaffolded first with mate-pairs, then Dovetail and finally BioNano.

### Illumina contig assembly

Two libraries were constructed for Illumina assembly. The first is a PCR-free library with insert size 500 bp ( $\pm 40\%$ ) which was sequenced with 250 bp paired-end reads on a single Illumina HiSeq run. We refer to this below as the DISCOVAR library. The coverage of the library was  $120\times$ . The second library is a PCR-free “Tight and Long Library” (TALL) with insert size 650 bp ( $\pm 20\%$ ) sequenced with 100 bp and 150 bp paired-end reads. The coverage of this library was  $135\times$ .

We analysed the TALL library reads with preqc, part of the SGA assembler [40], and it gave a genome size estimate at 722 Mbp, which agrees well with the 727 Mbp size of the potato genome assembly [34].

The TALL library was assembled with ABYSS [41] ( $k$ -mer size 113) and the DISCOVAR library using DISCOVAR *de novo* [45] producing contig assemblies *discoverar* and *abyss*, respectively. The results for these two Illumina assemblies are remarkably similar and shown in Table 1. These assemblies are more contiguous than the equivalent contig assemblies of the *S. tuberosum* genome [34].

### Illumina scaffolding

A Nextera long mate-pair (LMP) library was made with insert size 10 000 bp ( $\pm 20\%$ ) and sequenced on two lanes of an Illumina MiSeq with fragment size 500 bp and 300 bp reads. The total coverage of the LMP library was  $15\times$ . We scaffolded both the *discoverar* and *abyss* assemblies separately using Soapdenovo2 [27] producing *discoverar-mp* and *abyss-mp*, respectively. The contiguity of both was increased significantly as shown in Table 1. Here the *discoverar-mp* scaffolds were slightly better so we used this assembly to take forward for longer range scaffolding with other data types.

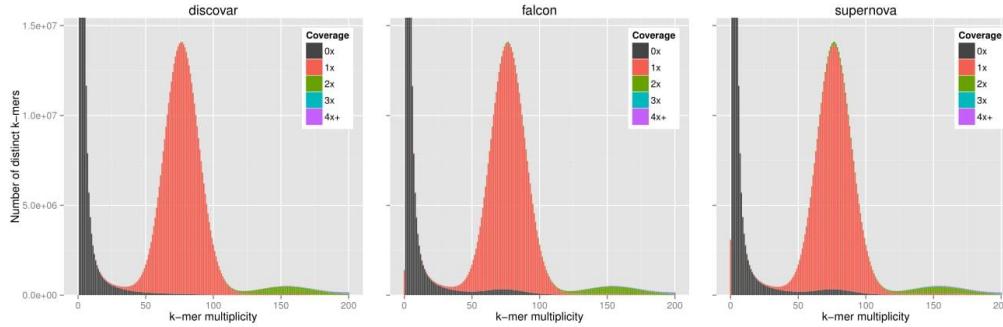
### PacBio assembly

A PacBio library with fragment lengths of at least 20 kbp was made giving a total coverage of  $50\times$ .

We conducted three long read assemblies on the same data using HGAP3 [7], part of smartanalysis (version 2.3.0p5), Canu [19] (version 1.0), and Falcon [8] (version 0.3.0) producing the *hgap*, *canu* and *falcon* assemblies, respectively. The assembly statistics for each is shown in Table 1.

Assembly	Number of contigs	N50 (kbp)	Max length (kbp)	Total length (Mbp)
abyss	33 146	75	642	702
abyss-mp	21 376	331	2 288	712
discoverar	25 216	77	498	646
discoverar-mp	8 074	858	4 266	665
hgap	5 446	585	4 876	716
canu	8 138	290	4 701	722
falcon	2 442	712	5 738	659

**Table 1:** Assembly statistics of Illumina and PacBio assemblies, with a minimum contig/scaffold size of 1 kbp. *abyss* uses the TALL library, *discoverar* uses the DISCOVAR library, and *hgap*, *canu* and *falcon* use the PacBio library.



**Figure 2:** KAT spectra-en plots comparing three *S. verrucosum* contig assemblies. The heights of the bars indicate how many *k*-mers of each multiplicity appear in the raw DISCOVAR reads. The colours indicate how many times those *k*-mers appear in the respective assemblies with black being zero times and red being one time. A coloured bar at zero multiplicity indicates *k*-mers appearing in the assembly which do not appear in the reads. The FALCON assembly has been polished with the Illumina reads using Pilon to reduce the effect of using a different sequencing platform.

The Canu assembly was made with reads that were first error-corrected by the HGAP3 pipeline because the first attempt using raw reads resulted in an excessive amounts of small scaffolds and a genome size more than 50 % longer than expected.

The canu and hgap assemblies contain considerably more content than all other assemblies. The falcon assembly has the highest N50, and is closest to the estimated genome length. FALCON also produced 9.9 Mbp of alternate contigs, likely from residual heterozygosity. We chose the falcon assembly to take forward to hybrid scaffolding. We first polished it using Quiver as part of SMRTanalysis (version 2.3.0p5).

### Longer-range scaffolding

To achieve higher contiguity, newer technologies have been developed to complement the previous methods and, in some cases, each other. In this section we investigate using longer range scaffolding methods to increase the contiguity of the Illumina *discover*-mp assembly and the *falcon* PacBio assembly. We also investigate the 10x Genomics Chromium platform, an integrated solution which can be used to generate short Illumina reads with long-range positional information.

### Dovetail

Dovetail Genomics provides a specialised library preparation method called Chicago and an assembly service using a custom scaffolder called HiRise. The Chicago library preparation technique is based on the Hi-C method, producing deliberately “chimeric” inserts linking DNA fragments from distant parts of the original molecule [35]. This is followed by standard Illumina paired-end sequencing of the inserts. Since the separation of the original fragments follows a well-modelled insert size distribution, the scaffolder is able to join contigs to form scaffolds spanning large distances, even up to 500 kbp [35].

Dovetail Genomics, LLC (Santa Cruz, CA, USA) received fresh leaf material from us from which they constructed a Chicago library. This was sequenced at Earlham Institute using Illumina 250 bp paired-end reads. The total read coverage of the Chicago library was 105  $\times$ . Dovetail used their HiRise software to further scaffold the *discover*-mp assembly, increasing the N50 from 825 kbp to 4700 kbp, and the *falcon* assembly, increasing

the N50 from 710 kbp to 2800 kbp. These assemblies are called *discover*-mp-dt and *falcon*-dt, respectively.

### BioNano

The BioNano Genomics Irys platform constructs a physical map using very large DNA fragments digested at known sequence motifs with a specific nicking enzyme, to which a polymerase adds a fluorescent nucleotide. The molecules are scanned, and the distance between nicks generates a fingerprint of each molecule which is then used to build a whole genome physical map. Sequence-based scaffolds or contigs can be integrated by performing the same digestion *in silico* then ordering and orienting the contigs according to the physical map [16].

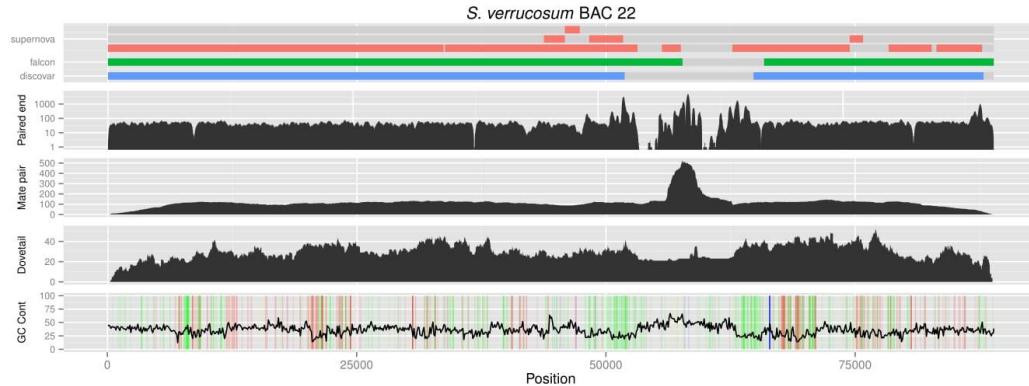
We collected BioNano data from 16 runs by repeatedly running the same chip. After filtering fragments less than 100 kbp, the yield varied from 0.8 Gb to 25.8 Gb, with the earlier runs yielding more whereas the molecule N50 was higher in later runs (ranging from 135 kbp to 240 kbp). The total yield of BioNano data was 252 Gbp which is roughly equivalent to 350  $\times$  coverage.

We performed hybrid scaffolding on the *discover*-mp and *falcon* assemblies. The *in silico* digest suggested a label density of 8.1/100 kbp for *discover*-mp and 8.4/100 kbp for *falcon* whilst the actual observed density was only 6.8/100 kbp. We used the BioNano pipeline (v2.0) to scaffold *discover*-mp, increasing the N50 from 825 kbp to 1260 kbp, and *falcon*, increasing the N50 from 710 kbp to 1500 kbp. These assemblies are called *discover*-mp-bn and *falcon*-bn, respectively.

### 10x Genomics

10x Genomics provides an integrated microfluidics based platform for generating linked reads (a cloud of non-contiguous reads with the same barcode from the same original DNA molecule) and customised software for their analysis [46]. Large fragments of genomic DNA are combined with individually barcoded gel beads into micelles in which library fragments are constructed and then sequenced as a standard Illumina library. Using the barcodes the reads from the same gel bead can be grouped together.

Unlike the previous two longer-range scaffolding approaches, the 10x Genomics platform constructs a new paired-end library



**Figure 3:** A difficult region of the genome which is contiguously assembled with a PacBio BAC but in none of our whole genome assemblies. The region was correctly scaffolded by Dovetail. The figure shows various alignments and information with respect to the BAC assembly. The top track shows the contigs which appear in the *discover*, *falcon* and *supernova* assemblies. The paired-end track shows read coverage of the DISCOVAR paired-end library. The mate-pair and Dovetail tracks show physical/fragment coverage of the mate-pair and Dovetail libraries, respectively. The bottom track shows GC content of the sequence as well as homopolymers sequences of at least 5 bp where A, C, G, and T are coloured as red, blue, yellow, and green, respectively.

which can be sequenced and then assembled into large scaffolds by one assembly program: SUPERNOVA.

A 10x Genomics Chromium library was made according to manufacturer's instructions and a lane of Illumina HiSeq 250 bp paired-end reads were generated with a coverage of about 92 $\times$ . SUPERNOVA (version 1.1.1) produced the *supernova* assembly with length 641 Mbp and a scaffold N50 of 2.33 Mbp. Trimming reads back to 150 bp or reducing sequencing depth to 56 $\times$ , which are the read length and depth recommended by 10x Genomics, generated very similar results (see Supplemental Section 2.3).

#### Hybrid scaffolding

It is possible to iteratively combine these longer-range scaffolding approaches. We tested several hybrid approaches using the *discover*, *falcon* and *supernova* assemblies. For example the *discover-mp* assembly was scaffolded using Dovetail and then BioNano producing *discover-mp-dt-bn* with an N50 of 7.0 Mbp, the highest contiguity of any assembly reported here. The *falcon* assembly when scaffolded with both produced scaffolds with an N50 of only 3.09 Mbp, lower than with BioNano alone. Finally we scaffolded the *supernova* assembly with BioNano producing *supernova-bn* which increased the N50 from 2.33 Mbp to 2.85 Mbp.

We also used long reads from PacBio to scaffold and to perform "gapfilling" on the assemblies, replacing regions of unknown sequence (N stretches) with a PacBio consensus sequence. This also presents an opportunity to use lower coverage PacBio data to improve an Illumina assembly, which may be more cost effective than a *de novo* assembly using PacBio. PBJelly (version 15.2.20) [11] was used to perform gapfilling using only 10 SMRTcells of PacBio data (8 $\times$  depth). The SUPERNOVA assembly increased in size from 641 Mbp to 671 Mbp, and N50 from 2.33 Mbp to 2.64 Mbp, and the amount of Ns present reduced from 7.58 % to 5.14 %. The *discover-mp-dt* assembly increased in size from 656 Mbp to 680 Mbp and N50 from 4.69 Mbp to 4.87 Mbp, with Ns reduced from 3.03 % to 1.28 %. However, how gaps and percentage Ns are generated differs between assembly methods (see

Discussion).

#### Assembly evaluation

Achieving a genome assembly with high levels of contiguity is potentially useless if it does not faithfully represent the original genome sequence. We assessed errors in assemblies by comparison to the raw data used to make the assemblies, as well as measuring gene content, local accuracy (BAC assemblies), and long-range synteny with the close relative *Solanum tuberosum*.

#### K-mer content

Analysis of the  $k$ -mer content of an assembly gives a broad overview of how well the assembly represents the underlying genome. We used the PCR-free Illumina DISCOVAR library as our reference for the  $k$ -mer content of the genome. Due to the high accuracy of the reads we expect the  $k$ -mer spectra for a library to form a number of distributions which correspond to read errors, non-repetitive, and repetitive content in the genome. These distributions can be seen by observing only the shapes and ignoring the colours in Figure 2. The reader is referred to the KAT documentation for further details [29].

In Figure 2 we compare the  $k$ -mer contents of the three contig assemblies—*discover*, *falcon*, and *supernova*—to the DISCOVAR library. To minimise the effects of the differences between Illumina and PacBio sequencing error profiles the *falcon* assembly has been polished with the Illumina reads using Pilon [44] (see Supplemental Figure S3.1 for the unpolished plot).

The small red bar on the origin in some plots shows content which appears in the assembly but not in the Illumina reads. The *discover* assembly is very faithful to the content in the library. The black area denotes sequences in the reads but not in the assembly: those clustering at the origin are predicted sequence errors in the reads, the small amount between 50–100 on the x-axis is sequence missing from the assembly. The dominant red peak (1 $\times$ , around multiplicity 77), which is the vast majority of all assemblies here, contains content in the Illumina reads which appears

once in the assembly (homozygous sample). Green areas on top of the main peak in FALCON and SUPERNOVA represents possible duplications in the assembly, whereas the green ( $2 \times$ ) small peak to the right of the main peak is probably true duplicates—as these sequences are present twice in the assembly and at twice the expected read counts. At the main peak ( $k$ -mer multiplicity 77), the amount of potentially duplicated content in the assemblies is 0.66 % in *falcon*, 1.3 % in *supernova*, and 0.15 % in *discover*.

### Gene content

We assessed the gene content of the three most contiguous assemblies—*discover*-*mp*-*dt*-*bn*, *falcon*-*dt*-*bn*, and *supernova*-*bn*—using two datasets. The first is with BUSCO and its *embryophyta*\_*odb9* (plants) dataset [39] and the second is all the predicted transcript sequences from the *S. tuberosum* genome [34].

We found that each of the three assemblies shows at least 95 % of BUSCOs as complete, with only 2–3 % missing. The difference is small but the *discover*-*mp*-*dt*-*bn* assembly is the most complete while *supernova*-*bn* is the worst performing. The results are shown in Figure 4.

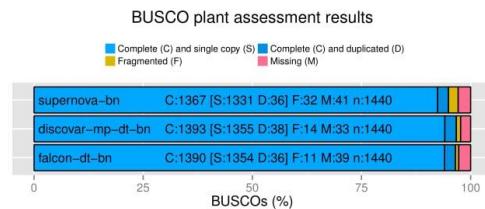
We aligned the *S. tuberosum* representative transcript sequences to each genome assembly using BLAST [2] and then measured how much of each transcript sequence was represented in the assembly according to various minimum percentage identity cutoffs. As expected when comparing between species, as the threshold approaches 100 % nucleotide identity the transcript completeness drops closer to zero. Using a threshold between 96–98 % we find the median transcript completeness is highest in *discover*-*mp*-*dt*-*bn*, followed by *falcon*-*dt*-*bn*, and then *supernova*-*bn*. However, the difference between the assemblies is small, Figure 5 shows a box and whisker plot of completeness of the representative transcript sequences.

### Local accuracy

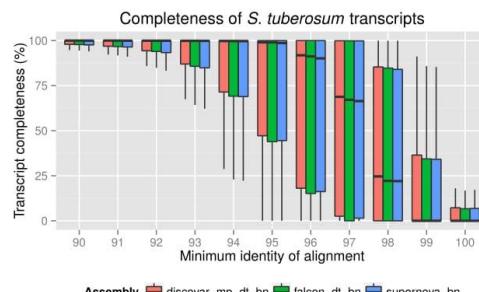
As BACs are easier to assemble due to smaller size and a much more limited amount of repetitive DNA content than a whole genome, we assessed the performance of our three assemblies at a local scale using BAC assemblies. We randomly selected, sequenced, and assembled 96 BAC clones from *S. verrucosum* BAC library. We chose 20 high-quality BAC assemblies (single scaffolds/contigs with Illumina or PacBio) to measure the accuracy of the whole genome assemblies.

We used dnadiff [20] to compare the BAC sequences to the *supernova*-*bn*, *discover*-*mp*-*dt*-*bn*, and *falcon*-*dt*-*bn* assemblies finding sequence identities of 99.40 %, 99.97 %, and 99.87 %, respectively. As in the previous section, the *discover*-*mp*-*dt*-*bn* assembly shows the highest accuracy, with *supernova*-*bn* the lowest, though the differences are small.

To illustrate the performance of the different technologies sequencing different genomic features we mapped whole genome reads and assemblies to single BACs as shown in Figure 3. None of our three whole genome assemblies are able to reconstruct BAC 22; each breaking at a large (more than 12 kbp) repeat. The DISCOVAR library (paired-end), mate-pair library and Dovetail library were each mapped and only reads mapping to a high quality and exhibiting up to one mismatch are shown in the figure. The mapping reveals several areas of high repetition, for example the arms and middle of a retrotransposon, and there are areas lacking coverage completely which suggests a sequence which is difficult



**Figure 4:** BUSCO analysis of *supernova*-*bn*, *discover*-*mp*-*dt*-*bn*, and *falcon*-*dt*-*bn* using the plant gene dataset.



**Figure 5:** Box and whisker plot showing completeness of the *S. tuberosum* transcripts in *supernova*-*bn*, *discover*-*mp*-*dt*-*bn*, and *falcon*-*dt*-*bn* with various levels of minimum percentage identity.

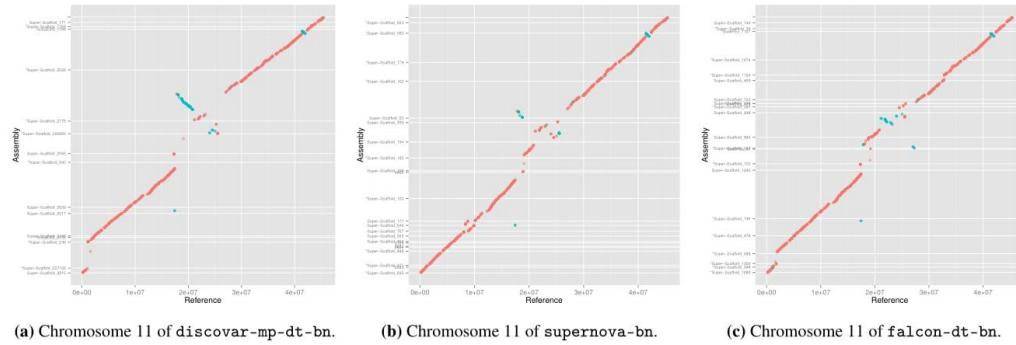
for our Illumina sequence data to resolve. We also see drops in coverage at some sites with high concentrations of homopolymers, as marked by coloured lines in the GC content, for example an A rich region of ~7 kbp. Interestingly the repeat arms are also rich in homopolymers.

We note that the *discover*-*mp*-*dt*-*bn* assembly leaves the largest gap around the repeat. The *falcon* assembly was able to completely cover an area with no mapping paired-end Illumina reads which explains some of extra  $k$ -mer content in Figure 2 noted earlier in this assembly. The *supernova*-*bn* assembly was able to reconstruct more of the difficult region, but it also contains duplications in the homopolymer rich flanking regions that is not seen in the other assemblies.

The mate-pair library was not able to scaffold the *discover* contigs due to the size of this repeat being larger than its 10 kbp insert size. The mate-pair fragments also map to a great depth in the repeat. Dovetail data, however, shows a much smoother fragment distribution and was able to scaffold the two *discover* contigs in the correct order and orientation as it could scaffold up to 50 kbp (the cutoff used by the HiRise scaffolder). However, the gap length was not estimated with Dovetail and was arbitrarily set to 100 Ns when in reality the gap is over 12 000 bp long.

### Long-range accuracy using synteny to *S. tuberosum*

As all our assemblies are *de novo*, in the sense that we used no prior information from other Solanaceae genomes, we reasoned that more accurate long range scaffolding would be apparent as longer syntenic blocks to a closely related species. We used nucmer [20]



**Figure 6:** Mummer plots showing alignment to chromosome 11 of the *S. tuberosum* reference. The *S. tuberosum* reference is shown on the x-axis and assembly scaffolds on the y-axis. Alignments shown are at least 10 kbp long and 90 % identical.

to analyse the synteny of our assemblies to the pseudomolecules of the *S. tuberosum* genome [38]. Figure 6 shows the mummer plot for chromosome 11 of *S. tuberosum* against our three assemblies. We saw the **falcon-dt-bn** assembly showed the best synteny with the **discoverar-mp-dt-bn** being the worst. The plots for the remaining chromosomes are shown in Supplemental Figures S3.2, S3.3, and S3.4.

