the plant journal

SEB Society for Experimental Biology

The Plant Journal (2014) **80,** 895–904 doi: 10.1111/tpj.12660

TECHNICAL ADVANCE

Using genic sequence capture in combination with a syntenic pseudo genome to map a deletion mutant in a wheat species

Laura-Jayne Gardiner[1], Piotr Gawroński[2], Lisa Olohan[1], Thorsten Schnurbusch[2], Neil Hall[1] and Anthony Hall[1,*]

[1]*Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool, UK, and*

[2]*Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstrasse 3, OT Gatersleben, D-06466 Stadt Seeland, Germany*

Received 2 May 2014; revised 29 August 2014; accepted 2 September 2014; published online 10 September 2014.
*For correspondence (e-mail Anthony.hall@liverpool.ac.uk).

SUMMARY

Mapping-by-sequencing analyses have largely required a complete reference sequence and employed whole genome re-sequencing. In species such as wheat, no finished genome reference sequence is available. Additionally, because of its large genome size (17 Gb), re-sequencing at sufficient depth of coverage is not practical. Here, we extend the utility of mapping by sequencing, developing a bespoke pipeline and algorithm to map an early-flowering locus in einkorn wheat (*Triticum monococcum* L.) that is closely related to the bread wheat genome A progenitor. We have developed a genomic enrichment approach using the gene-rich regions of hexaploid bread wheat to design a 110-Mbp NimbleGen SeqCap EZ in solution capture probe set, representing the majority of genes in wheat. Here, we use the capture probe set to enrich and sequence an F_2 mapping population of the mutant. The mutant locus was identified in *T. monococcum,* which lacks a complete genome reference sequence, by mapping the enriched data set onto pseudo-chromosomes derived from the capture probe target sequence, with a long-range order of genes based on synteny of wheat with *Brachypodium distachyon.* Using this approach we are able to map the region and identify a set of deleted genes within the interval.

Keywords: genomics, mapping by sequencing, next generation, target enrichment, wheat, technical advance.

INTRODUCTION

Genetic mapping of mutants is a lengthy process because of the time needed to perform crosses and generate marker data (Lukowitz *et al.*, 2000). A recent alternative to this is mapping by sequencing. Several pipelines have been introduced that streamline mapping-by-sequencing analyses in diploid species, including SHOREmap, MAQGene, MutMap, NGM-Next-generation EMS mutation mapping and CloudMap (Schneeberger *et al.*, 2009; Doitsidou *et al.*, 2010; Austin *et al.*, 2011; Abe *et al.*, 2012; Minevich *et al.*, 2012). SHOREmap is one of the most well-known pipelines and has formed the basis for many derivative mapping-by-sequencing methodologies, including MutMap and MAQGene. SHOREmap uses a combination of sequencing a phenotyped F_2 mapping population, a technique known as bulk segregation analysis (Michelmore *et al.*, 1991), and whole-genome re-sequencing to generate approximate mapping intervals that should contain the phenotype-inducing mutation. It was originally developed for mapping by sequencing *Arabidopsis thaliana* bulk segregant, mutant, mapping populations, and identified heterozygous single nucleotide polymorphism (SNP) markers to locate a region of marker scarcity (Schneeberger *et al.*, 2009). More recently SHOREmap is more comparable with CloudMap, and locates local differences in parental homozygous allele frequencies, across a reference sequence, with increased allele frequency of the mutant parent being introduced via mutant phenotypic selection in the mapping population (Galvao *et al.*, 2012; Hartwig *et al.*, 2012).

Mapping-by-sequencing analyses involve the generation of shotgun sequences of a phenotyped F_2 mapping population. In *Triticum monococcum* (wheat), because of its vast 17-Gbp genome and highly repetitive content of

approximately 80% (Choulet *et al.*, 2010), it is expensive to generate and challenging to analyze whole shotgun sequence data sets. To reduce this complexity, methods such as transcriptome sequencing (Trick *et al.*, 2012), restriction site-associated DNA sequencing (Baird *et al.*, 2008), or targeted enrichment sequencing (Winfield *et al.*, 2012) have been proposed to reduce the need for whole-genome re-sequencing. In this analysis, we demonstrate the utility of using the Nimblegen SeqCap EZ wheat exome capture probe set that was designed to target wheat genic regions prior to sequencing and to greatly reduce the cost associated with sequencing the wheat genome by eliminating much of the repetitive sequence from the analysis while still allowing mapping by sequencing.