Using synteny we identified two cases of chimerism, i.e. scaffolds that align well to two different pseudomolecules of *S. tuberosum* genome. Both cases are in **discoverar-mp-dt-bn** but not **falcon-dt-bn**. The first 1.5 Mbp of scaffold ScEqE3Q\_528 maps to pseudomolecule 7 while the last 2.9 Mbp map to pseudomolecule 2 in the *S. tuberosum* genome. There is no conflict reported with the BioNano Genomics optical map in this area, but we can exclude the possibility that these are real chromosome structural arrangements in *S. verrucosum* because we have GbS markers on each end of this scaffold which also map in an *S. verrucosum* cross to these different linkage groups (López-Girona unpublished). The other case is a scaffold ScEqE3Q\_633 in which the first 1.4 Mbp map to pseudomolecule 8 and the remainder to pseudomolecule 3, here BioNano Genomics does report a conflict which would highlight this error, and *S. verrucosum* genetic markers also support the chimera classification.

## Discussion

A DISCOVAR assembly is the cheapest and easiest to construct, and the resulting assembly is very accurate, albeit highly fragmented. Adding a long mate-pair library is a proven method of increasing the contiguity of a short read assembly by scaffolding. The 10x Genomics based assembly using SUPERNOVA was as easy to obtain as the DISCOVAR assembly. The two most remarkable features of this assembly are the low cost and input DNA requirement: for only slightly higher cost than a DISCOVAR assembly, and considerably less than with only one long mate-pair library, we obtained an assembly comparable to what one would expect from multiple long mate-pair libraries.

Our PacBio assembly using FALCON achieved contiguity similar to that of **discoverar-mp** (DISCOVAR plus long mate-pair scaffolding). PacBio sequencing has a considerably higher cost and material requirement than Illumina sequencing, but the **falcon** assembly contains truly contiguous sequence as opposed to

**discoverar-mp** which contains gaps patched with Ns. The PacBio read lengths ( $N50=13.5$  kbp) were similar to the insert size of mp library (mean 10 kb), and the read coverage was higher for PacBio ( $50\times$ ) than for the mp data ( $15\times$ ), but PacBio contigs ( $N50=712$  kbp) are slightly shorter than the **discoverar-mp** scaffolds ( $N50=858$  kbp).

The addition of Dovetail showed the most striking increase in contiguity by scaffolding. We note that our Dovetail scaffolds provided the order and orientation of the constituent contigs but no estimate for the length of the gaps between them. This should be taken into consideration if true physical length of sequences is important, and for specific downstream uses. Both Illumina (DISCOVAR+MP) and PacBio (FALCON) assemblies are amenable to the addition of Dovetail, but the scaffolds produced from the FALCON contigs ( $4\times$  increase) were not as long as those from the Illumina assembly ( $5.5\times$  increase). This could be because while the FALCON assembly has been polished with PacBio reads, it retains some PacBio errors and so some Dovetail (Illumina) reads do not pass stringent mapping filters. If true, Pilon polishing with Illumina reads could help, as it improved the  $k$ -mer spectra (Figure 2).

With BioNano Genomics restriction enzyme digest based optical maps we obtained less ( $\sim 2\times$  increase) scaffolding improvement than with Dovetail ( $4-5.5\times$  increase). This could be due to three issues: first that assembly gaps are not correctly sized which prevents real, and *in silico*, restriction maps matching (as information is purely encoded in the distances between sites). We see that the ungapped PacBio assemblies improve more than scaffolded Illumina, and Dovetail scaffolds (with arbitrary 100 bp gaps) hardly increase at all. Secondly, because the method produces low information density (one enzyme site per  $\sim 12$  kbp) long fragments with many sites are needed to create significant matches, and our DNA was not sufficiently long (best run  $N50$  was 240 kbp). Longer DNA (over 300 kbp), and perhaps multiple enzyme maps with iterative scaffolding could have improved the results. Thirdly we observe that the *in silico* restriction rates for Illumina and PacBio assemblies are similar ( $8.1-8.4$  sites /100 kbp) whereas the actual observed rates from the physical map is much lower at 6.8 sites/100 kbp, suggesting that there could be a fraction of the genome missing from our assemblies which is very low in sites such as centromeric or telomeric regions where the BioNano Genomics map can not scaffold through.

Gapfilling using PBJelly offers an attractive method of using

the long read data from PacBio to improve an existing Illumina based assembly. This closed many of the gaps in the scaffolds thereby decreasing the fraction of unknown sequence (Ns) and also increasing the contiguity. The increase in contiguity of the 10x Genomics assembly was the highest. It will be intriguing to see if an assembly approach combining Chromium data with long reads (directly on the assembly graph) can combine the best attributes of both data types to resolve complex regions.

Analysis of the  $k$ -mer content of the *supernova*, *discoverar*, and *falcon* assemblies showed that the  $k$ -mer spectra of each assembly is very clean. We see slightly higher level of sequence duplication in the *supernova* assembly, and to a lesser extent in the *falcon* assembly. All three assembly algorithms are diploid aware, meaning they are able to preserve both haplotypes. The gene content of each assembly was very similar with all three of our long assemblies showing a high percentage of the expected genes. The 10x Genomics based assembly showed a slightly lower count in both of our assessments but the difference is very small.

We used multiple BAC assemblies of ~100 kb insert size to illustrate the technical limitations of each method. Short read methods cannot resolve many areas of repetition within a WGS assembly. This is especially noticeable in a plant genome with higher repeat content, and is one of the major reasons for breaks in contiguity in these assemblies. In our example in Figure 3, the long mate-pair library alone is not sufficient. It takes the larger fragment lengths within the Dovetail Chicago library to finally make the join in the whole genome assembly.

Long read technologies do not suffer as much with repeats and, in the case of PacBio, tend to have more random rather than systematic errors [5]. We can see in our exemplar that the *falcon* assembly covers some of the repetitive region. The underlying BAC assembly was also obtained with PacBio and gave us a single true contig for the entire BAC. On close inspection we noticed that difficult region was spanned by reads of length 22–26 kbp. This shows that long reads are certainly able to span such regions of difficulty, and to assemble them. Recently ultra-long reads with an N50 of 99.7 kbp (max. 882 kbp) with ~92% accuracy have been produced with the new MinION R9.4 chemistry using high molecular weight DNA [17]. If this is also achievable on plant material the remaining repetitive fraction of genomes should become visible.

To evaluate the longer range accuracy of our genome assemblies we compared them to the closely related *S. tuberosum* pseudomolecule assembly, which revealed good synteny with all three of our longest assemblies (*discoverar-mp-dt-bn*, *falcon-dt-bn* and *supernova*). There are some disagreements especially in the centromeric areas, but as these appeared in all assemblies these could illustrate real structural variation. We detected two chimeric scaffolds in the *discoverar-mp-dt-bn* assembly but neither is present in the *falcon-dt-bn*. The two Dovetail scaffolding processes shared the same Hi-C sequence data but were conducted many months apart (*discoverar-mp* first and later *falcon*), so may use different versions of Dovetail's proprietary HiRise software. On detailed examination we see that the ScEqE3Q\_528 scaffold chimeric join is made by Dovetail hopping through a fragmented area of short (1–2 kbp) contigs. Such small contigs do not exist in the Falcon assembly, which maybe why we do not find chimeras. BioNano Genomics finds it hard to map to areas with many Dovetail gaps (as these are set to an arbitrary 100 bp size), and this region also has a high enzyme nicking rate (nearly twice the genome average), including two areas where nicks are less than 200 bp apart and so would be optically merged. In scaffold ScEqE3Q\_633

Library	Tissue type	Material/DNA amount	HMW	Fragment length (bp)
TALL	Frozen	3 µg	No	700
Discoverar	Frozen	0.6 µg	No	500
Mate-pair	Frozen	4 µg	No	10 000
PacBio	Young frozen	5 g	No	20 000
BioNano	Young fresh	2.5 µg	Yes	>100 000
Dovetail	Fresh	20 g	Yes	>100 000
Chromium	Flash frozen	0.5 g	Yes	>100 000

**Table 2:** Material requirements for each library. Amounts in grams are for fresh/frozen material and amounts in micrograms for DNA. In each case where frozen or flash frozen is stated, fresh material is also acceptable.

we detect that *discoverar-mp* scaffold123 was correctly split by Dovetail data as chimeric (also highlighted by BioNano Genomics and genetic markers) but the scaffold was not broken at the exact chimeric join, and the remaining sequence from the wrong chromosome was sufficient for Dovetail to propagate the error. Whilst we did not detect a high level of systematic errors in any of our assembly methods, the importance of using BioNano Genomics and genetic markers to identify chimeras that then can be broken is apparent.

## Materials and Methods

### Project requirements

Each of the assembly methods we have used comes with its own requirements. We have broken this down into material requirements, that is plant and DNA material, monetary requirements, that is the cost of preparation and sequencing, and computational requirements. Table 2 lists the material requirements for each library.

We calculated costs taking into consideration the costs of consumables, laboratory time, and machine overheads, but not bioinformatics time. For sequencing costs we used the Duke University cost as much as possible to provide comparative figures. Since several of the projects share common methods, such as sequencing a lane on a HiSeq 2500, we have broken down the costs into individual components. See Table 3 for our full costs calculations.

In many cases the assemblies can be performed with modest scientific computing facilities. In some cases, notably for *SUPERNOVA*, a very large amount of memory is required. In this case the computing requirement will not be available to most laboratories and will need to be sourced elsewhere. Table 4 shows the computational requirements of each assembly method.

### Library preparation and sequencing

In this section we briefly describe methods for library preparation and sequencing. For a comprehensive description, please see the supplementary material.

*S. verrucosum* accession Ver-54 was grown in the glass house in James Hutton Institute in Scotland. Both fresh and frozen leaves from this accession and its clones were used for DNA extraction.

The TALL library was prepared using 3 µg of DNA and fragments of 650 bp were sequenced with a HiSeq2500 with a 2×150 bp read metric. The DISCOVAR library was prepared using 600 ng of DNA and fragments of 500 bp were sequenced with a HiSeq2500 with a 2×250 bp read metric.

Assembly	Paired-end	Mate-pair	PacBio	Chromium	Dovetail	BioNano	HiSeq2500	MiSeq	PacBio RSII	Total (tsn)
discover	X						X			3,273
discover-mp	X	X				X	X	X		7,854
discover-mp-bn	X	X				X	X	X		8,803
discover-mp-dt	X	X			X		XX	X		32,793
discover-mp-dt-bn	X	X			X	X	XX	X		33,742
falcon				X					X	25,499
falcon-bn				X					X	26,448
falcon-dt			X		X		X		X	50,438
falcon-dt-bn			X		X	X	X		X	51,387
supernova				X			X			4,299
supernova-bn				X		X	X			5,248
Cost (USD)	209	595	474	1,235*	21,875	949*	3,064	3,986	25,025	

**Table 3:** The overall cost of each assembly project. We show which library preparations and sequencing runs are required for each assembly with a checkmark (X). Individual costs are given at the bottom, and total costs of each assembly on the right. All costs are according to Duke University as of April 2017 and in USD, except those marked with \* which were according to the Earlham Institute and converted from GBP to USD at an exchange rate of 0.804 GBP/USD. Paired-end, mate-pair, PacBio, and Chromium are library preparations including DNA extraction. Dovetail includes Chicago library preparation and HiRise scaffolding. BioNano is the cost of building the optical map. HiSeq2500 is for a rapid run half flowcell (one lane) with 250 bp reads. MiSeq is for two runs with 300 bp reads. PacBio RSII is for 65 SMRT cells.

Name of assembly	Approximate runtime	Peak memory	Average memory	System
Supernova	3 d	1300 GB		Large memory
Canu (Uncorr)	12 d	47 GB	20 GB	HPC cluster
Canu (Corr)	4 d	34 GB	14 GB	HPC cluster
Falcon	5 d	120 GB	60 GB	Large memory
HGAP	2 m	280 GB		Large memory
Discover	22 h	260 GB	134 GB	Large memory
ABYSS	1 w	64 GB		HPC cluster
BioNano (Asm)	8 h	64 GB	64 GB	HPC cluster
BioNano (Scaf)	1 d	64 GB	64 GB	HPC cluster

**Table 4:** Computational requirements.

The mate-pair library was prepared using 4 µg of DNA and fragments of 10 kbp were circularised, fragmented and sequenced on a MiSeq with a 2×300 bp read metric.

A PacBio library was prepared using 5 g of frozen leaf material. A 20 kbp fragment length library was prepared according to manufacturer's instructions and sequenced on 65 SMRT cells with the P6C4 chemistry on a PacBio RSII.

The 10x Chromium library was prepared according to the manufacturer's instructions and sequenced on a HiSeq2500 with a 2×250 bp read metric.

For BioNano, DNA was extracted using the IrysPrep protocol. 300 ng was used in the Nick, Label, Repair and Stain reaction and loaded onto a single flow cell on a BioNano chip. The chip was run eight times to generate 252 Gb of raw data.

### Assembly and evaluation

All tools and scripts that were used to perform the evaluation and produce the figures are available on GitHub in the georgek/potato-figures repository.

We used RAMPART [28] to run ABYSS [41] multiple times with different  $k$  values. DISCOVAR *de novo* was run with normal parameters.

Long mate-pair reads were first processed with NextClip [22] to remove the Nextera adapter. Soapdenovo2 was then used to perform scaffolding with both the paired-end and mate-pair libraries.

$k$ -mer content was analysed with the `kat comp` tool [29]. We used default parameters with manually adjusted plot axes to show the relevant information.

We used the BUSCO core plant dataset to evaluate the gene content. The *S. tuberosum* representative transcripts were aligned to the assemblies using BLAST and the coverage of transcripts at various thresholds using a tool we developed.

The BACs were sequenced with the Earlham Institute BAC pipeline [3] and were assembled with DISCOVAR *de novo* using normal parameters after filtering for *E. coli* and the BAC vector. The PacBio BAC was assembled using HGAP. We used GNU parallel [42] for concurrent assembly and analysis.

20 BACs which assembled into a single contig were selected to use as a reference. These BACs are non-redundant to the extent that they do not share any lengths of sequence of more than 95 % identity and over 5000 bp long. Short reads were aligned to the BACs using Bowtie2 [21] with default parameters. The assemblies were mapped to the BACs using `bwa mem` [23]. The mapped sequences were sorted and filtered for quality using `sambamba` [43]. Fragment coverage was calculated using `samtools` [24] and `bedtools` [36].

Synteny was analysed with `mummer` [10]. We used `nuclmer` to align the assemblies to the *S. tuberosum* reference v4.04 [15]. Alignments less than 10 kbp and 90 % identity were filtered out.

### Data Access

All read data generated in this study have been submitted to the EMBL-EBI European Nucleotide Archive under the project PRJEB20860.

### Acknowledgements

We thank Lawrence Percival-Alwyn and Walter Verweij for their assistance in library preparation and analysis, and Michael Bevan for critical reading of this manuscript. This work was funded with BBSRC project grants (BB/K019325/1) and (BB/K019090/1). This work was strategically funded by the BBSRC, Core Strategic Programme Grant (BB/CSP17270/1) at the Earlham Institute. High-throughput sequencing and library construction

was delivered via the BBSRC National Capability in Genomics (BB/CCG1720/1) at the Earlham Institute (EI, formerly The Genome Analysis Centre, Norwich), by members of the Platforms and Pipelines Group. This research was supported in part by the NBI Computing Infrastructure for Science (CiS) group through the HPC cluster and UV systems. We thank Duke University for providing sequencing costs via DUGSIM (<https://dugsim.net/>).

*Authors' contributions:* GB, ELG, IH, and GW prepared the sample. MDC, GK, and PP designed the analysis. DB, GB, ELG, MG, DH, IH, AL, and IM constructed libraries and performed sequencing. GK and PP made the assemblies and GK, ELG, and PP performed the evaluation. MDC, GK, GB, ELG and PP wrote and prepared the manuscript. All authors read and approved the final manuscript.

## References

- [1] The 1000 Genomes Project Consortium. “An integrated map of genetic variation from 1,092 human genomes”. In: *Nature* 491.7422 (2012), pp. 56–65. ISSN: 0028-0836. DOI: 10.1038/nature11632.
- [2] S. F. Altschul et al. “Basic local alignment search tool”. In: *Journal of Molecular Biology* 215.3 (1990), pp. 403–410. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80360-2.
- [3] S. Beier et al. “Construction of a map-based reference genome sequence for barley, *Hordeum vulgare L.*” In: *Scientific Data* 4 (2017). DOI: 10.1038/sdata.2017.44.
- [4] E. Callaway. “‘Platinum’ genome takes on disease”. In: *Nature News* 515.7527 (2014), p. 323. DOI: 10.1038/515323a.
- [5] M. O. Carneiro et al. “Pacific biosciences sequencing technology for genotyping and variation discovery in human data”. In: *BMC Genomics* 13.1 (2012), p. 375. ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-375.
- [6] M. J. P. Chaisson et al. “Resolving the complexity of the human genome using single-molecule sequencing”. In: *Nature* 517.7536 (2015), pp. 608–611. ISSN: 0028-0836. DOI: 10.1038/nature13907.
- [7] C.-S. Chin et al. “Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data”. In: *Nature Methods* 10.6 (2013), pp. 563–569. ISSN: 1548-7091. DOI: 10.1038/nmeth.2474.
- [8] C.-S. Chin et al. “Phased diploid genome assembly with single-molecule real-time sequencing”. In: *Nature Methods* 13.12 (2016), pp. 1050–1054. ISSN: 1548-7091. DOI: 10.1038/nmeth.4035.
- [9] F. Choulet et al. “Structural and functional partitioning of bread wheat chromosome 3B”. In: *Science* 345.6194 (2014), pp. 1249721–1249721. ISSN: 0036-8075. DOI: 10.1126/science.1249721.
- [10] A. L. Delcher, S. L. Salzberg and A. M. Phillippy. “Using MUMmer to identify similar regions in large sequence sets”. In: *Current Protocols in Bioinformatics* (2003), pp. 10–3. DOI: 10.1002/0471250953.bi1003s00.
- [11] A. C. English et al. “Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology”. In: *PLOS ONE* 7.11 (2012), e47768. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0047768.
- [12] A. Fleischmann et al. “Evolution of genome size and chromosome number in the carnivorous plant genus *Gentisea* (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms”. In: *Annals of Botany* 114.8 (2014), pp. 1651–1663. ISSN: 0305-7364. DOI: 10.1093/aob/mcu189.
- [13] E. A. Friar. “Isolation of DNA from plants with large amounts of secondary metabolites”. In: *Methods in Enzymology. Molecular Evolution: Producing the Biochemical Data* 395 (2005), pp. 1–12. ISSN: 0076-6879. DOI: 10.1016/S0076-6879(05)95001-5.
- [14] “Genome in a bottle—a human DNA standard”. In: *Nature Biotechnology* 33.7 (2015), pp. 675–675. ISSN: 1087-0156. DOI: 10.1038/nbt0715-675a.
- [15] M. A. Hardigan et al. “Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum tuberosum*”. In: *The Plant Cell* (2016), TPC2015-00538-RA. ISSN: 1532-298X. DOI: 10.1105/tpc.15.00538.
- [16] A. R. Hasty et al. “Rapid genome mapping in nanochannel arrays for highly complete and accurate *de novo* sequence assembly of the complex *Aegilops tauschii* genome”. In: *PLOS ONE* 8.2 (2013), e55864. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0055864.
- [17] M. Jain et al. “Nanopore sequencing and assembly of a human genome with ultra-long reads”. In: *bioRxiv* (2017). DOI: 10.1101/128835. eprint: <https://www.biorxiv.org/content/early/2017/04/20/128835.full.pdf>.
- [18] W.-B. Jiao and K. Schneeberger. “The impact of third generation genomic technologies on plant genome assembly”. In: *Current Opinion in Plant Biology* 36 (2017), pp. 64–70. ISSN: 1369-5266. DOI: 10.1016/j.pbi.2017.02.002.
- [19] S. Koren et al. “Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation”. In: *Genome Research* 27.5 (2017), pp. 722–736. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.215087.116.
- [20] S. Kurtz et al. “Versatile and open software for comparing large genomes”. In: *Genome Biology* 5.2 (2004), R12. DOI: 10.1186/gb-2004-5-2-r12.
- [21] B. Langmead and S. L. Salzberg. “Fast gapped-read alignment with Bowtie 2”. In: *Nature Methods* 9.4 (2012), pp. 357–359. DOI: 10.1038/nmeth.1923.
- [22] R. M. Leggett et al. “Nextclip: an analysis and read preparation tool for Nextera long mate pair libraries”. In: *Bioinformatics* 30.4 (2014), pp. 566–568. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt702.
- [23] H. Li. “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM”. In: *arXiv preprint arXiv:1303.3997* (2013).
- [24] H. Li et al. “The sequence alignment/map format and SAMtools”. In: *Bioinformatics* 25.16 (2009), pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352.
- [25] R. Li et al. “*De novo* assembly of human genomes with massively parallel short read sequencing”. In: *Genome Research* 20.2 (2010), pp. 265–272. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.097261.109.
- [26] R. Li et al. “The sequence and *de novo* assembly of the giant panda genome”. In: *Nature* 463.7279 (2010), pp. 311–317. ISSN: 0028-0836. DOI: 10.1038/nature08696.
- [27] R. Luo et al. “SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler”. In: *GigaScience* 1 (2012), p. 18. ISSN: 2047-217X. DOI: 10.1186/2047-217X-1-18.
- [28] D. Mapleson, N. Drou and D. Swarbreck. “RAMPART: a workflow management system for *de novo* genome assembly”. In: *Bioinformatics* 31.11 (2015), pp. 1824–1826. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv056.
- [29] D. Mapleson et al. “KAT: a k-mer analysis toolkit to quality control NGS datasets and genome assemblies”. In: *Bioinformatics* (2016). DOI: 10.1093/bioinformatics/btw663.
- [30] M. Mascher et al. “A chromosome conformation capture ordered sequence of the barley genome”. In: *Nature* 544.7651 (2017), pp. 427–433. ISSN: 0028-0836. DOI: 10.1038/nature22043.
- [31] Y. Mostovoy et al. “A hybrid approach for *de novo* human genome sequence assembly and phasing”. In: *Nature Methods* 13.7 (2016), pp. 587–590. ISSN: 1548-7091. DOI: 10.1038/nmeth.3865.
- [32] J. Pellicer, M. F. Fay and I. J. Leitch. “The largest eukaryotic genome of them all?” In: *Botanical Journal of the Linnean Society* 164.1 (2010), pp. 10–15. ISSN: 0024-4074. DOI: 10.1111/j.1095-8339.2010.01072.x.

- [33] M. Pendleton et al. “Assembly and diploid architecture of an individual human genome via single-molecule technologies”. In: *Nature Methods* 12.8 (2015), pp. 780–786. ISSN: 1548-7091. DOI: 10.1038/nmeth.3454.
- [34] The Potato Genome Sequencing Consortium. “Genome sequence and analysis of the tuber crop potato”. In: *Nature* 475.7355 (2011), pp. 189–195. ISSN: 0028-0836. DOI: 10.1038/nature10158.
- [35] N. H. Putnam et al. “Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage”. In: *Genome Research* 26.3 (2016), pp. 342–350. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.193474.115.
- [36] A. R. Quinlan and I. M. Hall. “BEDTools: a flexible suite of utilities for comparing genomic features”. In: *Bioinformatics* 26.6 (2010), pp. 841–842. DOI: 10.1093/bioinformatics/btq033.
- [37] J. M. Rothberg et al. “An integrated semiconductor device enabling non-optical genome sequencing”. In: *Nature* 475.7356 (2011), pp. 348–352. ISSN: 0028-0836. DOI: 10.1038/nature10242.
- [38] S. K. Sharma et al. “Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps”. In: *G3&#38; GenesGenomes|Genetics* 3.11 (2013), pp. 2031–2047. ISSN: 2160-1836. DOI: 10.1534/g3.113.007153.
- [39] F. A. Simão et al. “BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs”. In: *Bioinformatics* 31.19 (2015), pp. 3210–3212. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv351.
- [40] J. T. Simpson and R. Durbin. “Efficient *de novo* assembly of large genomes using compressed data structures”. In: *Genome Research* 22.3 (2012), pp. 549–556. DOI: 10.1101/gr.126953.111.
- [41] J. T. Simpson et al. “ABySS: a parallel assembler for short read sequence data”. In: *Genome Research* 19.6 (2009), pp. 1117–1123. DOI: 10.1101/gr.089532.108.
- [42] O. Tange. “GNU parallel—the command-line power tool”. In: *login: The USENIX Magazine* 36.1 (2011), pp. 42–47. DOI: 10.5281/zenodo.16303.
- [43] A. Tarasov et al. “Sambamba: fast processing of NGS alignment formats”. In: *Bioinformatics* 31.12 (2015), pp. 2032–2034. DOI: 10.1093/bioinformatics/btv098.
- [44] B. J. Walker et al. “Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement”. In: *PLOS ONE* 9.11 (2014), pp. 1–14. DOI: 10.1371/journal.pone.0112963.
- [45] N. I. Weisenfeld et al. “Comprehensive variation discovery in single human genomes”. In: *Nature Genetics* 46.12 (2014), pp. 1350–1355. ISSN: 1061-4036. DOI: 10.1038/ng.3121.
- [46] N. I. Weisenfeld et al. “Direct determination of diploid genome sequences”. In: *Genome Research* (2017). ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.214874.116.

---

## Supplemental material for *A critical comparison of technologies for a plant genome sequencing project*

Pirita Paajanen<sup>1\*</sup>, George Kettleborough<sup>1\*</sup>, Elena López-Girona<sup>2</sup>, Michael Giolai<sup>1</sup>, Darren Heavens<sup>1</sup>, Dave Baker<sup>1</sup>, Ashleigh Lister<sup>1</sup>, Gail Wilde<sup>2</sup>, Ingo Hein<sup>2</sup>, Iain Macaulay<sup>1</sup>, Glenn J. Bryan<sup>2</sup>, and Matthew D. Clark<sup>1+</sup>

\*contributed equally to this work

+corresponding author: matt.clark@earlham.ac.uk

<sup>1</sup>Earlham Institute, Norwich, UK; <sup>2</sup>The James Hutton Institute, Invergowrie, Dundee, UK.

### Contents

<b>1 DNA Extraction and Library Preparation</b>	<b>3</b>
1.1 Germplasm . . . . .	3
1.2 DNA extraction . . . . .	3
1.2.1 CTAB DNA extraction . . . . .	3
1.2.2 IrysPrep DNA extraction . . . . .	3
1.3 Illumina library preparation . . . . .	4
1.3.1 Tight and Long Library (TALL) paired-end library construction protocol . . . . .	4
1.3.2 Amplification free paired-end library construction (DISCOVAR) . . . . .	4
1.3.3 Long mate pair library construction protocol . . . . .	5
1.4 PacBio library preparation . . . . .	6
1.5 10x library preparation . . . . .	6
1.6 BioNano preparation . . . . .	6
1.7 Dovetail extraction and library preparation . . . . .	6
1.8 BAC library . . . . .	6
<b>2 Assembly and Scaffolding</b>	<b>8</b>
2.1 Illumina . . . . .	8
2.1.1 ABySS . . . . .	8
2.1.2 DISCOVAR . . . . .	8
2.1.3 Mate-pair scaffolding . . . . .	8
2.2 PacBio . . . . .	8
2.2.1 Falcon . . . . .	8
2.2.2 Canu . . . . .	9
2.2.3 HGAP . . . . .	9
2.3 10x Genomics Supernova . . . . .	9
2.4 BioNano . . . . .	9
2.4.1 Optical map assembly . . . . .	10
2.4.2 Scaffolding . . . . .	10
2.5 Dovetail . . . . .	10
2.6 BAC assembly . . . . .	10
2.6.1 Illumina . . . . .	10
2.6.2 PacBio . . . . .	10
<b>3 Evaluation</b>	<b>11</b>
3.1 K-mer content . . . . .	11
3.2 Synteny . . . . .	11
3.3 Gene content . . . . .	11

3.4	BUSCO	11
<b>4</b>	<b>Data availability</b>	<b>15</b>
4.1	Short reads	15
4.2	Long reads	15
4.3	Optical map	15
4.4	Assemblies	15
<b>5</b>	<b>Authors' contributions</b>	<b>16</b>
<b>6</b>	<b>File list</b>	<b>16</b>

Library name	Library type	Read pair count	Read length (bp)	Insert size (bp)	Read coverage	Fragment coverage
LIB6268	TALL	308 350 323	150	650	137.25	274.50
LIB6487	Mate-pair	44 864 446	300	10 000	39.94	665.64
LIB12786	Discovar	160 305 585	250	500	118.92	118.92
LIB17395	Chicago (Dovetail)	142 321 242	250		105.58	
LIB24104	Chromium (10x)	141 344 719	250	900	104.86	188.74

Table S1.1: Summary of short read library sequencing. Insert size is given where a mode is appropriate.

## 1 DNA Extraction and Library Preparation

### 1.1 Germplasm

*S. verrucosum* accession Ver-54 was grown in the glass house in James Hutton Institute in Scotland. Both fresh and frozen leaves from this accession and its clones were used for DNA extraction.

### 1.2 DNA extraction

#### 1.2.1 CTAB DNA extraction

Genomic DNA was extracted using CTAB lysis, phenol-chloroform, and Qiagen MagAttract HMW DNA Kit (QIAGEN, Manchester, UK) purification. Young frozen *S. verrucosum* Ver-54 leaves (5 g) were ground to a fine powder using a liquid nitrogen cooled pestle and mortar, distributed over two 50 ml falcon tubes and mixed with 20 ml CTAB buffer (100mM Tris-HCl, pH 8.0, 2 % (w/v) CTAB, 1.4 M NaCl, 20 mM EDTA) containing 20 µg/ml proteinase K and incubated at 55 °C for 20 minutes. Next, 0.5 volumes (8 ml) chloroform was added and carefully mixed by 15× inversion followed by centrifugation at 2990×g on an Eppendorf Centrifuge 5810 R (Eppendorf, Stevenage, UK) for 30 minutes. The aqueous phase was carefully transferred into a new tube to which 1 volume phenol/chloroform/isoamyl alcohol was added followed by centrifugation for 30 minutes at 2990×g. The aqueous phase was ethanol precipitated by addition of 3 M sodium acetate (1/10 of DNA volume), (pH 5.2) and 2.5 volumes of ethanol, mixed and precipitated at 2990×g and 4 °C. DNA pellets were washed with ice-cold 70 % ethanol, air dried and resuspended in 350 µl of 1× TE buffer (10 mM Tris-HCl, 1 mM EDTA pH 8.0) containing 10 U RNase One (Promega, Southampton, UK). The DNA was dissolved overnight at 4 °C with occasional mixing by inversion. Finally, we purified the DNA with the QIAGEN MagAttract HMW DNA Kit with minor changes to the manufacture instructions at the beginning; we added 150 µl AL Buffer to 200 µl DNA (10 µg in TE) and 15 µl magnetic bead suspension. From this point we followed the Quick-Start Protocol and we eluted the DNA in 200 µl of 1× TE buffer.

#### 1.2.2 IrysPrep DNA extraction

Fresh young leaves of the *S. verrucosum* Ver-54 accession were collected after 48-hour treatment in the dark. Earlham Institute's Platforms and Pipelines group followed IrysPrep "Fix'n'Blend" Plant DNA extraction protocol supplied by BioNano Genomics. 2.5 g of fresh young leaves were fixed with 2 % formaldehyde. After washing, leaves are disrupted and homogenized in the presence of isolation buffer. The isolation buffer contains PVP10 and BME to prevent oxidation of polyphenols. Triton X-100 is added to facilitate the release of nuclei from the broken cells. The nuclei are then purified on a Percoll cushion. A nuclei phase is taken and washed several times in isolation buffer before embedding into low melting point agarose. 2 plugs of 90 µl were cast using the CHEF Mammalian Genomic DNA Plug Kit (Bio-Rad 170-3591). Once set at 4 °C the plugs were added to a lysis solution containing 200 µl proteinase K (QIAGEN 158920) and 2.5 ml of BioNano lysis buffer in a 50 ml conical tube. These were put at 50 °C for 2 hours on a thermomixer, making a fresh proteinase K solution to incubate overnight. The 50 ml tubes were then removed from the thermomixer for 5 minutes before 50 µl RNase A (Qiagen158924) was added and the tubes returned to the thermomixer for a further hour at 37 °C. The plugs were then washed 7 times in Wash Buffer supplied in Chef kit and 7 times in 1xTE. One plug was removed and melted for 2 minutes at 70 °C followed by 5 minutes at 43 °C before adding 10 µl of 0.2 U /µl of GELase (Cambio Ltd G31200). After 45 minutes at 43 °C the melted plug was dialysed on a 0.1 µM membrane (Millipore VCWP04700) sitting on 15 ml of 1xTE in a small petri dish. After 2 hours the sample was removed with a wide bore tip and mixed gently 5 times and left overnight at 4 °C.

### 1.3 Illumina library preparation

#### 1.3.1 Tight and Long Library (TALL) paired-end library construction protocol

DNA was extracted using the CTAB protocol given in Section 1.2.1. A total of 3 µg of DNA was sheared in a 60 µl volume on a Covaris S2 (Covaris, Massachusetts, USA) for 1 cycle of 40 seconds with a duty cycle of 5 %, cycles per burst of 200 and intensity of 3. The fragmented DNA was then subjected to size selection on a Blue Pippin (Sage Science, Beverly, USA). The 40 µl in each of collection wells was replaced with fresh buffer and the separation and elution current checked prior to loading the sample. To 30 µl of the end repaired molecules 10 µl of R2 marker solution was added and then loaded onto a 1.5 % Cassette. The Blue Pippin was configured to collect fragments at 800 bp using the tight settings. Post size selection, the 40 µl from the collection well was recovered and the size isolated estimated on High Sensitivity BioAnalyser Chip and DNA concentration determined using a Qubit HS Assay.

The size selected molecules were then end repaired in 100 µl volume using the NEB End Repair Module (NEB, Hitchin, UK) incubating the reaction at 22 °C for 30 minutes. Post incubation 100 µl beads of CleanPCR beads (GC Biotech, Alphen aan den Rijn, The Netherlands) were added and the DNA precipitated onto the beads. They were then washed twice with 70 % ethanol and the end repaired molecules eluted in 25 µl Nuclease free water (Qiagen, Manchester, UK).

End repaired molecules were then A tailed in 30 µl volume using in the NEB A tailing module (NEB) incubating the reaction at 37 °C for 30 minutes. To the A tailed library molecules 1 µl of an appropriate Illumina TruSeq Index adapter (Illumina, San Diego, USA) was added and mixed then 31 µl of Blunt/ TA ligase (NEB) added and incubated at 22 °C for 10m. Post incubation 5 µl of stop ligation was added and then the reaction incubated at room temperature for 5 minutes. Following this incubation 67 µl beads of CleanPCR beads (GC Biotech, Alphen aan den Rijn, The Netherlands) were added and the DNA precipitated onto the beads. They were then washed twice with 70 % ethanol and the end repaired molecules eluted in 100 µl nuclease free water. Two further 1 × CleanPCR bead based purifications were undertaken to remove any adapter dimer molecules that may have formed during the adapter ligation step and the final library eluted in 25 µl Resuspension Buffer (Illumina).

Library QC was performed by running a 1 µl aliquot on a High Sensitivity BioAnalyser chip (Agilent, Stockport, UK) and the DNA concentration measured using the High Sensitivity Qubit (Thermo Fisher, Cambridge, UK). To determine the number of viable library molecules the library was subjected to quantification by the Kappa qPCR Illumina quantification kit (Kapa Biosystems, London, UK) and then sequenced on the Hiseq2500 (Illumina) with a 2 × 150 bp read metric.

#### 1.3.2 Amplification free paired-end library construction (DISCOVAR)

DNA was extracted using the CTAB protocol given in Section 1.2.1. A total of 600 ng of DNA was sheared in a 60 µl volume on a Covaris S2 (Covaris, Massachusetts, USA) for 1 cycle of 40 seconds with a duty cycle of 5 %, cycles per burst of 200 and intensity of 3. The fragmented molecules were then end repaired in 100 µl volume using the NEB End Repair Module (NEB, Hitchin, UK) incubating the reaction at 22 °C for 30 minutes. Post incubation 58 µl beads of CleanPCR beads (GC Biotech, Alphen aan den Rijn, The Netherlands) were added using a positive displacement pipette to ensure accuracy and the DNA precipitated onto the beads. They were then washed twice with 70 % ethanol and the end repaired molecules eluted in 25 µl Nuclease free water (Qiagen, Manchester, UK).

End repaired molecules were then A tailed in 30 µl volume using in the NEB A tailing module (NEB) incubating the reaction at 37 °C for 30 minutes. To the A tailed library molecules 1 µl of an appropriate Illumina TruSeq Index adapter (Illumina, San Diego, USA) was added and mixed then 31 µl of Blunt/TA ligase (NEB) added and incubated at 22 °C for 10m. Post incubation 5 µl of stop ligation was added and then the reaction incubated at room temperature for 5 minutes. Following this incubation 67 µl beads of CleanPCR beads (GC Biotech, Alphen aan den Rijn, The Netherlands) were added and the DNA precipitated onto the beads. They were then washed twice with 70 % ethanol and the end repaired molecules eluted in 100 µl nuclease free water. Two further CleanPCR bead based purifications were undertaken to remove any adapter dimer molecules that may have formed during the adapter ligation step. The first with 0.9 × volume beads, the second with 0.6 × and the final library eluted in 25 µl Resuspension Buffer (Illumina).

Library QC was performed by running a 1 µl aliquot on a High Sensitivity BioAnalyser chip (Agilent, Stockport, UK) and the DNA concentration measured using the High Sensitivity Qubit (Thermo Fisher, Cambridge, UK). To determine the number of viable library molecules the library was subjected to quantification by the Kappa qPCR Illumina quantification kit (Kapa Biosystems, London, UK) and then sequenced at a loading concentration of 9 pM on the Hiseq2500s (Illumina) with a 2 × 250bp read metric.

### 1.3.3 Long mate pair library construction protocol

DNA was extracted using the CTAB protocol given in Section 1.2.1. For the Tagmentation reactions 4 µg of Genomic DNA was prepared in 308 µl and then 80 µl 5 × Tagment Buffer Mate Pair (Illumina, San Diego, USA) added followed by 12 µl Mate Pair Tagmentation Enzyme (Illumina) and the reaction gently vortexed to mix. This was then incubated for 30 minutes at 55 °C, 100 µl of Neutralize Tagment Buffer (Illumina) added and then incubated at room temperature for 5 minutes. A 1 × volume bead clean-up was performed with CleanPCR beads (GC Biotech, Alphen aan den Rijn, The Netherlands) and the DNA eluted in 165 µl of Nuclease free Water . A 1 µl aliquot was run on a BioAnalyser 1200 chip and DNA concentration determined using a Qubit HS Assay.

Strand Displacement was performed by combining 162 µl of tagmented DNA, 20 µl 10× Strand Displacement Buffer (Illumina), 8 µl dNTPs (Illumina) and 10 µl Strand Displacement Polymerase (Illumina). This was then incubated at room temperature for 30 minutes. A 0.75 × volume bead clean-up was performed with CleanPCR beads and the DNA eluted in 16 µl of Nuclease free Water and the eluted DNA from the 3 µg and 6 µg reactions pooled. A 1 µl aliquot was diluted 1:6 and run on a BioAnalyser 1200 chip chip (Agilent, Stockport, UK) and DNA concentration determined using a Qubit HS Assay (Thermo Fisher, Cambridge, UK).

Size selection was performed on a Sage Science Blue Pippin (Sage Science, Beverly, USA). The 40 µl in the collection well was replaced with fresh buffer and the collection and elution current checked prior to loading the sample. To 30 µl of the pooled Strand Displaced reaction 10 µl of loading solution was added and then loaded onto a 0.75 % Cassette which was configured to collect fragments at 10 kbp using the tight setting. Post size selection, the 40 µl from the collection well was recovered and the DNA concentration determined using a Qubit HS Assay.

Circularisation was performed by combining 40 µl of size fractionated DNA, 12.5 µl of 10× circularisation buffer (Illumina), 3 µl Circularisation Enzyme (Illumina) and 75 µl nuclease free water.

This was then incubated at 30 °C overnight. Linear DNA was digested by adding 3.75 µl Exonuclease (Illumina) and incubating at 37 °C for 30 minutes followed by 70 °C for 30 minutes to denature the enzyme and 5 µl of stop ligation (Illumina) added. During exonuclease treatment 20 µl of M280 Dynabeads (Thermo Fisher) were prepared by washing twice with 100 µl Bead Bind Buffer (Illumina) before resuspending in 130 µl Bead Bind Buffer. Circularised DNA was then sheared in 130 µl volume on a Covaris S2 for 2 cycles of 37 seconds with a duty cycle of 10 %, cycles per burst of 200 and intensity of 4.

To 130 µl fragmented DNA 130 µl of washed M280 beads was added, mixed and then placed on a lab rotator at room temperature for 20 minutes. Library molecules bound to M280 beads were then washed four times with 200 µl Bead Washer Buffer (Illumina) and twice with 200 µl Resuspension Buffer (Illumina).

A master mix containing 85 µl nuclease free water (Qiagen, Manchester, UK), 10 µl 10× End Repair Reaction Buffer (NEB, Hitchin, UK) and 5 µl end repair enzyme mix (NEB) was prepared and added to the tube, mixed with the beads and incubated at room temperature for 30 minutes. End repaired library molecules bound to M280 beads were then washed four times with 200 µl Bead Washer Buffer and twice with 200 µl Resuspension Buffer.