Although mapping-by-sequencing analyses have been implemented without a reference sequence, it is acknowledged that for a streamlined and simplistic pipeline a reference sequence is beneficial (Galvao *et al.*, 2012; Nordstrom *et al.*, 2013). As such, many of the current methodologies rely on reference strain sequence or variant information. For wheat, like many crop species, no finished genome reference sequence is available. Here, a shotgun analysis of the wheat genome (Brenchley *et al.*, 2012) was used to develop the gene capture array. The long-range order of the short assemblies was approximated using synteny between wheat and *Brachypodium*, which diverged from wheat approximately 35–40 million years ago (Bossolini *et al.*, 2007). This allowed the construction of seven wheat pseudo-chromosome sequences that could be used as a reference genome for a sliding window mapping-by-sequencing analysis. Moreover, adding homeologous SNP information would allow the association of sequence data with each of the three wheat genomes in hexaploid wheat. This work extends a proof-of-principle approach where an Arabidopsis cDNA sequencing dataset was assembled into *Brassica rapa* based pseudo-chromosomes using synteny between the two species that diverged approximately 20 million years ago (Yang *et al.*, 1999). In a mapping-by-sequencing analysis two mutant intervals were identified as pseudo-chromosome positions in *B. rapa* using allele-frequency estimates at 4375 marker positions in an enriched subset of Arabidopsis. These *B. rapa* intervals were later translated back to a single Arabidopsis position (Galvao *et al.*, 2012).

To test the utility of combined genic enrichment and a sliding window mapping-by-synteny analysis, we mapped an early-flowering mutant earliness *per se* (*Eps-3A*^{m}) in the monocot species *T. monococcum* (Gawroński *et al.*, 2014). The *Eps-3A*^{m} locus comes from einkorn wheat line KT3-5 that is an X-ray mutant (Shindo and Sasakuma, 2001). *T. monococcum* is a close relative of *Triticum urartu,* the A-genome progenitor of modern hexaploid bread wheat (AA BB DD), with *T. monococcum* diverging from *T. urartu* between 0.5 and 1 million years ago (Huang *et al.*, 2002). As such, *T. monococcum* has been used successfully here as a model for the A genome progenitor of hexaploid wheat (Wicker *et al.*, 2003). *Eps-3A*^{m} is an early-flowering mutant with an altered circadian clock phenotype, with previous mapping suggesting that the phenotype results from the deletion of the circadian clock gene, an ortholog of the Arabidopsis circadian clock gene *LUX ARRHYTHMO/PHYTOCLOCK 1 (LUX)*, which is thought to play an important role in the evening complex within the circadian clock mechanism (Hazen *et al.*, *2005;* Gawroński *et al.*, 2014).

We demonstrate the feasibility of enrichment of a divergent species combined with a sliding window mapping-by-synteny approach, by first developing an enrichment approach to capture the genic portion of the wheat genome. Secondly, by generating wheat pseudo-chromosomes based on synteny between *Brachypodium* and wheat to form a 'reference genome' and thirdly, using a bespoke pipeline and algorithm to map the *Eps-3A*^{m} mutation to a small deletion on chromosome 3 in *T. monococcum*. We intend to ultimately apply this methodology to polyploid wheat. This intended application necessitates the development of a novel pipeline that prioritizes adaptability to polyploid plant species, as current tools are tailored to diploid species, typically detecting regions where allele frequency tends towards 1.

RESULTS

Enrichment of the genic portion of wheat

To reduce the size and complexity of the wheat genome we used the NimbleGen SeqCap EZ in solution custom capture probe method. In collaboration with NimbleGen we developed a custom probe set for the wheat genome (approximately 110 Mbp; see Figure S1). This was achieved, using the low copy number genome assembly (LCG) of the wheat cultivar Chinese Spring (http://mips.helmholtz-muenchen.de/plant/wheat/uk454survey/download/index.jsp). The LCG has chloroplast, mitochondria and transposon sequences removed, and also contains homoeologous copies of genes collapsed into a single contiguous sequence or contig (Brenchley *et al.*, 2012). The LCG is 3.8 Gb in size and, as such, is still too large for a custom capture target sequence. To reduce the size still further we used BLASTN (*e*-value $< 1e^{-10}$) to identify LCG contiguous sequences that were similar to *Brachypodium* gene sequences. We then used the same LCG sequence library (BLASTN *e*-value $< 1e^{-20}$) to identify LCG contiguous sequences that matched a set of non-redundant wheat cDNA and expressed sequence tag (EST) sequences (Galvao *et al.*, 2012), and transcriptome assemblies generated by 454 sequencing rounds of nine diverse wheat cultivars (Cavanagh *et al.*, 2013), not present in the *Brachypodium* gene set. Finally, to remove sequence duplications from the contiguous sequences set, we compared the set against

itself using BLASTN. Similar sequences were identified (95% identity over 100 bp), and the longest contiguous sequence of a matching pair was retained. This process resulted in a target sequence that was made up of 169 345 contigs that varied in size from 100 to 13 168 bp. The capture probes ranged in size from 50 to 100 bp, with an average length of 75 bp; they were generated with an average spacing of 49 bp across the target sequence (measured from the 5′ oligo starting position to the next 5′ oligo starting position).