A master mix containing 25 µl nuclease free water, 3 µl A Tailing 10× Reaction Buffer (NEB) and 2 µl A tailing enzyme mix (NEB) was prepared and added to the tube, mixed with the beads and incubated at 37 °C for 30 minutes. To the A tailed library molecules 1 µl of the appropriate Illumina Index adapter (Illumina) was added and mixed then 31 µl of Blunt/TA ligase (NEB) added and incubated at room temperature for 10 minutes. Post incubation 5 µl of stop ligation added and then the adapter ligated library molecules bound to M280 beads were then washed four times with 200 µl Bead Washer Buffer and twice with 200 µl Resuspension Buffer.

A master mix containing 20 µl nuclease free water, 25 µl 2× Kappa HiFi (Kappa Biosystems) and 5 µl Illumina Primer Cocktail (Illumina) was prepared and added to each tube, mixed with the beads and the contents, including beads, transferred to a 200 µl PCR tube. Each sample was then subjected to amplification on a Veriti Thermal Cycler (Thermo Fisher) with the following conditions:- 98 °C for 3 minutes, 15 cycles of PCR of 98 °C for 10 seconds, 60 °C for 30 seconds, 72 °C for 30 seconds followed by 72 °C for 5 minutes and Hold at 4 °C.

Post amplification the PCR tube was placed on a magnetic plate, the beads allowed to pellet and then 45 µl of the PCR transferred to a 2 ml Lobind Eppendorf Tube. To this 31.5 µl beads of CleanPCR beads were added to precipitate the DNA, the beads washed twice with 70 % ethanol and the final library eluted in 20 µl resuspension buffer. Library QC was performed by running a 1 µl aliquot on a High Sensitivity BioAnalyser chip and the DNA concentration measured using the High Sensitivity Qubit. The quantification of the pool was determined by the Kappa qPCR Illumina quantification kit with the pool run at 10 pM on a MiSeq with a 2×300 bp reads read metric.

## 1.4 PacBio library preparation

DNA was extracted using the CTAB protocol given in Section 1.2.1. We created a 20 kbp fragment length library according to the manufacturer's instructions as 20 kbp Template Preparation Using BluePippinTM Size-Selection System, and sequenced 65 SMRT cells with the P6C4 chemistry on the PacBio RSII instrument.

The total yield was 32 Gb of Data, the final N50 of read length was 13 499 bp. The total coverage of the raw data was 50 $\times$ .

## 1.5 10x library preparation

DNA was extracted using the CTAB protocol given in Section 1.2.1. DNA material was diluted to 1.1ng/ul with EB (Qiagen) and checked with a QuBit Flurometer 2.0 (Invitrogen) using the QuBit dsDNA HS Assay kit. The Chromium User Guide was followed as per the manufacturer's instructions (10X Genomics, PN-120229).

The final library was quantified using qPCR (KAPA Library Quant kit (Illumina), ABI Prism qPCR Mix, Kapa Biosystems). Sizing of the library fragments were checked using a Bioanalyzer (High Sensitivity DNA Reagents, Agilent). Samples were pooled based on the molarities calculated using the two QC measurements.

The library was clustered at 8 pm with a 1 % spike in of PhiX library (Illumina). The pool was run on a HiSeq2500 250 bp Rapid Run V2 mode (Illumina). The following run metrics were applied: Read 1: 250 cycles, Index 1: 8 cycles, Index 2: 0 cycles and Read 2: 250 cycles.

## 1.6 BioNano preparation

DNA was extracted using the IrysPrep protocol given in Section 1.2.2. A small amount was removed to QC on an Opgen Argus Q-Card and Qubit HS for the DNA concentration. 300 ng of DNA was taken into the NLRS (Nick, Label, Repair and Stain) reaction using 1  $\mu$ l Nt.BspQ1 (NEB R0644S). Following the NLRS reaction 16  $\mu$ l was loaded onto a single flow cell on a BioNano chip. The Chip loading was optimised and run for 30 cycles on the BioNano Irys using ICS1.6. The same chip was run a total of 8 times on each side to generate 252 Gb of raw data with molecule length over 100 kbp with a nick density of 6.79/100 kbp. Images were converted to .bnx files using AutoDetect 2.1.0.6656 before analysis.

HMW DNA was isolated and the nicking endonuclease Nt.BspQ1 (New England BioLabs) was used label high-quality HMW DNA molecules at specific sequence motifs (GCTCTTC). The nicked DNA molecules were then stained according to the instructions of IrysPrep Reagent Kit (BioNano Genomics).

The HMW DNA with fluorescent labels was loaded onto the nanochannel array of the IrysChip (BioNano Genomics) and was automatically imaged by the Irys system (BioNano Genomics). Raw DNA molecules of at least 100 kbp were collected and converted into BNX files by AutoDetect software to obtain basic labeling and DNA length information. The filtered raw DNA molecules in BNX format were aligned, clustered, and assembled into the BNG map by using the BioNano Genomics assembly pipeline as described in previous publications (40, 41). The P value thresholds used for pairwise assembly, extension/refinement, and final refinement stages were 1 $\times$ 10<sup>9</sup>, 1 $\times$ 10<sup>10</sup>, and 1 $\times$ 10<sup>10</sup>, respectively. The initial BNG map was then checked for potential chimeric BNG contigs and was further refined. To compare the draft sequence assembly with the BNG map, sequences were digested *in silico* according to the restriction site of Nt.BspQ1 by using Knickers (BioNano Genomics). The alignment of sequence assemblies with the BNG map was computed with RefAligner, and the visualization of the alignment was performed with snapshot in IrysView (<http://bionanogenomics.com/support/software-downloads/>).

## 1.7 Dovetail extraction and library preparation

20 g of fresh leaf material was sent to Dovetail Genomics, LLC (Santa Cruz, CA, USA). They extracted HMW DNA that was used to construct a Chicago library.

## 1.8 BAC library

Very young, partially expanded leaves from dark treated *S. verrucosum* Ver-54 plantlets were harvested and immediately frozen in liquid nitrogen. A 5 $\times$  pooled BAC library was generated by Bio S&T (Canada) and contained 192 pools of approximately 600 independent recombinants per pool. For the library generation, genomic DNA was subjected to HindIII restriction enzyme digest, ligated in to the vector pindigoBAC-5 (HindIII-Cloning Ready) (Epicentre) and transformed into compatible DH10B cells (Invitrogen). The average insert size was estimated to be 125 kbp following

NotI restriction enzyme digest and PFGE separation of nine randomly selected clones. Each well from one plate was then grown on a single plate and a single colony from each was selected for sequencing.

## 2 Assembly and Scaffolding

### 2.1 Illumina

#### 2.1.1 ABYSS

The Tight Amplification free Library (TALL) was assembled using ABYSS. We used RAMPART [1] to automate the running of ABYSS and select an appropriate  $k$  value ( $k = 113$  was selected). The configuration files for RAMPART is given in Supplemental File **S1** (pair\_end\_potato.xml). We used ABYSS version 1.5.1 and RAMPART version 0.10.3. The command use to run Rampart was:

```
rampart pair_end_potato.xml
```

#### 2.1.2 DISCOVAR

The DISCOVAR library was assembled using DISCOVAR de novo version 51828. We ran it with the following command:

```
DiscoverExp READS=DI_LIB12786_L1_R1.fastq.gz,DI_LIB12786_L1_R2.fastq.gz \
    OUT_DIR=potato NUM_THREADS=300 MAX_MEM_GB=3000
```

#### 2.1.3 Mate-pair scaffolding

The raw reads were processed using NextClip which filters out paired-end reads (reads which do not include the Nextera adaptor), and removes the adaptor from the mate-pair reads. We used NextClip version 1.3 which depends on R version 2.15.2, bwa version 0.6.2 and txlive version 1.2.2013. We used the following command for NextClip:

```
nextclip -d -m 30 -t 0 -n 600000000 -l output.log \
    -i mp_R1.fastq -j mp_R2.fastq -o nextclip_potato
```

The number of reads before processing was 44 864 446, which were divided into four categories based on whether the Nextera adaptor was found in the reads or not. The reads with adaptor were further filtered for sequences that were long enough (less than 25 bp), and total of 33 044 388 (73.65 %) were deemed usable. Total 11 173 807 647 (15.4 $\times$ ) was written for a file to be used in the scaffolding stage.

The scaffolding was done with SOAPdenovo version 2.4 (r240) with the following commands and the configuration file given in Supplemental File **S2**:

```
finalFusion -D -K 71 -s soap-discover.config -c assembly.fasta -g scaffolds
SOAPdenovo-127mer map -s soap-discover.config -g scaffolds
SOAPdenovo-127mer scaff -g scaffolds
```

## 2.2 PacBio

Python version 2.7.9 was used for running the PacBio software.

### 2.2.1 Falcon

We used Falcon version 0.3.0 and the following command line:

```
fc_run.py fc_run_ver.cfg
```

The configuration file is in **fc\_run\_ver.cfg** (Supplemental File **S6**) which needs the list of input files **input.fofn** (Supplemental File **S3**).

## 2.2.2 Canu

We used Canu version 1.0 with the following command line:

```
canu -d run2-18.2.16/ \
-p verrucosum \
errorRate=0.06 \
genomeSize=670m \
-pacbio-raw filtered_subreads.fastq \
useGrid=false
```

We initially used filtered subreads from the standard PacBio pipeline and then did a second assembly using subreads from the HGAP 3 pipeline (the filtered reads are the result of running the HGAP pipeline, below, up to the P\_PreAssemblerDagcon stage).

## 2.2.3 HGAP

We used HGAP 3 as part of smrtanalysis 2.3.0p5 and the following command line:

```
smrtpipe.py --params=params_v1.xml \
xml:input.xml > smrtpipe.log
```

params\_v1.xml (Supplemental File **S5**) contains the configuration for HGAP and input.xml (Supplemental File **S4**) lists the raw PacBio input files.

## 2.3 10x Genomics Supernova

To assemble the 10x Genomics library we used Supernova version 1.1.1 and the following commands:

```
supernova run --id=verrusocum5 \
--fastqs=potato/10x/supernova/fastq \
--sample=potato
supernova mktopput --asmdir=genome-assembly/10x/verrusocum5/outs/assembly \
--outprefix=pseudohap_ver_500 \
--style=pseudohap --minsize=500
```

The manufacturer recommends the use of 150 bp paired-end reads with a coverage of between 38 $\times$  to 56 $\times$ . Since our library was higher coverage, and with 250 bp reads, we generated subsamples of the read sets to simulate the use of 150 bp reads and coverages of 48 $\times$  and 52 $\times$ . The reads for each subsample were selected using a pseudorandom number generator with a known seed. Three different seeds were produced for each subsample to provide triplicates for testing. We found that in each case the subsampled libraries with shorter read lengths produced assemblies with significantly less contiguity. Our original assembly has an N50 of 2.38 Mbp, while the 52 $\times$  version has 2.15 Mbp and the 48 $\times$  version has 1.94 Mbp.

The subsample program can be found here:

<https://github.com/georgek/bio-tools/blob/master/fastq-subsample.cc>.

## 2.4 BioNano

We used BioNano Irys version 2.0 (<http://bionanogenomics.com/support/software-downloads/>), Python version 2.7.12 as part of Anaconda version 2.5.0, and Perl version 5.16.3.

#### 2.4.1 Optical map assembly

First we built a BioNano *de novo* optical map assembly using the following command line:

```
python bionano/scripts_UV2K/pipelineCL.py \
-U -d -T 32 -j 16 -N 4 -i 5 -w \
-a potato_arguments_v1.xml \
-t bionano/tools \
-l output \
-b Molecules.bnx \
-C clusterArguments.xml
```

XML configuration is given in Supplemental Files **S7** and **S8**.

#### 2.4.2 Scaffolding

Scaffolding is performed using the optical map assembly and one of our *S. verrucosum* genome assemblies.

```
perl bionano/scripts_SLURM/HybridScaffold/hybridScaffold.pl \
-n supernova.fasta \
-b EXP_REFINEFINAL1.cmap \
-c potato_hybrid_scaffold_parameters.xml \
-o output \
-B 1 -N 1 \
-r bionano/tools/RefAligner
```

`EXP_REFINEFINAL1.cmap` is the optical map assembly from the previous section. `supernova.fasta` is our Supernova assembly (see Section 4.4). `potato_hybrid_scaffold_parameters.xml` is the configuration file (Supplemental File **S9**).

### 2.5 Dovetail

Dovetail scaffolding was performed for the `discover-mp` and `falcon` assemblies by Dovetail Ltd. using HiRise.

### 2.6 BAC assembly

#### 2.6.1 Illumina

The BAC reads were filtered to remove any phiX, *E. coli*, and the pIndigo5 BAC vector sequence. The filtered reads were then assembled using DISCOVAR *de novo* and the assembled contigs were filtered to remove any residual pieces of BAC vector and contigs under 500 bp in length.

#### 2.6.2 PacBio

The BAC pools were assembled with HGAP as part of smrtanalysis 2.3.0p5 with standard settings.

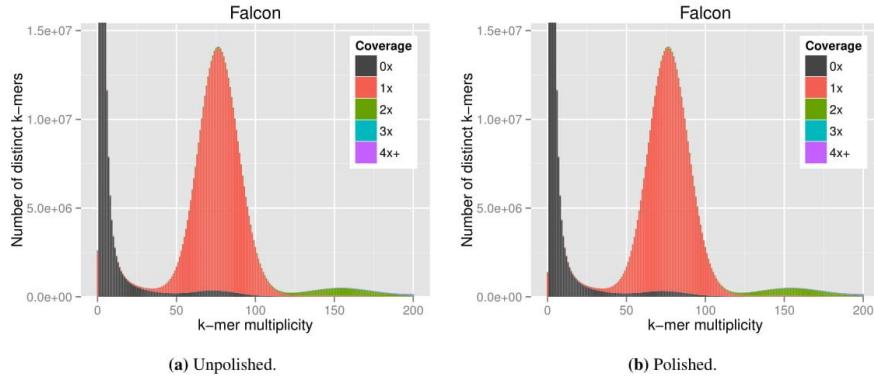


Figure S3.1: KAT plots for Falcon assembly with and without polishing using Pilon.

### 3 Evaluation

All of the code used to perform the evaluation is available on GitHub in the [georgek/potato-figures](https://github.com/georgek/potato-figures) repository. The complete process from raw data to figures can be found at <https://georgek.github.io/potato-figures/figures.html>.

#### 3.1 K-mer content

In Section we used KAT to see the *k*-mer content of each assembly. Since we were comparing the assemblies to the Illumina reads, we polished the Falcon assembly using Pilon to reduce the difference due to the sequencing platform. Figure S3.1 shows the KAT plots for Falcon with and without the polishing step for comparison.

#### 3.2 Synteny

The complete set of mummer plots showing alignments to the *S. tuberosum* reference v4.04 for the Falcon assembly, DISCOVAR assembly, and Supernova assembly are shown in Figures S3.2, S3.3, and S3.4, respectively.

#### 3.3 Gene content

#### 3.4 BUSCO

BUSCO was run for each assembly using the following command (in this case for `discover-mp-dt-bn`):

```
python BUSCO.py -i discover-mp-dt-bn.fasta -o discover_plant \
-l embryophyta_odb9 -m genome -c 16 -sp arabidopsis
```

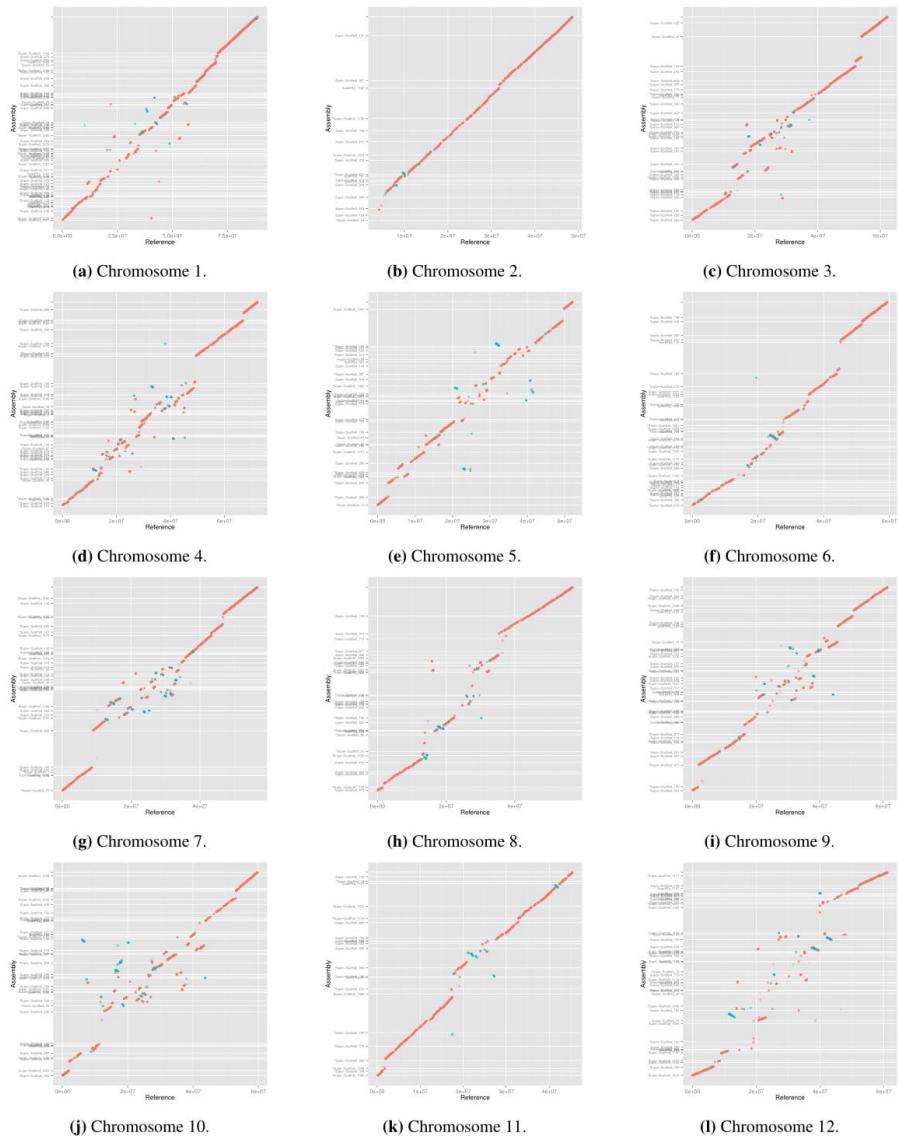
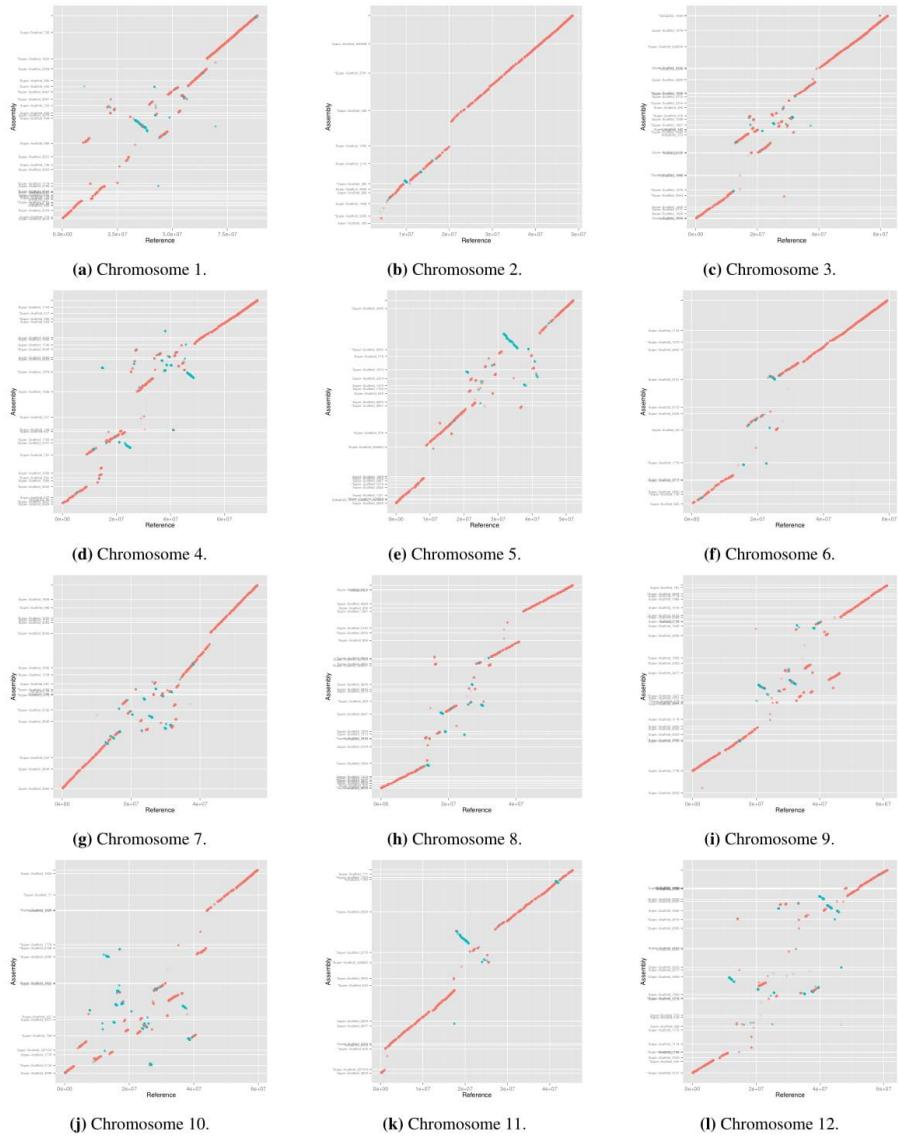


Figure S3.2: Mummer plots for Falcon assembly against the *S. tuberosum* reference.



**Figure S3.3:** Mummer plots for Discovar assembly against the *S. tuberosum* reference.

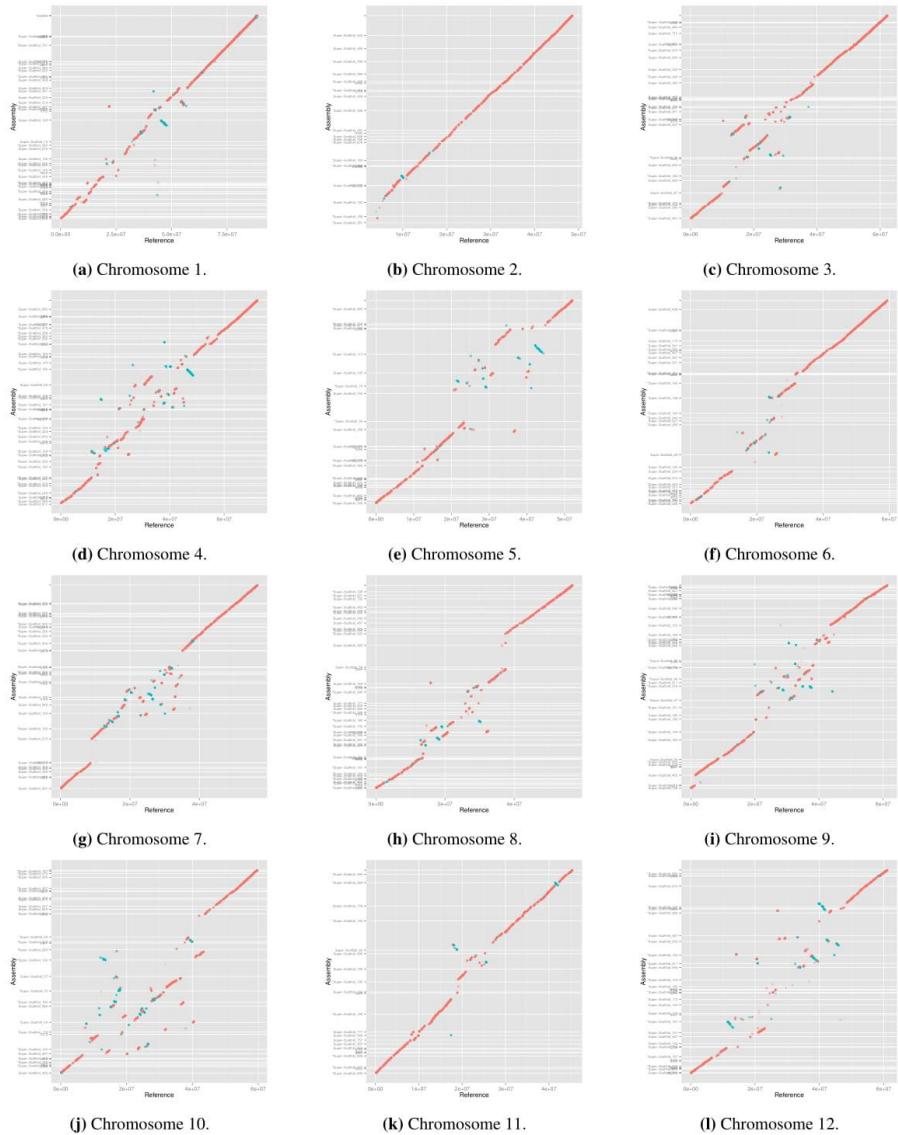


Figure S3.4: Mummer plots for Supernova assembly against the *S. tuberosum* reference.

## 4 Data availability

All read data is available in the European Nucleotide Archive (ENA) under the project PRJEB20860. The following subsections give the individual accession numbers for runs or analyses.

### 4.1 Short reads

DISCOVAR reads: ERR1988833. Dovetail reads: ERR1988846. Mate-pair reads: ERR1988848, ERR1988849. TALL reads: ERR1988850, ERR1988851. 10x reads: ERR1990231, ERR1990232, ERR1990233, ERR1990234.

### 4.2 Long reads

PacBio reads were generated in the following runs, each containing one or more SMRT cell: ERR1988882, ERR1988883, ERR1988884, ERR1988885, ERR1988886, ERR1988887, ERR1988888, ERR1989452, ERR1989453, ERR1989454, ERR1989455, ERR1989456, ERR1989457.

### 4.3 Optical map

The BioNano optical map is available in analysis ERZ438893.

### 4.4 Assemblies

- The Illumina BAC assemblies are in:

Accession	Assembly	Accession	Assembly
PRJEB21134	BAC 13	PRJEB21144	BAC 45
PRJEB21135	BAC 14	PRJEB21145	BAC 47
PRJEB21136	BAC 21	PRJEB21146	BAC 49
PRJEB21137	BAC 22	PRJEB21147	BAC 53
PRJEB21138	BAC 23	PRJEB21148	BAC 63
PRJEB21139	BAC 28	PRJEB21149	BAC 66
PRJEB21140	BAC 34	PRJEB21150	BAC 71
PRJEB21141	BAC 41	PRJEB21151	BAC 74
PRJEB21142	BAC 42	PRJEB21152	BAC 84
PRJEB21143	BAC 43	PRJEB21153	BAC 93

- The PacBio assembly of BAC 22 is in PRJEB21154;

- The whole genome assemblies are in:

Accession	Assembly	Accession	Assembly
PRJEB21112	10x-asm	PRJEB21122	falcon-asm
PRJEB21113	abyss113-asm	PRJEB21123	hgap-asm
PRJEB21114	abyss113-mp-asm	PRJEB21124	hgap-mp-asm
PRJEB21115	abyss77-asm	PRJEB21125	10x-bn-asm
PRJEB21116	abyss77-mp-asm	PRJEB21126	canu-bn-asm
PRJEB21117	canu-asm	PRJEB21127	discover-mp-dt-bn-asm
PRJEB21118	discover-contig-asm	PRJEB21128	falcon-bn-asm
PRJEB21119	discover-mp-dt-asm	PRJEB21129	falcon-dt-bn-asm
PRJEB21120	discover-mp-asm	PRJEB21130	hgap-bn-asm
PRJEB21121	falcon-dt-asm		

## 5 Authors' contributions

GB, ELG, IH, and GW prepared the sample. MDC, GK, and PP designed the analysis. DB, GB, ELG, MG, DH, IH, AL, and IM constructed libraries and performed sequencing. GK and PP made the assemblies and GK, ELG, and PP performed the evaluation. MDC, GK, GB, ELG and PP wrote and prepared the manuscript. All authors read and approved the final manuscript.

## 6 File list

- S1** `pair_end_potato.xml`. Input file for RAMPART used to automate ABySS assemblies,
- S2** `soap-discovar.config`. Configuration file for SOAPdenovo.
- S3** `input.fofn`. Raw input for PacBio assembly.
- S4** `input.xml`. Raw input for PacBio assembly in XML format.
- S5** `params_v1.xml`. Configuration for HGAP assembler.
- S6** `fc_run_ver.cfg`. Configuration for Falcon assembler.
- S7** `clusterArguments.xml`. Cluster arguments for BioNano *de novo* optical map assembly.
- S8** `potato_arguments_v1.xml`. Configuration for BioNano optical map assembly for *S. verrucosum*.
- S9** `potato_hybrid_scaffold_parameters.xml`. Configuration file for BioNano scaffolding.

## References

- [1] D. Mapleson, N. Drou and D. Swarbreck. "RAMPART: a workflow management system for *de novo* genome assembly". In: *Bioinformatics* 31.11 (2015), pp. 1824–1826. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv056.

# Independent assessment and improvement of wheat genome assemblies using Fosill jumping libraries.

Fu-Hao Lu<sup>1\*</sup>, Neil McKenzie<sup>1\*</sup>, George Kettleborough<sup>2</sup>, Darren Heavens<sup>2</sup>, Matthew D Clark<sup>2</sup>,  
Michael W Bevan<sup>+1</sup>

<sup>1</sup>John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK

<sup>2</sup>The Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, UK

\*Joint first Authors

<sup>+</sup>Corresponding Author

## Abstract

### Background

The accurate sequencing and assembly of very large, often polyploid, genomes remain a challenging task, limiting long-range sequence information and phased sequence variation for applications such as plant breeding. The 15 Gb hexaploid bread wheat genome has been particularly challenging to sequence, and several contending approaches recently generated accurate long-range assemblies. Understanding errors in these assemblies is important for optimising future sequencing and assembly approaches and for comparative genomics.

### Results

Here we use a Fosill 38 Kb jumping library to assess medium and longer-range order of different publicly available wheat genome assemblies. Modifications to the Fosill protocol generated longer Illumina sequences and enabled comprehensive genome coverage. Analyses of two independent BAC-based chromosome-scale assemblies, two independent Illumina whole genome shotgun assemblies, and a hybrid long read (PacBio) and short read (Illumina) assembly were carried out. We revealed a variety of discrepancies using Fosill mate-pair mapping and validated several of each class. In addition, Fosill mate-pairs were used to scaffold a whole genome Illumina assembly, leading to a three-fold increase in N50 values.

### Conclusions

Our analyses, using an independent means to validate different wheat genome assemblies, show that whole genome shotgun assemblies are significantly more accurate by all measures compared to BAC-based chromosome-scale assemblies. Although current whole genome assemblies are reasonably accurate and useful, additional steps will be needed for the rapid, cost-effective and complete sequencing and assembly of wheat genomes.

## Keywords

Wheat genome/assembly methods/Fosills/long-range genome assembly/Illumina/PacBio

## Background

Genome sequence assemblies are key foundations for many biological studies, therefore the accuracy of sequence assemblies and their long-range order is a fundamental prerequisite for their use. Multiple types of differences in the information content of DNA molecules, from single nucleotide polymorphisms (SNPs) to large-scale structural variation (SV), form part of natural genetic variation that can cause phenotypic variation [1,2]. Distinguishing such *bona fide* variation from apparent variation generated by sequence and assembly methods is therefore a critically important activity in genomics.

Sequence assemblies are generally incomplete and contain multiple types of errors, reducing their information content. Gaps in assemblies can occur where no sequence reads were generated for that region, but this is now increasingly unlikely given the very deep coverage achievable by short read sequencing, improved sequence chemistry, and template preparation methods that avoids bias, such as that introduced by PCR [3]. Closely related repetitive DNA sequences can lead to incorrect joins in assemblies, or to an unresolvable assembly graph that breaks an assembly. Assemblies can be either joined or broken inadvertently by closely related or polymorphic sequences that cause alternate, multiple, or collapsed assemblies, for example in assemblies of polyploid organisms [4]. Errors and incompleteness can obscure important genomic information such as the correct order (phasing) of sequence variants.

A broad spectrum of sequence and assembly artefacts can be distinguished from natural sequence variation, structural variants identified, and sequence variation phased, using long-range sequence information. Fosmid clones, which have large precisely-sized inserts due to lambda phage packaging, have been sequenced to close gaps in human genome assemblies [5] and to establish longer-range sequence haplotypes [6,7]. Earlier uses of fosmid clones for bulk sequencing [8,9] were supplanted by sequencing libraries of larger insert Bacterial Artificial Chromosome (BAC) clones [10]. Sequences of long single molecules generated by PacBio Single Molecule Real Time (SMRT) and Nanopore technologies are increasingly used for defining long-range gene order and for *de novo* genome assembly [11,12]. Linked read technologies such as 10X Genomics reads are also beginning to be widely used for long-range ordering of scaffolds assembled from short reads, and for identifying structural variation [13]. SMRT is also often used in hybrid approaches that utilise Illumina assemblies to improve the accuracy of single molecule reads [14]. Thus, there are several approaches available for generating and assessing the long-range integrity of genome assemblies.

These improvements in sequence read length and assembly procedures are enabling the creation of genomic resources for even the largest genomes. These include the genomes of grasses and gymnosperm trees, which have massive repetitive DNA tracts comprising about 80% of their genomes. The 22 Gb genome of loblolly pine (*Pinus taeda*), initially assembled from Illumina paired end sequence reads [15], has been significantly improved using SMRT sequencing [16]. 10X Genomics linked reads were used to generate an eight-fold increase in scaffold NG50 sizes of sugar pine (*P. lambertiana*) genome assemblies to nearly 2 Mb [17]. Bread wheat (*Triticum aestivum*) has a large 15 Gb allohexaploid genome consisting of three closely-related and separately maintained A, B and D genomes [18]. Assemblies of Illumina whole genome shotgun sequences were assigned to their correct genome [19], but the assembly was still highly fragmented. A near-complete and highly contiguous assembly of Illumina paired-end and mate-pair reads from wild emmer wheat (*Triticum turgidum*), a tetraploid progenitor of bread wheat has also recently been published [20]. Finally, long SMRT sequence reads integrated with Illumina sequence coverage increased the size and contiguity of maize [21], a diploid wheat progenitor [14] and hexaploid wheat genome assemblies [22]. It is not known how these assemblies differ in error types which may obscure true genetic variation.

Generating and assessing accurate long-range genome information from the large genomes of crop plants is necessary for identifying haplotypes used by breeders and for mapping large-scale structural variation contributing to agronomic performance [23]. Therefore, assessing the fidelity of longer-range genome assemblies is important for their applications to crop improvement. Here we use mate-paired sequences of wheat fosmid clones to assess three different wheat whole genome assemblies and two BAC-based wheat chromosome assemblies. Our analyses have identified a range of assembly issues and may help to identify optimal approaches to wheat genome assembly. Integrating fosmid end-sequences into scaffolds also increased scaffold sizes of both fragmentary and more contiguous assemblies.

## Results

### Creating and assessing a wheat fosmid clone library

Fosmid clone libraries have been used to assess genome assemblies and identify structural variation in human [24,25] and pine genomes [16]. Fosmids are used because DNA is cloned in a precise range of  $38\pm3$  Kb by efficient packaging of phage lambda and cohesive end circularisation. Fosmid clone inserts have been converted to Illumina sequencing templates to generate 38 Kb mate-pair “jumping libraries” for improving the assemblies of the mouse

genome [26]. In genomes with extensive tracts of closely-related repeats that have been challenging to assemble, fosmid jumping libraries could provide an independent means to both assess the fidelity of assemblies and to improve them. Several different hexaploid and tetraploid wheat chromosome and whole genome assemblies have been generated using different approaches [19,20,22,27], and assessing these could provide information needed to identify optimal approaches to wheat genome assembly.

To explore the potential of fosmid jumping libraries for assessing and improving wheat genome assemblies, we first carried out a simulation, using 38 Kb mate-pairs, of whole genome shotgun assemblies of three long 3.5 - 4.1 Mb scaffolds of wheat chromosome 3B generated by sequencing a manually curated physical map of BACs [27]. Simulation settings used different paired-end distances, read lengths and sequence coverage on the chromosome 3B scaffolds to assess how 38 Kb mate-paired reads, read-depth and read-length contributed to re-assembly (Additional File 1). Addition of 38 Kb mate-pair reads was required for accurate and complete reconstruction of all three scaffolds under these conditions. Paired-end read lengths between 100 – 250 bp were then assessed using a common combination of mate pair distances and sequence coverage. Reads of over 200 bp were required for consistent re-assembly of all three scaffolds. Finally, simulation of sequence coverage of 38 kb mate pair reads of length 250 bp showed that consistent re-assembly of all three scaffolds required sequence coverage of approximately 0.75x (Additional File 1). Taken together, these simulations showed that 38 Kb paired-end 250 bp reads with a sequence coverage of approximately 0.75x (>50x physical coverage) could be used to guide and assess assemblies of the wheat genome.

The Fosill vector system was developed for converting fosmid clones to Illumina paired-end read templates [28]. We modified this Fosill conversion protocol to generate long paired-end 250 bp Illumina reads, to maximise library complexity, and to minimise clonal- and PCR-based amplification bias. Both of these modifications were required to maximise unique matches of paired-end reads to the highly repetitive polyploid wheat genome, and to maximise sequence coverage of the large genome. Additional File 2 describes the modified protocols for library preparation and paired-end read analyses. These involved increasing the time of nick-translation to between 50-60 minutes on ice to generate inverse PCR products with a peak size distribution of 785-860 bp (Additional File 2). This minimised overlap of 250bp reads from either end of the PCR product. For each pool of 5-10M Fosill clones, a small sample of the circularised template was amplified for up to 16 cycles, and the minimum number of cycles required (generally 12-13) to generate sufficient template for sequencing was estimated.

Table 1 in Additional File 2 summarises the Fosill libraries produced and the paired-end sequences generated from them. Paired-end reads that overlapped each other on the template were discarded (2.61%), while 11.91% of the raw reads were excluded after vector/adapter sequence and quality trimming. The final number of 576 M paired-end sequences (85.5% of the total reads) were generated from 54.61 M Fosill clones (124x physical coverage). These were then mapped to the chromosome 3B pseudomolecule to measure the insert size distribution of the libraries (Additional File 3). Figure 1A shows the size distribution of 588,268 mapped read pairs, which had a mean estimated insert size of approximately 37,725 bp. This is the expected insert size range in the Fosill4 vector [28], and demonstrated successful size selection during packaging. Figure 1B shows the distribution of mate-pairs mapped in 100 kb windows across chromosome 3B BAC pseudomolecule. Reads with a depth of  $\leq 5$  covered 494 Mb of the total 833 Mb chromosome, accounting for 59% of the chromosome sequence. Their even distribution across the pseudomolecule indicated that the libraries were representative of the entire chromosome. There were approximately 30 distinct peaks of greatly increased read-depth (Figure 1B) in the 100 kb windows across chromosome 3B. These probably correspond to mate-pairs spanning approximately 40 Kb repeated regions common to multiple genomic loci. These reads accounts for 80% of the alignments but covered only 4.3% of chromosome 3B. For all subsequent analyses only Fosill mate-pairs of sequence depth  $\leq 5$  were used. Finally, reads that mapped to multiple locations, which lacked a paired read in the expected genomic location, or which had a paired read in the incorrect orientation, were removed.

#### **Using Fosill mate pairs to assess wheat chromosome sequence assemblies**

The even representation of long mate-paired reads across the chromosome 3B pseudomolecule indicated their suitability for assessing wheat sequence assemblies and for making new joins in wheat sequence scaffolds. For assessing assemblies, a windows-based filter was developed to identify sets of  $\geq 5$  unique neighbouring Fosill sequence reads in a “driver” window of  $< 10$  kb and their  $\geq 5$  mate-pair reads in a “follower” window of  $< 20$  kb on chromosome and genome assemblies. The vast proportion of mate-paired reads fell within this distance distribution (Additional File 3, Figure 1). Using this approach to map Fosill reads, we aimed to identify different types of paired-end matches to genome sequence assemblies. These can be used to identify genome assemblies consistent with the 37.7 kb mate-pair distances  $\pm$  sd, to identify possible new joins between assemblies, and to identify different

types of inconsistencies in the range of current publicly available wheat genome assemblies. Figure 2A illustrates the possible types of Fosill paired-end matches to assemblies. Tables 1A-1E show the outcomes of mapping Fosill paired-end reads to BAC-based bread wheat chromosome assemblies of chromosome 3B [27], TGACv1 Illumina assemblies of 3B [19], the Triticum 3.0 whole genome assembly of Pacbio SMRT and Illumina sequences of chromosomes 3B and 3DL [22], and DeNovoMagic assemblies of Illumina sequences from wild emmer wheat (WEW) chromosome 3B [20]. We also assessed an assembly of hexaploid wheat chromosome 3DL from sequenced BACs in a minimal tiling path using an automated pipeline (Additional File 4). A set of larger whole genome assemblies of the TGACv1 Illumina wheat genome were also assessed. These assemblies represent diverse approaches to sequencing wheat chromosomes and chromosome arms, including manually curated and automated BAC-based assemblies, two different Illumina-based assembly methods, and a combined Illumina and Pacific Biosciences SMRT assembly of wheat chromosomes.

Variation in Fosill insert sizes were consistent across the TGACv1, Triticum 3.0 and DeNovo Magic whole genome assemblies and the 3DL BAC assemblies. In contrast, chromosome 3B BAC assemblies had a higher variation of insert sizes (Table 1A). This may be due to a higher proportion of mis-assemblies in the 3B BAC assembly that could have introduced or removed small tracts of sequences, and possibly due to the use of a mixture of 454 and Illumina sequences. This variation in Fosill mate-pair matches did not contribute to assessment of assembly accuracy. The accuracy of assemblies was estimated by counting the bases included in correctly-sized windows (mean insert size  $\pm$ sd) of Fosill mate-pair reads, and by the proportion of assemblies/scaffolds that were fully consistent with Fosill mate-pair windows along their length. The un-edited BAC-based scaffolds of chromosome 3DL were the least accurate, with only 17% of the assemblies covered with consistent fossil mate-pair matches, and 57% of the sequence included under consistent mate-pair matches (Table 1A). The 3B BAC assemblies, which have been extensively manually edited, were considerably more accurate, with 66% consistent assemblies and 85% of sequences in consistent windows. Looking at the TGACv1 3B assemblies, 61% of scaffolds were consistent and 80% of sequences were contained within consistent Fosill windows. In comparison, larger TGACv1 assemblies from the whole genome were all consistent with mate-pair windows and 99% of the sequences were in consistent windows. The differences with TGACv1 3B assemblies are most likely due to many shorter assemblies being included in the 3B assembly that limit the potential for 37 Kb mate-pair mapping, for example, there will be a low proportion of matches at the ends of assemblies. The Triticum 3.0 WGS assembly of 3B had 92% consistent assemblies, and 90% of sequences within consistent Fosill windows. Similarly, the Triticum 3.0 WGS assembly of chromosome 3DL had 99.5% assemblies and 80% of sequences in

consistent windows. The DeNovo Magic WGS assembly of *T. turgidum* 3B contained 99.6% of sequences in consistent Fosill windows. As these assemblies were integrated into a single pseudomolecule the measure of the number of correct scaffolds was 100%.

Four different classes of discrepancies that may be due to assembly problems were assessed using Fosill mate pair mapping to assemblies: failed scaffolding, in which scaffolds had matches to only one end of Fosill end-sequences, and which may need to be broken; orientation errors in which the direction of one region of a scaffold is consistently reversed with respect to flanking regions; insertions, in which the span of Fosill mate-pair matches is greater than expected; and deletions, in which mate-pair spans are less than expected. These results are summarised in Tables 1B-1E. Of these potential error types, the most frequent were the potential erroneous joining of assemblies. These were highest in the BAC assemblies, and lowest in the DenovoMAGIC assembly of 3B. An example of this is shown in Figure 1B, where two BAC-based scaffolds were assembled at either end of chromosome 3B. Fosill mapping evidence, supported by TGACv1 assemblies, showed that the two scaffolds can be merged in opposite orientation to that originally assembled. Figure 2C reveals a 12 Kb deletion in a TGACv1 assembly that was due to a missing tandem duplication of the repeat, as validated by comparison with the Triticum 3.0 assembly. An aberrant insertion in a TGACv1 scaffold identified by Fosill mate-pair mapping was also validated by comparison with the Triticum 3.0 assembly (Figure 2D).

The TGACv1 large assemblies have relatively low numbers of mis-assemblies. The Triticum 3.0 assemblies of both 3B and 3DL had a consistently large number of potential mis-assemblies, with about 400-500 per chromosome or chromosome arm, affecting about 10 Mb of sequence region. Potential deletion errors, in which assemblies may be missing sequences, were most frequent in the BAC assembly of chromosome 3B, and were also the most frequent type of error in the DenovoMAGIC assembly. Deletions were least frequent in the TGACv1 whole genome assembly. Potential erroneous insertions were less frequent than deletions, with the highest rates of both types of potential error in BAC-based assemblies. In general, potentially erroneous deletions were more common in all assemblies than insertions. Mis-orientations were the rarest potential errors and were most prevalent in manual assembled 3B BAC scaffolds and essentially absent from TGACv1 and Triticum 3.0 assemblies, but were more frequent in the DeNovo Magic 3B assembly.

#### **Using Fosill mate-pairs to create more contiguous assemblies**

The wheat Fosill library was also used to create new joins in different assemblies. Table 2A shows that Fosill mate-pair reads made 267 new links between 477 chromosome 3B BAC scaffolds. Where available, TGACv1 3B assemblies spanning the new links precisely (124 cases), supporting the new join, and no examples were found where the new Fosill joins linked the wrong neighbours or the wrong strand. We then applied the Fosill mate pairs to make new joins in chromosome 3B TGACv1 assemblies and chromosome 3DL BAC assemblies. Table 2B shows the total assembly sizes were increased, while the number of scaffolds in the assemblies was decreased, and the scaffold n10 more than doubled in size. This showed, as predicted by simulations (Figure1), that 38 kb mate-pair reads can make new links that substantially improve contiguity of both WGS and BAC-based assemblies. Where available, independent assemblies supported these new Fosill-based links. Figure 3 shows the distribution of scaffold sizes and numbers before and after Fosill linking on TGACv1 chromosome 3B (panel A) and chromosome 3DL BAC (Panel B) assemblies. Increases in the numbers of larger assemblies and concomitant reduction in the numbers of smaller assemblies after Fosill joining was more apparent in the chromosome 3B WGS scaffolds than in the 3DL BAC scaffolds. This may reflect the fewer joins needed in the less fragmentary 3B assembly (2,808 scaffolds) than the very fragmented 3DL assembly (23,433 scaffolds).

Based on these improvements in both BAC- based and WGS scaffold contiguity by integrating Fosill mate-pair reads, we re-scaffolded the complete TGACv1 WGS assembly of the wheat variety Chinese Spring 42 [19]. Figure 4 and Supplemental File 1 show the scaffold sizes of each chromosome arm before and after integration of Fosill mate-pairs. Substantial increases in scaffold N50 of between 2.7 - 3.2-fold were achieved. The largest scaffolds increased in size between 1.5 - 3.2-fold, with the largest scaffold of 2.8 Mb on chromosome 3B.

## Discussion

Bread wheat is one of the three major cereals that we depend on for our nutrition, and generating accurate long-range assemblies is essential for new genomics-led approaches to crop improvement. However, its genome has been exceptionally challenging to sequence due to its polyploid composition of three closely-related large genomes, and extensive tracts of closely related repetitive sequences. Two strategies have been followed to deal with this genomic complexity: the first used BAC clones made from purified chromosomal DNA to reduce the complexity of chromosome-specific assemblies [27]; the second set of approaches uses different types of whole genome shotgun sequence technologies and assembly methods [19,20,22]. At this stage of wheat genome sequencing, when these complementary and contending approaches have been published, it is timely to assess the long-range accuracy

of these different assemblies. For this, we mapped precise 38 Kb Fosill long mate pair reads to measure errors in different assemblies of chromosome 3B and the long arm of chromosome 3DL. We also used these Fosill mate pair reads to increase genome contiguity.

In order to maximise the accuracy of Fosill mate-pair read mapping to the A, B or D genomes and to repetitive regions of the hexaploid wheat genome, we modified the template conversion protocol of the Fosill 4 vector system [28] to generate longer paired 250 bp Illumina sequence reads. Nick-translation reactions to extend Nb.BbvCI nicks were optimised to generate an Illumina sequencing template between 750 - 1,000 bp. PCR amplification of re-circularised products was optimised to reduce amplification to the minimum required for efficient sequencing of a large library. Overall, 576.5M read pairs were generated from 55.1M clones (Additional File 2). When reads were mapped to chromosome 3B sequence assemblies [29], a consistent size distribution around 37.7 kb was observed (Figure 1A), demonstrating correct packaging and processing. Read depth varied several thousand-fold along chromosome 3B, likely due to matches of read-pairs to highly repetitive regions from across the genome. Consequently, only reads with depth  $\leq 5$  were used. Using this filter, we obtained sequence coverage of nearly 60% of the 833 Mb BAC-based chromosome 3B assembly. Simulations indicated that 0.75x sequence coverage of paired-end 250 bp reads was effective in creating long-range assemblies of wheat (Additional File 1), therefore we used Fosill read mapping for subsequent analyses.

Fosill reads were mapped to different assemblies of chromosome 3B and the long arm of chromosome 3D in order to compare the full range of current publicly available wheat genome assemblies. Several types of inconsistencies spanning a wide range of scales have been detected by mapping long-range mate-pairs to human genome assemblies [24,30]. Tables 1A-1E) show the types of inconsistencies detected in wheat assemblies using this approach. Looking first at the proportion of bases in different assemblies that were fully consistent with mapped 38 Kb mate-pair reads (Table 1A), the DenovoMAGIC Illumina-based assembly and the SMRT long-read Triticum 3.0 had respectively 99.6% and 90% of bases in consistent Fosill windows. The manually curated BAC-based assembly of 3B had 85% of consistent assembled sequences, while the TGACv1 3B assembly had 80% of assembled sequence in consistent windows, while the larger TGACv1 assemblies were 99% consistent. This difference may reflect the more fragmentary state of TGACv1 assemblies. The non-curated BAC assembly of chromosome 3DL was the least accurate according to this measure, with only 57% of bases in consistent windows. These data demonstrate both the superior accuracy of *de novo* whole genome sequencing strategies that incorporate deep and long 250 bp Illumina paired-end and

mate-pair sequencing, and the relative accuracy of long-range assemblies generated by mate-pair assembly strategies [19,20,22], compared to BAC-based strategies ([27].

The most frequent type of inconsistency identified by Fosill mapping was the potential incorrect joining of assemblies (Table 1B). Illumina strategies produced the fewest incorrect joins, while BAC-based assemblies produced the most. Interestingly, assemblies of both 3B and 3DL made from PacBio SMRT reads combined with 150 bp Illumina paired end reads (forming mega-reads) [22] had more possible assembly issues than Illumina- only assemblies. While a more complete assembly and relatively long assemblies were achieved from SMRT sequences, these assembly issues suggest that reads longer than 10 kb, or including long Illumina mate-pair libraries, could further improve SMRT-based assemblies. Assembly methods may also need further optimisation to utilize fully the potential of SMRT long reads. Furthermore, integrating long 250 bp Illumina reads into mega-reads may improve assemblies by distinguishing very closely related sequences, such as repeat regions from homoeologous chromosomes.

Potential deletion events were also quite common in all assemblies, and were the most common inconsistencies detected in DenovoMAGIC assemblies of 3B. The sizes of these events are not known precisely, but they have a minimum size of 12 Kb (Table 1E). These probably arise from missing tracts of near-identical sequence in assemblies. Similarly, potential insertions may arise from the incorrect integration of near-identical sequences into assemblies. The observation that potential deletions are more frequent than potential insertions suggests that all WGS-alone assembly strategies could achieve more complete assemblies of the wheat genome such as that achieved using PacBio SMRT sequence assemblies. Finally, potential mis-orientations/inversions of assemblies are more common in the DevoMAGIC assembly of 3B than the other whole- genome assemblies. Although this approach has yet to be fully described, mis-orientations may reflect more relaxed criteria for linking scaffolds than related Illumina-based assembly and scaffolding approaches ([19].

How much more accurate can the best current assemblies of bread wheat and wild emmer wheat be, judging by their assemblies of chromosome 3B? Fosmid end-mapping to 2005 versions of human genome assemblies [24] identified 297 longer range discrepancies in the 3.2 Gb genome. Scaling from chromosome 3B (0.8 Gb) with 127 potential inconsistencies, our analyses predict 480 discrepancies per 3Gb of wild emmer wheat genome assembly- roughly twice the error frequency of 2005 versions of the human genome. It is highly likely that a DenovoMAGIC version of the hexaploid bread wheat genome will achieve similar high levels of accuracy and coverage.

Three-fold increases in the scaffold N50 sizes of the TGACv1 whole genome assembly were achieved by an additional scaffolding step using Fosill mate-pairs. In addition to making a more useful genomic resource, this additional scaffolding shows the relatively fragmentary but highly accurate TGACv1 assembly has the potential for substantial further improvement. Considering the urgent need to generate accurate long-range assemblies of multiple elite bread wheat genomes and wild progenitor species for crop improvement programmes, linked read technologies [31] and Nanopore long reads [32] provide promising new opportunities for efficient, cost-effective and open-source approaches to identifying of a wide range of structural and phased sequence variation in wheat genome assemblies.

## Methods

Detailed descriptions of experimental and computational procedures are shown in Additional Files. These describe simulation of 38 Kb mate-pair reads for assembly (Additional File 1), Production and sequencing of Fosill libraries (Additional File 2) and physical mapping and sequencing of BACs from chromosome 3DL (Additional File 3).

### General bioinformatics

All analytical pipelines have been deposited in GitHub, and relevant links are shown in the manuscript and Additional Files. Joinable read pairs from Illumina Miseq or HiSeq sequencing were removed using FLASH v1.2.11 [33]. Ligation adaptors in reads were trimmed off using CutAdapt v1.6 [34]. Sequencing primer sequences and low-quality sequences in reads were removed using Trimmomatic v0.32 (parameter) [35]. Then the resulting reads were evaluated using FastQC v1.2.11 [36].

Trimmed reads were further filtered using ReadCleaner4Scaffolding pipeline (<https://github.com/lufuhao/ReadCleaner4Scaffolding>). Both mates in each pair was mapped to chr3B BAC scaffolds using bowtie v1.0.1 [37]. And then the Picard MarkDuplicates (v1.108, <http://broadinstitute.github.io/picard>) was used to remove the duplicates as single reads. A read depth threshold was used to remove the repeat-like reads by plotting the summary of output from samtools depth, and the reads mapped to those regions with higher depth were not used for scaffolding. The remaining reads were subjected to removal again as pairs.

Those reads mapped to multiple positions, whose mates were not mapped, or had the wrong orientation, were removed. A window size filter was applied to identify sets of  $\geq 5$  neighbouring reads in sliding windows of less than 10 Kb that had all their mates in a following window of less than 20 kb. Variations of the expected distance between mate-pairs (average  $\pm$  sd) of approximately 3 sd was used to identify potential assembly discrepancies.

## Data Availability

Fosill mate-pair reads from Chinese Spring 42 in this study have been submitted to the EBI European Nucleotide Archive (ENA), and are available in study accession PRJEB23322. Chromosome 3DL BAC scaffolds are available in ENA study accession PRJEB23358.

## Declarations

The authors declare they have no competing interests.

## Funding

This work was supported by a Biological and Biotechnological Sciences Research Council (BBSRC) strategic LOLA award to MWB (BB/J00328X/1 and MDC (BB/J003743/1), The FP7 Triticeae Genome Project to MWB, and a BBSRC Institute Strategic Programme Grant (GEN) BB/P013511/1 to MWB. BBSRC Institute Strategic Programme Grant (BB/J004669/1) and Core Strategic Programme Grant (BB/CSP17270/1) also supported work at the Earlham Institute. Sequencing was delivered via the BBSRC National Capability in Genomics (BB/J010375/1) at the Earlham Institute and performed by members of the Genomics Pipelines Group.

## Authors' contributions.

MWB conceived and coordinated the project, and wrote the manuscript. F-HL planned and carried out bioinformatics analyses, NMCK constructed the Fosill libraries, GK and MDC sequenced chromosome 3DL BACs and managed sequence data, and DH managed all sequencing library production and Illumina sequencing.

## Acknowledgements

We are grateful to Louise Williams (Broad Institute) for Fosill vectors and detailed advice.

## References

1. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet.* 2003;33 Suppl:228–37. doi:10.1038/ng1090.
2. Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, et al. Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. *Proc Natl Acad Sci U S A.* 2011;108:10249–54. doi:10.1073/pnas.1107739108.
3. Weisenfeld NI, Yin S, Sharpe T, Lau B, Hegarty R, Holmes L, et al. Comprehensive variation discovery in single human genomes. *Nat Genet.* 2014;46:1350–5. doi:10.1038/ng.3121.
4. Chaisson MJP, Wilson RK, Eichler EE. Genetic variation and the *de novo* assembly of human genomes. *Nat Rev Genet.* 2015;16:627–40. doi:10.1038/nrg3933.
5. Bovee D, Zhou Y, Haugen E, Wu Z, Hayden HS, Gillett W, et al. Closing gaps in the human genome with fosmid resources generated from multiple individuals. *Nat Genet.* 2008;40:1:96–101. doi:10.1038/ng.2007.34.
6. Kitzman JO, Mackenzie AP, Adey A, Hiatt JB, Patwardhan RP, Sudmant PH, et al. Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat Biotechnol.* 2011;29:59–63. doi:10.1038/nbt.1740.
7. Duitama J, McEwen GK, Huebsch T, Palczewski S, Schulz S, Verstrepen K, et al. Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. *Nucleic Acids Res.* 2012;40 5:2041–53. doi:10.1093/nar/gkr1042.
8. Ammiraju JSS, Yu Y, Luo M, Kudrna D, Kim H, Goicoechea JL, et al. Random sheared fosmid library as a new genomic tool to accelerate complete finishing of rice (*Oryza sativa* spp. *Nipponbare*) genome sequence: sequencing of gap-specific fosmid clones uncovers new euchromatic portions of the genome. *Theor Appl Genet.* 2005;111:1596–607. doi:10.1007/s00122-005-0091-3.
9. Park T-H, Park B-S, Kim JA, Hong JK, Jin M, Seol Y-J, et al. Construction of random sheared fosmid library from Chinese cabbage and its use for *Brassica rapa* genome sequencing project. *J Genet Genomics.* 2011;38:47–53. doi:10.1016/j.jgg.2010.12.002.
10. Luo M-C, Gu YQ, You FM, Deal KR, Ma Y, Hu Y, et al. A 4-gigabase physical map unlocks the structure and evolution of the complex genome of *Aegilops tauschii*, the wheat D-genome progenitor. *Proc Natl Acad Sci U S A.* 2013;110 19:7940–5. doi:10.1073/pnas.1219082110.
11. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Meth.* 2016; 13 12:1050-4. doi:10.1038/nmeth.4035.
12. Jain M, Koren S, Quick J, Rand AC, Sasani TA, Tyson JR, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *BioRxiv* 2017; <https://dx.doi.org/10.1101/128835>

13. Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol.* 2016;34:303–11. doi:10.1038/nbt.3432.
14. Zimin AV, Puiu D, Luo M-C, Zhu T, Koren S, Marçais G, et al. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* 2017;27:5:787–92. doi:10.1101/gr.213405.116.
15. Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol.* 2014;15:3:R59. doi:10.1186/gb-2014-15-3-r59.
16. Zimin AV, Stevens KA, Crepeau MW, Puiu D, Wegrzyn JL, Yorke JA, et al. An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *GigaScience.* 2017;6:1:1–4. doi:10.1093/gigascience/giw016.
17. Crepeau MW, Langley CH, Stevens KA. From pine cones to read clouds: rescuffolding the megagenome of sugar pine (*Pinus lambertiana*). *G3: Genes, Genomes, Genetics.* 2017;7:1563–8. doi:10.1534/g3.117.040055.
18. Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, International Wheat Genome Sequencing Consortium, et al. Ancient hybridizations among the ancestral genomes of bread wheat. *Science.* 2014;345:1250092. doi:10.1126/science.1250092.
19. Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, et al. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Res.* 2017;27:5:885–96. doi:10.1101/gr.217117.116.
20. Avni R, Nave M, Barad O, Baruch K, Twardziok SO, Gundlach H, et al. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science.* 2017;357:6346:93–7. doi:10.1126/science.aan0032.
21. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, et al. Improved maize reference genome with single-molecule technologies. *Nature.* 2017;546:524–7. doi:10.1038/nature22971.
22. Zimin AV, Puiu D, Hall R, Kingan S, Clavijo B, Salzberg SL. The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *GigaScience.* 2017; <https://doi.org/10.1093/gigascience/gix097>
23. Ma J, Stiller J, Berkman PJ, Wei Y, Rogers J, Feuillet C, et al. Sequence-based analysis of translocations and inversions in bread wheat (*Triticum aestivum* L.). *PLoS ONE.* 2013;8:11:e79329. doi:10.1371/journal.pone.0079329.
24. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, et al. Fine-scale structural variation of the human genome. *Nature Genet.* 2005;37:727–32. doi:10.1038/ng1562.
25. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. *Nature.* 2008;453:7191:56–64. doi:10.1038/nature06862.
26. Williams LJS, Tabbaa DG, Li N, Berlin AM, Shea TP, MacCallum I, et al. Paired-end sequencing of Fosmid libraries by Illumina. *Genome Res.* 2012;22:11:2241–9.

doi:10.1101/gr.138925.112.

27. Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, et al. Structural and functional partitioning of bread wheat chromosome 3B. *Science*. 2014;345 6194:1249721. doi:10.1126/science.1249721.

30. Rasekh ME, Chiatante G, Miroballo M, Tang J, Ventura M, Amemiya CT, et al. Discovery of large genomic inversions using long range information. *BMC Genomics*. 2017;18 1:65. doi:10.1186/s12864-016-3444-1.

31. Hulse-Kemp AM, Maheshwari S, Stoffel K, Hill TA, Jaffe D, Williams S, et al. Reference quality assembly of the 3.5 Gb genome of *Capsicum annuum* from a single linked-read library. *bioRxiv*. 2017; <http://dx.doi.org/10.1101/152777>.

32. Schmidt MHW. Reconstructing the gigabase plant genome of *Solanum pennellii* using Nanopore sequencing. *bioRxiv*. 2017; <http://dx.doi.org/10.1101/129148>.

33. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*. 2011;27 21:2957–63. doi:10.1093/bioinformatics/btr507.

34. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. 2017; 17 1:10-12. doi:10.14806/ej.17.1.200.

35. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30 15:2114–20. doi:10.1093/bioinformatics/btu170.

36. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2011. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

37. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Meth*. 2012;9 4:357–9. doi:10.1038/nmeth.1923.

## Figure Legends

### **Figure 1. Determination of Fosill mate-pair distance distributions on chromosome 3B.**

**A.** 576 M quality controlled paired-end sequences were mapped to the chromosome 3B pseudomolecule. 588,268 read pairs were mapped and the insert sizes calculated. The mean insert size was 37,725 Kb.

**B.** Fosill mate-pairs were mapped in 100 Kb bins along chromosome 3B to assess the depth and evenness of coverage. Coverage was generally even across the entire chromosome, with approximately 30 very high copy peaks that are probably due to Fosill mate-pairs from highly related 40 Kb+ regions from across the genome. Most mate-pairs mapped to a depth of  $\leq 5$  and were used for subsequent analyses.

### **Figure 2. Using Fosill mate-pair matches to identify discrepancies in wheat chromosome and genome assemblies.**

**A.** The schematic describes different classes of matches of Fosill mate-pair sequences to wheat chromosome and genome assemblies. Consistent assemblies matched a span of  $\geq 5$  mate-pairs in a sliding 10 Kb “driver” window that matched their mate in a 20 Kb “follower” window at a distance of 37 Kb  $+$ / $-$  sd in the correct orientation. Where mate-pairs spanned more than 50 Kb (approximately 3 sd) this was construed to be due to an aberrant insertion in the underlying assembly. Spans  $< 25$  Kb (approximately 3 sd) were construed to be due to an aberrant deletion in the assembly. Mis-orientations of the mate-pairs indicated a mis-oriented assembly, and no span a mis-join in the assembly. New joins were also identified. Drawing not to scale.

**B.** An example of a mis-join of the BAC-based assembly of chromosome 3B. Two scaffolds, v443\_0362 and v443\_0787, were originally assembled at opposite ends of chromosome 3B 730 Mb apart. Matches to Fosills indicated that these two scaffolds could be re-assembled together with v443\_0362 in the opposite orientation. The Mummer plot shows that this join is supported by TGACv1 scaffold\_220633\_3B. Drawing not to scale.

**C.** An example of an aberrant deletion in TGACv1 scaffold 220602 on chromosome 3B. Assembly missed a duplicate copy of a 12 Kb repeat (represented by an arrow) that was identified as a discrepancy in Fosill mate-pair matches. Comparison to a Triticum 3.0 scaffold identifies the predicted missing copy of the repeat. Drawing not to scale.

**D.** An example of an aberrant insertion in TGAv1 scaffold 591781 on chromosome 7BS detected by Fosill mate-pair matches of  $> 50$  Kb. Comparison to the Triticum 3.0 assembly of the same regions identifies the mis-assembled insertion. Drawing not to scale.

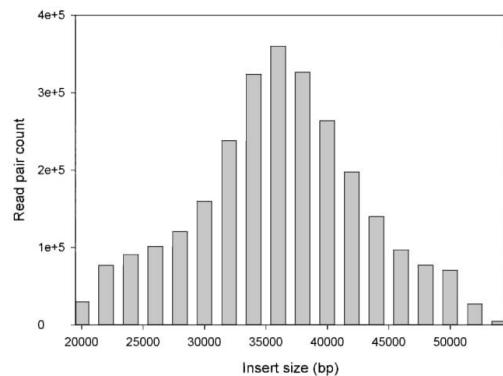
**Figure 3. Increasing assembly contiguity using Fosill matches.**

**A.** Fosill mate-pair reads were used to link scaffolds of TGACv1 Illumina assemblies from chromosome 3B. The distribution of scaffold lengths and the number of scaffolds in each size range is shown before (dark bars) and after (grey bars) Fosill scaffolding. The numbers of smaller scaffolds are reduced, and the numbers of larger scaffolds are increased, by Fosill scaffolding, showing successful further assembly.

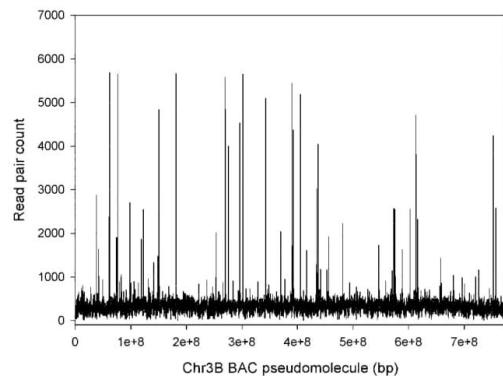
**B.** Fosill mate-pair reads were used to link scaffolds of BAC-based assemblies of chromosome 3DL. The distribution of scaffold lengths and the number of scaffolds in each size range is shown before (dark bars) and after (grey bars) Fosill scaffolding. The numbers of smaller scaffolds are reduced, and the numbers of larger scaffolds are increased, by Fosill scaffolding, showing successful further assembly.

**Figure 4. Fosill-mediated scaffolding of TGACv1 Illumina assemblies of the wheat genome.** The 21 chromosomes are shown with their scaffold N50 values before (black bars) and after (grey bars) Fosill-mediated scaffolding.

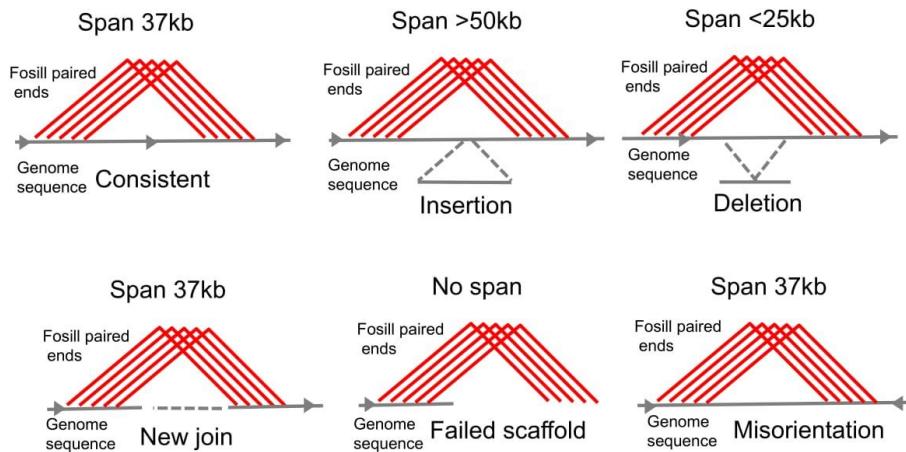
**Figure 1A.**



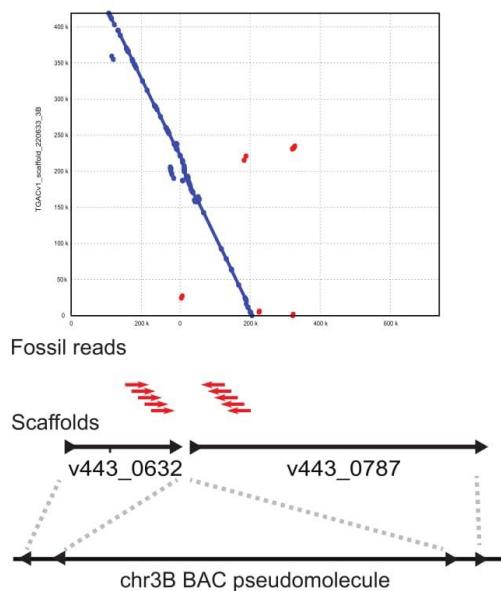
**Figure 1B.**



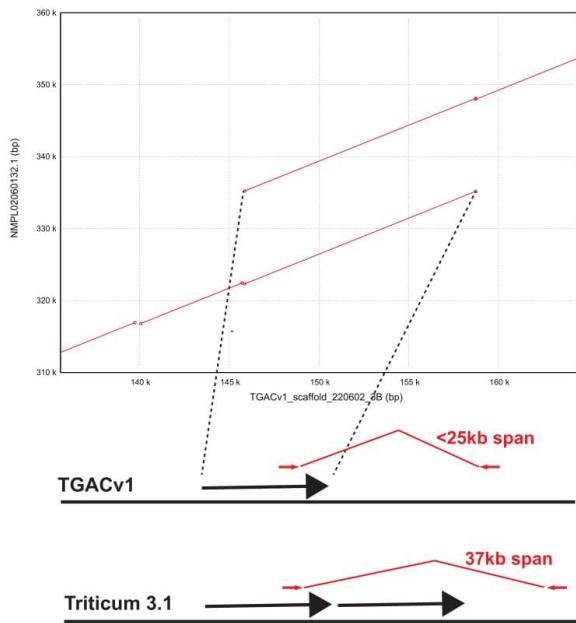
**Figure 2A.**



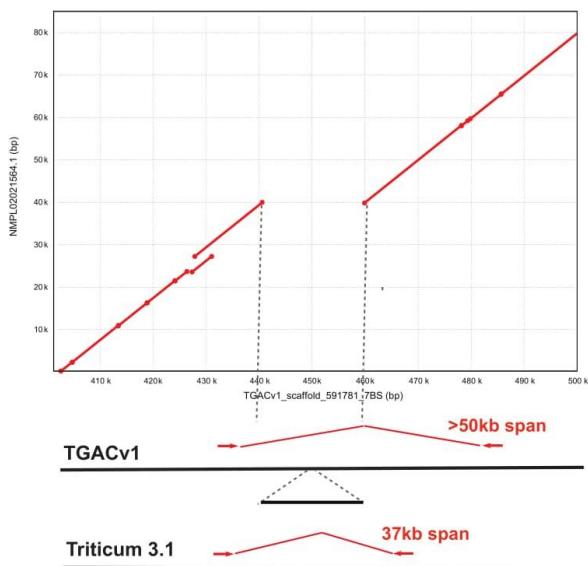
**Figure 2B.**



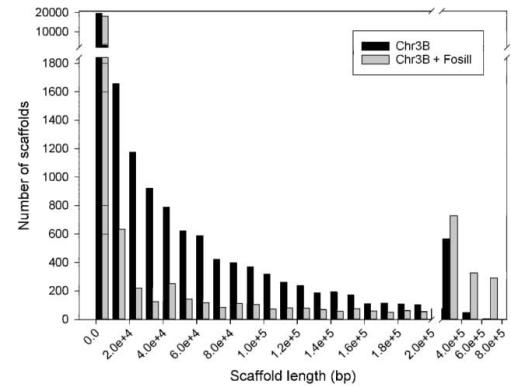
**Figure 2C**



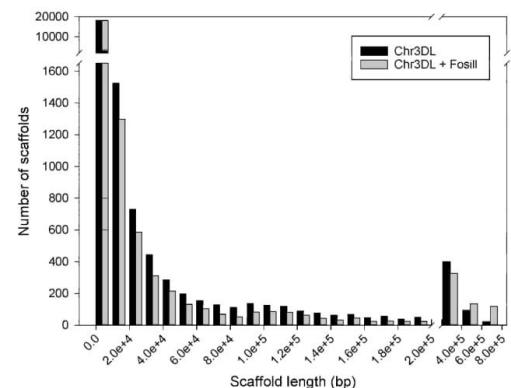
**Figure 2D**



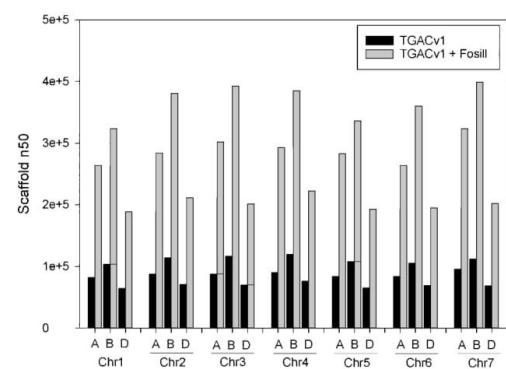
**Figure 3A.**

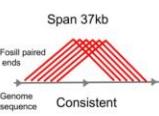


**Figure 3B.**

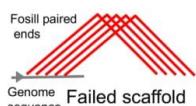


**Figure 4.**



 Span 37kb Fossil paired ends Genome sequence Consistent	Mean Fosill Insert Size bp	Std Dev	Assembly Size Mb	Scaffold N50 Kb	Total Scaffolds	Consistent Scaffolds	Consistent bases (% of total)
3B BAC Assembly	37,683	11,421	832	892	2,808	1859 (66%)	85%
3B TGAC v1 Assembly	37,177	4,608	789	116	29,000	17,730 (61%)	80%
3B Triticum 3.0	37,303	3,892	782	372	3,750	3,518 (92%)	90%
3B DenovoMAGIC2	37,661	3,968	841	6,373	271	271(100%)	99.6%
3DL BAC Assembly	37,254	5,160	453	154	23,433	4,040 (17%)	57%
3DL Triticum 3.0	37,247	3,887	409	279	2,703	2,691 (99.5%)	80%
All TGAC v1 (500Kb)	37,489	4,549	943	-	159	159 (100%)	99%

**Table 1A. Summary of Fosill mate-pair alignments to different publically-available assemblies of chromosome 3B and 3DL, and the TGACv1 whole genome assembly.** The consistency of mapping is shown according to the number of assemblies with consistent matches, and the percentage of bases included in consistent matches to Fosill mate-pairs.

No span	Scaffolding failures	Assemblies involved	Bases Involved Mb
 <p>Fossil paired ends Genome sequence</p>			
3B BAC Assembly	642	520	2.7
3B TGAC v1 Assembly	314	314	3.6
3B Triticum 3.0	517	499	11.4
3B DenovoMAGIC2	4	1	0.034
3DL BAC Assembly	536	491	17.23
3DL Triticum 3.0	444	442	9.7
All TGAC v1 (500Kb)	11	11	0.054

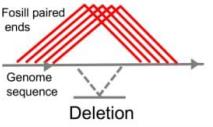
**Table 1B. Potential failed scaffolding**

Span 37kb	Mis-orientation errors	Assemblies involved	Bases Involved Mb
 <p>Genome sequence</p>			
3B BAC Assembly	92	78	2.7
3B TGAC v1 Assembly	6	6	0.094
3B Triticum 3.0	1	1	0.049
3B DenovoMAGIC2	21	1	1.06
3DL BAC Assembly	8	8	0.214
3DL Triticum 3.0	0	0	0
All TGAC v1 (500Kb)	1	1	0.007

**Table 1C. Potential mis-orientations in scaffolds**

Span >50kb	Insertion errors	Assemblies involved	Bases Involved Mb
 <p>Fossil paired ends Genome sequence</p>			
3B BAC Assembly	255	177	8.06
3B TGAC v1 Assembly	31	27	0.712
3B Triticum 3.0	88	66	1.7
3B DenovoMAGIC2	30	1	0.358
3DL BAC Assembly	78	63	2.225
3DL Triticum 3.0	18	17	0.396
All TGAC v1 (500Kb)	4	4	0.163

**Table 1D. Potential erroneous insertions in scaffolds**

Span <25kb	Deletion errors	Assemblies involved	Bases Involved Mb
			
3B BAC Assembly	626	381	15.97
3B TGAC v1 Assembly	129	116	2.58
3B Triticum 3.0	108	89	2.17
3B DenovoMAGIC2	72	1	0.366
3DL BAC	58	53	1.09
3DL Triticum 3.0	41	36	0.758
All TGAC v1 (500Kb)	13	9	0.307

**Table 1E. Potential erroneous deletions in scaffolds**

267 new links	Strand	Validated by TGACv1
37 links ≤40kb on pseudomolecule	20 correct strands	12
	17 reverse strands	7
147 links >40kb on pseudomolecule	73 correct strands	31
	74 reverse strands	38
83 links between scaffolds not assigned to the pseudomolecule	-	36

**Table 2A. Summary of new links made between BAC scaffolds on chromosome 3B.**

	<b>chr3B TGACv1</b>		<b>chr3DL BACs</b>	
	<b>Before</b>	<b>After</b>	<b>Before</b>	<b>After</b>
Bases (bp)	789,970,040	846,817,359	452,947,627	463,673,958
Assemblies	29,090	22,014	23,433	21,985
Num>n50	2,020	644	790	721
Min	500	500	501	501
n10	293,318	947,439	463,110	933,142
n50	116,546	398,569	154,985	286,993
Max	739,616	2,867,878	1,240,092	1,942,124

**Table 2B. Summary of changes in assemblies of chromosome 3B TGACv1 and chromosome 3DL BAC assemblies.**

## Additional File 1

### Genome assembly simulation

Before adopting Fosill jumping libraries for analysing and improving wheat genome assemblies, we simulated assembly processes using Fosill mate-pair reads on three of the largest scaffolds of the BAC-based assembly of chromosome 3B [1]. Simulations used Next-Generation illumina SIMulation PipeLinE (NGSimple, <https://github.com/lufuhao/NGSimple>) to assess various parameters, including library types, fragment sizes, read length, and sequencing depth. Simulated reads were generated by the Mason program [2] and then trimmed by Trimmomatic version 0.32 [3]. Quality control was applied to these datasets before and after, in order to confirm the removal of the low quality and low complexity bases in reads. Velvet v1.2.10 [4] was used to assemble these reads from different parameter settings in one run. And finally, MUMmer v3.23 [5] was used to map the assembled contigs back to its original scaffold to evaluate the quality of the faux assemblies. Results for a single representative scaffold are shown below.

#### Chromosome 3B scaffolds used for simulation

Scaffold ID	Length
v443_0936	4,169,843 bp
v443_0903	3,662,090 bp
v443_0899	3,466,457 bp

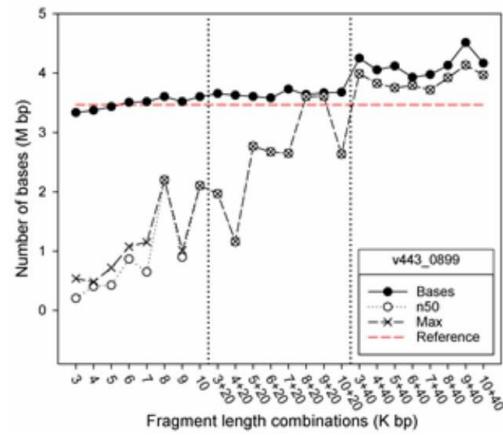
#### Simulation parameter settings

Parameters	Settings		
Library types	Paired end	Mate pair	Long mate pair
Fragment sizes (bp)	600	3 - 10 Kb	20 K, 40 Kb
Read lengths	100, 120, 150, 175, 200, 250 bp		
Sequencing depths	50X	2-10X	0.1-2X

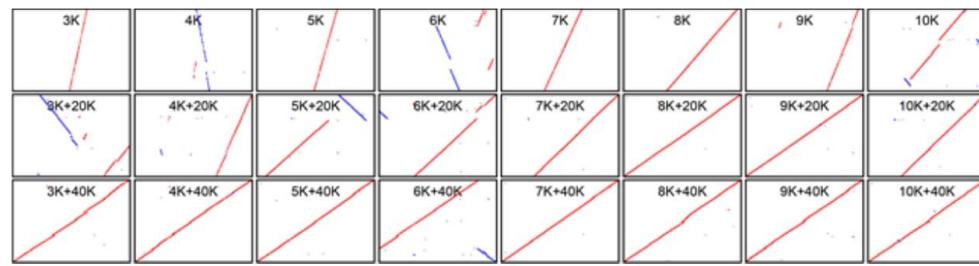
#### 1.1 Optimizing mate-pair fragment length

##### Simulation settings for mate-pair library fragment sizes

	Insert size	Coverage	Read length
Paired-end	600 bp	50x	250 bp
Mate pair	3,000-10,000 bp	10x	250 bp
Long mate pair	20 Kb, 40 Kb	2x	250 bp



**Fig.1** Summary of the *de novo* assemblies based on scaffold v443\_0899

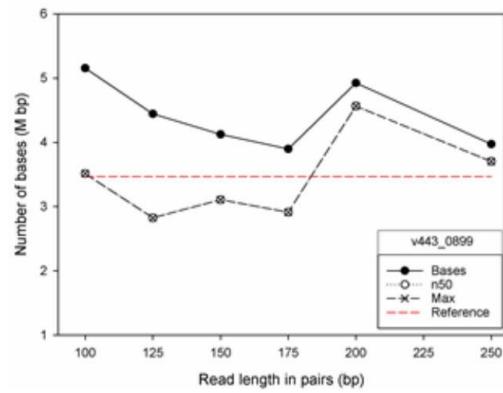


**Fig.2** Plot of longest contig against its reference v443\_0899

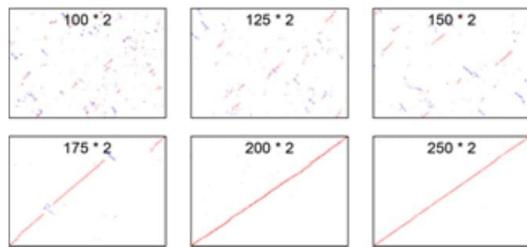
### 1.2 Optimizing mate-pair sequence read length

**Table 2.** Simulation settings for sequence read length

	Insert size bp	Coverage	Read length
Paired-end	600	30x	100-300 bp
Mate paired	7,000	10x	100-300 bp
Long mate paired	40,000	2x	100-300 bp



**Fig.3** Summary of the *de novo* assemblies based on v443\_0899

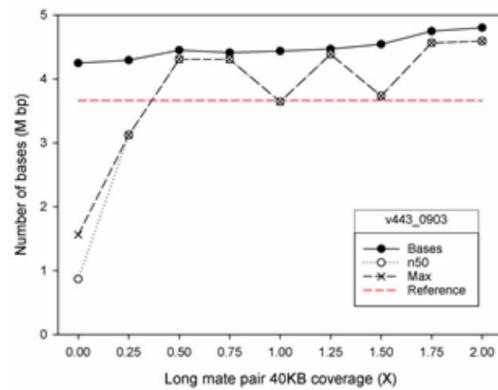


**Fig.4** Plot of longest contig against its reference v443\_0899

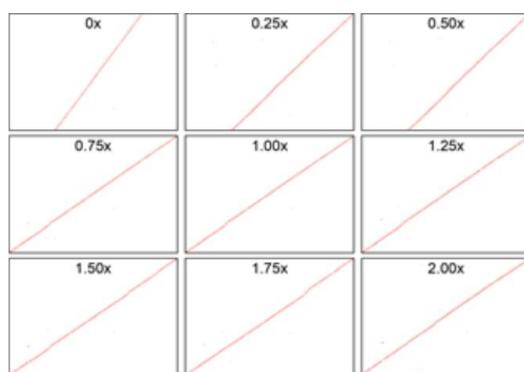
### 1.3 Optimizing long mate-pair coverage

**Table 3.** Simulation Setting for coverage by long mate-pair libraries

	Insert size bp	Coverage	Read length
Paired-end	600	50x	250 bp
Mate paired	7,000	10x	250 bp
Long mate paired	40,000	0-2x	250 bp



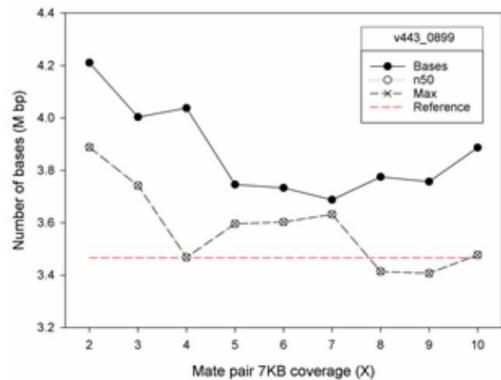
**Fig.5** Summary of the *de novo* assemblies based on v443\_0903



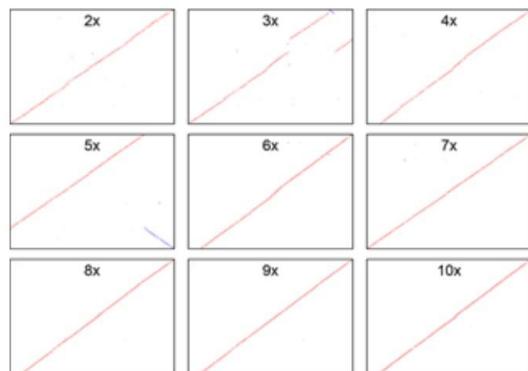
**Fig.6** Plot of longest contig against its reference v443\_0903

**Table 4.** Simulation Setting for coverage by mate-pair libraries

	Insert size bp	Coverage	Read length
Paired-end	600	50x	250 bp
Mate paired	7,000	2-10x	250 bp
Long mate paired	40,000	1x	250 bp



**Fig.7** Summary of the *de novo* assemblies based on v443\_0899



**Fig.8** Plot of longest contig against its reference v443\_0899

## References

1. Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, et al. Structural and functional partitioning of bread wheat chromosome 3B. *Science*. 2014;345:1249721.
2. Holtgrewe M. 2010. Mason - a read simulator for second generation sequencing data. Diploma Thesis. Repository: Freie Universität Berlin, Math Dept
3. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30:2114–20.
4. Zerbino DR, Birney E. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research*. 2008;18:821–9.
5. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. *Genome Biology*. 2004;5:R12.

## Additional File 2

### Fosill library production

The pFosill 4 cloning vector was used in this study was kindly provided by Louise Williams (Broad Institute, Cambridge, MA, USA). The construction, preparation and downstream steps for non-size-selected DNA fragments were carried out as described [1]. Only differences in the methods are described here.

A single-seed-descent line of *Triticum aestivum* Chinese Spring (CS42) was used for high molecular weight DNA extraction as described [2]. 30 µg of Genomic DNA was sheared to approximately 40 Kb average fragment size by HydroShear (Digilab, Marlborough, MA, USA) using the Large Shearing Assembly set at speed code 40 for 20 cycles. Sheared DNA was assessed by Pulse Field Gel Electrophoresis. 0.5-1 µg of sheared and non-sheared DNA was run on BioRad CHEF DRII. 120 degrees, 6V/cm, 1 to 10 second ramp switch, 17 hours at 14 °C. DNA was visualised by staining gel with Ethidium Bromide (Figure 1).

10 µg batches of sheared DNA was end-repaired in 175 µl reactions containing 1 X T4 ligase buffer, 0.25 mM dNTPs, 15 units T4 DNA polymerase, 50 units T4 polynucleotide Kinase, and 5 units Klenow fragment (all NEB) for 30 mins at 20 °C. TE was added to the DNA up to 400 µl, which was then cleaned and concentrated through Amicon Ultra 0.5 ml 100k concentrator (Millipore) at 2,000 g to approximately 30 µl. Recovered DNA was measured on the Qubit fluorometer (Invitrogen) using Quant-iT dsDNA BR kit.

t-b index linker A: GATCTCTACCAGG and t-b index linker B: CCTGGTAGAG were annealed and multiple ligations were set up using 500 ng end repaired genomic DNA and 200-fold molar excess of annealed linker. DNA was pooled and between 5 (500 µl) to 20 (2000 µl) ligations were cleaned and concentrated with one Amicon column.

Multiple 10 µl ligations were set up containing 250 ng linker DNA and 500ng cut and dephosphorylated pFosill 4 vector. 10 µl ligation was packaged with 2 successive 50 µl MaxPlax λ packaging extract (Epicentre) for 90 mins at 30 °C. 1,850 µl Phage dilution buffer and 140 µl DMSO were added (2,100 µl total volume) The libraries were titered and stored at -80 °C. λ-competent GC10 (Sigma) was used for processing packaged libraries into fosmid DNA. A proportion of 1.5 ml λ packaged sample to 40 ml of cells was found to give best transformation efficiency. Cultures were grown overnight at 30 °C in LB.

Fosmid DNA was isolated from LB culture using Qiagen's Plamid Maxi Purification Kit. 20 ml of Solutions P1, P2 and P3 were used and the supernatant was transferred to a new tube through a layer of Miracloth prior to addition to the Maxi column. Fosmid DNA was eluted with 500 µl TE and quantified using a Qubit fluorometer.

### **Conversion of Fosmids into Fossils.**

Pools of approximately 2.5 million independent Fosill clones were collected (see Table 1) and 10 µg of DNA from each pool was processed, with the following modifications. 900 ng was nicked with Nb.BbvCI for between 55-60 mins, before S1 nuclease treatment. 300 ng of DNA was re-circularised in 650 µl containing 1x T4 ligase buffer and 8,000 units of T4 ligase (NEB) at 16 °C for 16 hours. Products were purified using a Qiagen PCR cleanup kit. Columns were washed twice with 750 µl wash buffer and eluted with 55 µl TE.

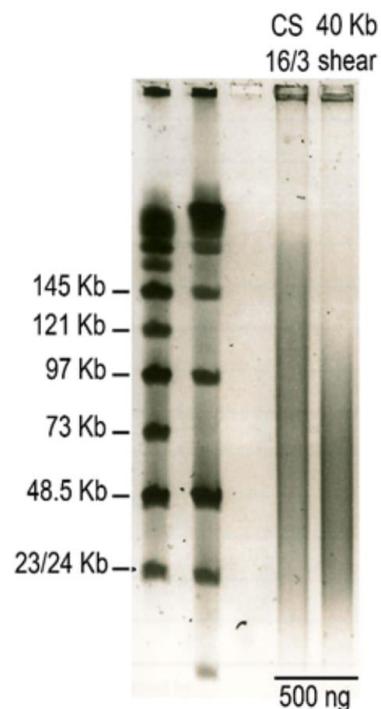
A trial PCR was used to determine minimal amplification required for Illumina template preparation. 2 µl of re-circularised DNA was amplified in 25 µl total volume of 1x Phusion HF master mix and 0.5 µM PCR primers:

SBS3: 5'AATGATAACGGCGACCACCGAGATCTACACTCTTCCCTACACGACGC 3'

SBS12:

5'AAGCAGAAGACGGCATACGAGATGATCGATCGTACTGGAGTTCAGACGTGTGC 3'

Cycling parameters were 98 °C for 3 mins, 16 and 18 cycles respectively of 98 °C for 15 secs, 65 °C for 30 secs, 72 °C for 30 secs and a final extension at 72 °C for 7 mins. PCR products were analysed with a MultiNA Bioanalyser (Shimadzu) using DNA 12000 reagent kit in on-chip mode. An intensity measurement of 5-8 MV, which equated to approximately 7.5 ng/µl to 12ng/µl for the 700 - 950 bp peak, was optimal. Following analysis of MultiNA data to determine minimal cycling conditions for each pool, Super-Pools of approximately 10 million independent Fosill clones were selected from the pools and minimal cycle number calculated for each pool to give sufficient material for sequencing. A total of 24 50 µl PCR reactions each containing 4 µl of Fosill DNA for each Super-Pool. Cycling parameters, primers and primer concentration were same as for trial PCR (except for varied cycle numbers). PCR products from Super-Pools were combined (1,200 µl) and purified with AMPure XP beads and eluted with 40 µl of TE. 4 µl of sample was used for MultiNA analysis to confirm size range and quantity. 30 µl of sample was size-selected on 1.5% agarose cassette with R2 marker using Sage Science BluePippin (Beverly, MA, USA) set to collect fragments between 650 – 1,000bp. Successful size selection was confirmed using TapeStation size measurement, and DNA was purified with AMPure XP beads and eluted in 25 µl TE. Sequencing was performed using 2 x 250 bp pair-end sequencing chemistry on an Illumina HiSeq 2500 sequencer.



**Figure 1. PFGE analysis of sheared DNA for Fosill vector cloning**

Library	Titre (M)	Platform	Num_Read	PCR redundancy
Lib17562	0.90	MiSeq	7,283,029	6.05
Lib18185	5.50	HiSeq	51,668,481	9.74
Lib18186	5.80	HiSeq	55,092,287	9.35
Lib19454	10.09	HiSeq	124,755,368	11.68
Lib19455	9.99	HiSeq	117,879,434	9.14
Lib19456	11.59	HiSeq	111,113,269	6.94
Lib19457	11.64	HiSeq	108,714,323	6.89
Total	55.51	-	576,506,191	8.51 average

**Table 1. Summary of Fosill libraries and paired-end reads generated**

## References

1. Williams LJS, Tabbaa DG, Li N, Berlin AM, Shea TP, MacCallum I, et al. Paired-end sequencing of Fosmid libraries by Illumina. *Genome Research*. 2012;22:2241–9.
2. Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, et al. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Research*. 2017;27:885–96.

### **Additional File 3.**

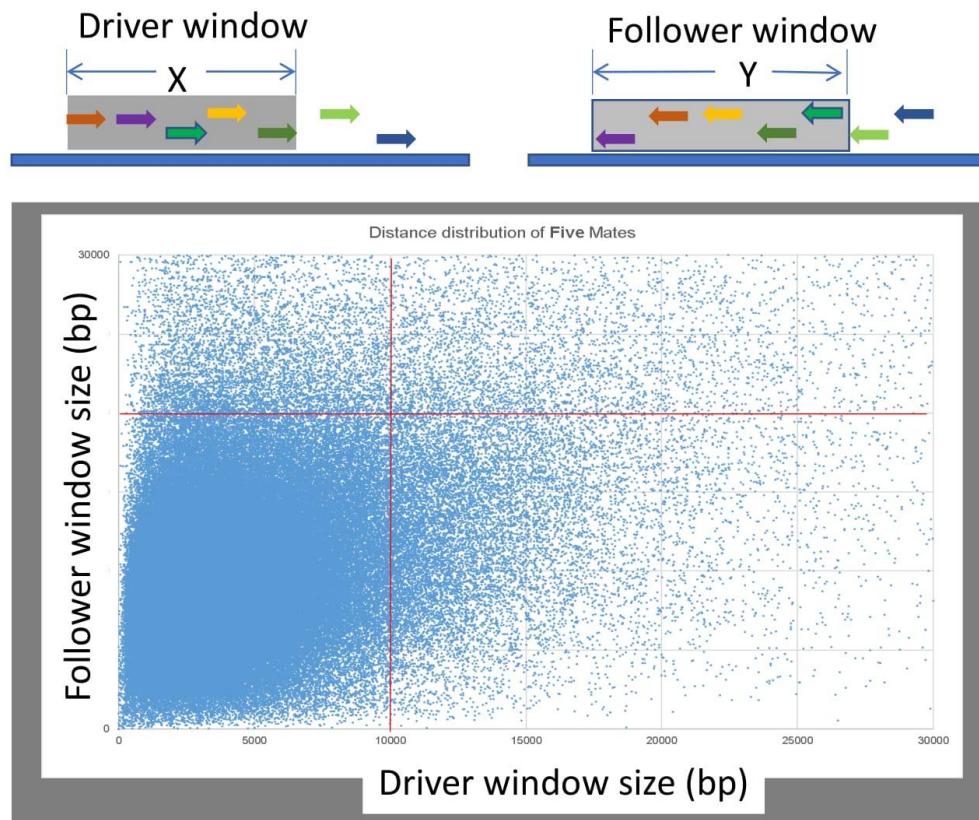
#### **Fosill mate-pair mapping**

Read pairs from Illumina Miseq or HiSeq sequencing that joined were removed using FLASH v1.2.11 [1]. Ligation adaptors and vector sequences in reads were trimmed off using CutAdapt v1.6 [2]. Sequencing primer sequences and low- quality sequences in reads were removed using Trimmomatic v0.32 (parameter) [3]. Resulting reads were then evaluated using FastQC v1.2.11 [4].

Trimmed Fosill mate-pair reads were filtered using the ReadCleaner4Scaffolding pipeline (<https://github.com/lufuhao/ReadCleaner4Scaffolding>). Both mates of each pair were mapped to chr3B BAC scaffolds using bowtie v1.0.1 [5]. Picard MarkDuplicates (v1.108, <http://broadinstitute.github.io/picard>) was then used to remove duplicates as single reads. A read depth threshold was determined to remove any highly repetitive reads by plotting the summary of output from samtools depth, and all the reads mapped to those regions with depth  $>5$  were not considered for scaffolding. Reads mapping to multiple positions, whose mates were not mapped, or were in the wrong orientation, were removed. A window sizing method was used to map mate-pairs to genomic regions. A group of  $\geq 5$  neighbouring mate reads within a “driver” window of less than 10 Kb were linked by the average 37.7 Kb insertion size  $\pm$  sd to a “follower” window of 20 Kb. Figure 1 shows that nearly all mate-pairs mapped to chromosome 3B using these criteria. Reads mapping within these window criteria were used to define regions of chromosomes that were consistent with the average insertion size, or had inconsistent matches.

To generate coordinates of each scaffold on the 3B pseudomolecule, BAC scaffolds were mapped to the pseudomolecule and plotted using SyntenyDraw (available on <https://github.com/lufuhao/SyntenyPlot>). These coordinates were compared with our evidence from ReadCleaner4Scaffolding pipeline. To validate mapping, mate-pairs were mapped TGAC v1 chr3B contigs.

Filtered Fosill mate-pair reads were mapped to the TGACv1 whole genome assembly of Chinese Spring 42 as described above, using SSPACE v3.0 to join scaffolds with five fossils links.



**Figure 1. Mapping Fosill mate pairs to genomic scaffolds**

The distribution of mate-pair links of five adjacent Fosill mate-pairs in different sized windows to their corresponding mate-pair read at the average insert size  $\pm$  sd was mapped on chromosome 3B BACs. The vast majority of mate-pairs in a 10 Kb window were found in 20 Kb windows at the correct distance. These window sizes were used to map Fosill mate-pairs to genomic scaffolds.

#### References

1. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 2017;17:10-12.
2. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*. 2011;27:2957–63.
3. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30:2114–20.
4. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. <http://wwwbioinformaticsbabraham.ac.uk/projects/fastqc/> doi:citeulike-article-id:11583827.
5. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Meth*. 2012;9:357–9.

## **Additional File 4.**

### **Sequencing Chromosome 3DL BAC minimal tiling path**

#### **BAC library preparation and sequencing**

The 3DL BAC library was prepared from flow sorted chromosomes [1] at The Institute of Experimental Botany, Olomouc, Czech Republic, and was fingerprinted at CNRGV (Toulouse, France) using SNaPshot-based high information content methods [2]. The raw fingerprint data was processed according to IWGSC guidelines. LTC [3] was used to build the physical map and generate a minimum tiling path (MTP). The final path consisted of 620 fingerprint contigs (FPCs) containing 5 or more BACs. 6,338 BACs were selected for sequencing (6,252 MTP clones plus 86 bridge clones).

Paired-end and long mate-pair (LMP) libraries were prepared and sequenced to generate PE reads for each BAC and a pool of LMP reads for each 384 well plate of BACs. LMP reads were processed as described in [4]. After standard QC, filtering and de-multiplexing, the reads were ready for assembly.

#### **BAC assembly and mate-pair preparation**

Reads were aligned to *E. coli* DH10B, wheat chloroplast and mitochondria sequences using Bowtie 2 [5]. Read pairs with one or more reads mapping with 95% identity or above were removed. Reads were also aligned to the pIndigoBAC-5 BAC vector sequence. Read pairs where one or more reads mapped to the middle of the vector sequence were removed while pairs where a read mapped to the end of a vector sequence were kept. This identified vector insert ligation sites. BACs were assembled individually using ABySS [6]. The BAC assemblies had an average insert size of 112,886 bp and an average N50 of 25,214 bp. In addition to the pooled LMPs, we used the 9 Kb and 12 Kb whole genome wheat LMPs from [4]. These were first filtered for non-3DL reads by alignment to the IWGSC CSS assembly where all 3DL contigs were replaced with our BAC assemblies. Reads were assigned to individual BACs as a side effect of this process.

Reads were then assigned from the pooled LMPs to each BAC. A Jellyfish [7] 31-mer hash table was generated from each assembly and these were combined to create a table of 31-mers found in the BACs on each plate. To identify LMP reads matching BACs, the “sect” function of the Kmer Analysis Toolkit (KAT) [8] v1.0.5 was used to generate a *k*-mer coverage profile of each LMP read in each pool using plate-specific PE *k*-mer hash tables. The plate-

specific LMP reads were then classified to individual BACs on that plate using  $k$ -mers from individual BAC assemblies.

### Chromosome arm assembly

Before any scaffolding, the BAC assemblies belonging to each FPC were then merged to remove redundancy. This was done first using CD-HIT [9] and then BLAST [10]. Any overlapping sequence at the end of two BAC contigs of at least 98 % identity and 1000 bp in length resulted in the two contigs being merged into one new contig. Following this procedure each FPC had an average size of 460 Kb and an N50 of 17 Kb.

The non-redundant FPCs were then scaffolded using Soapdenovo [11] with the assigned pooled and whole genome LMP reads. This resulted in an average FPC size of 782 Kb and an N50 of 180 Kb. Finally, the FPC sequences were combined and the merging process was run again. This resulted in a total size of 455 Mb for the whole chromosome arm and an N50 of 145 Kb.

### References

1. Safar J, Bartos J, Janda J, Bellec A, Kubaláková M, Valárik M, et al. Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. *Plant J.* 2004;39:960–8.
2. Luo M-C, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, et al. High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics.* 2003;82:378–89.
3. Frenkel Z, Paux E, Mester D, Feuillet C, Korol A. LTC: a novel algorithm to improve the efficiency of contig assembly for physical mapping incomplex genomes. *BMC Bioinformatics.* 2010;11:584.
4. Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, et al. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Research.* 2017;27:885–96.
5. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Meth.* 2012;9:357–9.
6. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABYSS: A parallel assembler for short read sequence data. *Genome Research.* 2009;19:1117–23.
7. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of  $k$ -mers. *Bioinformatics.* 2011;27:764–70.
8. Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. KAT: a  $K$ -mer

analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*. 2017;33:574–6.

9. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28:3150–2.
10. S. F. Altschul et al. “Basic local alignment search tool”. *Journal of Molecular Biology* 215.3 1990;215:403–410.
11. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*. 2012;1:18.