The final design was non-redundant and covered the majority of the wheat genes, with each probe being potentially capable of enriching all three homoeologous gene copies in hexaploid wheat (Gu *et al.*, 2004). Although originally designed based on hexaploid wheat, the probe set is also effective, as demonstrated here, in enriching the genic portion of the diploid *T. monococcum* that is the domesticated relative of the A-genome progenitor *T. urartu*.

BLAST 2.2.17 alignments of the capture probe target sequences against existing data sets hit 94% of the genes in *Brachypodium distachyon* and 76% of the 97 481 full-length wheat cDNA contigs that were identified in the transcriptome assembly (Brenchley *et al.*, 2012; BLASTN *e*-value $< 1e^{-5}$).

Generating genome references for each parent

The first step in building reference genomes for two parental recombinant inbred lines (RILs) was to construct a set of wheat pseudo-chromosomes that were representative of the three genomes, using the contiguous sequences used to design the capture probe set. BLASTN 2.2.17 was used to place the contigs on the *Brachypodium* genome. For each contig the top BLAST gene hit was used and hits were prioritized as follows: highest BLAST score, longest length (minimum 100 bp), lowest *e*-value (cut-off $1e^{-3}$) and highest sequence identity (minimum 90%). This allowed us to place 115 250 out of the original 169 345 contigs on the *Brachypodium* genome. A set of 807 wheat markers was used to associate parts of the *Brachypodium* genome with the corresponding wheat genome (Wilkinson *et al.*, 2012). The midpoint of each contiguous sequence was taken as the 'probe position' to enable calculation of the nearest wheat marker as a measure of distance along the relevant *Brachypodium* chromosome. The probes could then be ordered into Chinese Spring pseudo wheat chromosomes based on association with wheat chromosome marker positions.

The pipeline to generate 'reference genomes' based on the two parents RIL71 and RIL25 is shown in Figure 1(a). Genomic DNA was isolated from each of the parents and separately enriched using our Nimblegen SeqCap EZ wheat exome capture probe set before sequencing using an Illumina HiSeq 2000 (http://res.illumina.com/documents/products/datasheets/datasheet_hiseq2000.pdf). The short read sequences were mapped to the Chinese Spring pseudo-wheat chromosomes using short-read mapping software BWA (Li and Durbin, 2009). Across the two sequencing data sets, coverage was highly conserved with on average 70% of the reference base space being mapped to uniquely after removal of duplicate reads, at an average depth of approximately 70X coverage (see Table S1). Coverage of the mapping reference was acceptable considering that we were mapping *T. monococcum* reads to a reference sequence that was designed based on the hexaploid wheat Chinese Spring. In RIL25 and RIL71 this mapping extended across 108 218 and 108 263 of the capture probe target sequence contigs that made up the pseudo-chromosomes, respectively. Of the 7031 and 6986 target sequence contigs that were unmapped by the RIL25 and RIL71 sequencing data sets, the majority, 6072, were the same contigs. To enable comparison sequencing data for the hexaploid wheat, Chinese Spring was mapped to the pseudo-chromosome sequences gaining approximately 98% coverage of the reference sequence. This encompassed 114 981 capture probe target sequence contigs, with only 268 unmapped; 186 of these were also unmapped in the two RIL *T. monococcum* lines.

The SNPs were scored between the Chinese Spring reference and the two RILs using Samtools MPILEUP 0.1.18 (Li *et al.*, 2009) and VARSCAN 2.2.11 (Koboldt *et al.*, 2012). SNPs with a mapping coverage of <10 and sequencing reads with mapping quality scores of less than 15 were filtered out (see Table S1). We identified 881 860 homozygous SNPs in relation to the Chinese Spring reference that were shared between the two RIL lines. These were removed, leaving 96 651 RIL25- and 131 409 RIL71-specific homozygous SNPs. The RIL-specific SNPs could then be used to generate RIL25 and RIL71 'reference genomes'.

Generation of a bulk segregant mapping population for mapping-by-sequencing

Mapping-by-sequencing relies on a local skewing of allelic frequency close to the locus responsible for the mutant phenotype. The analysis benefits from the generation of two contrasting bulks based on phenotypic scores in the F_2 mapping population and subsequent genotyping by sequencing. The breeding and phenotyping of the lines that were analyzed here was performed as previously reported (Gawroński *et al.*, 2014). Two parental RIL lines, RIL25 (early flowering) and RIL71 (wild type), were formed from a series of crosses between the early flowering *T. monococcum* mutant KT3-5 with a wild accession KT1-1 of *Triticum boeoticum*. To generate an F_2 mapping population, RIL25 and RIL71 were then crossed and the resulting F_1 plants were allowed to self-pollinate. This mapping population was phenotypically classified into two groups: wild type (parent RIL71) phenotype, named Bulk A, and early-flowering (parent RIL25) phenotype, named Bulk B. Each

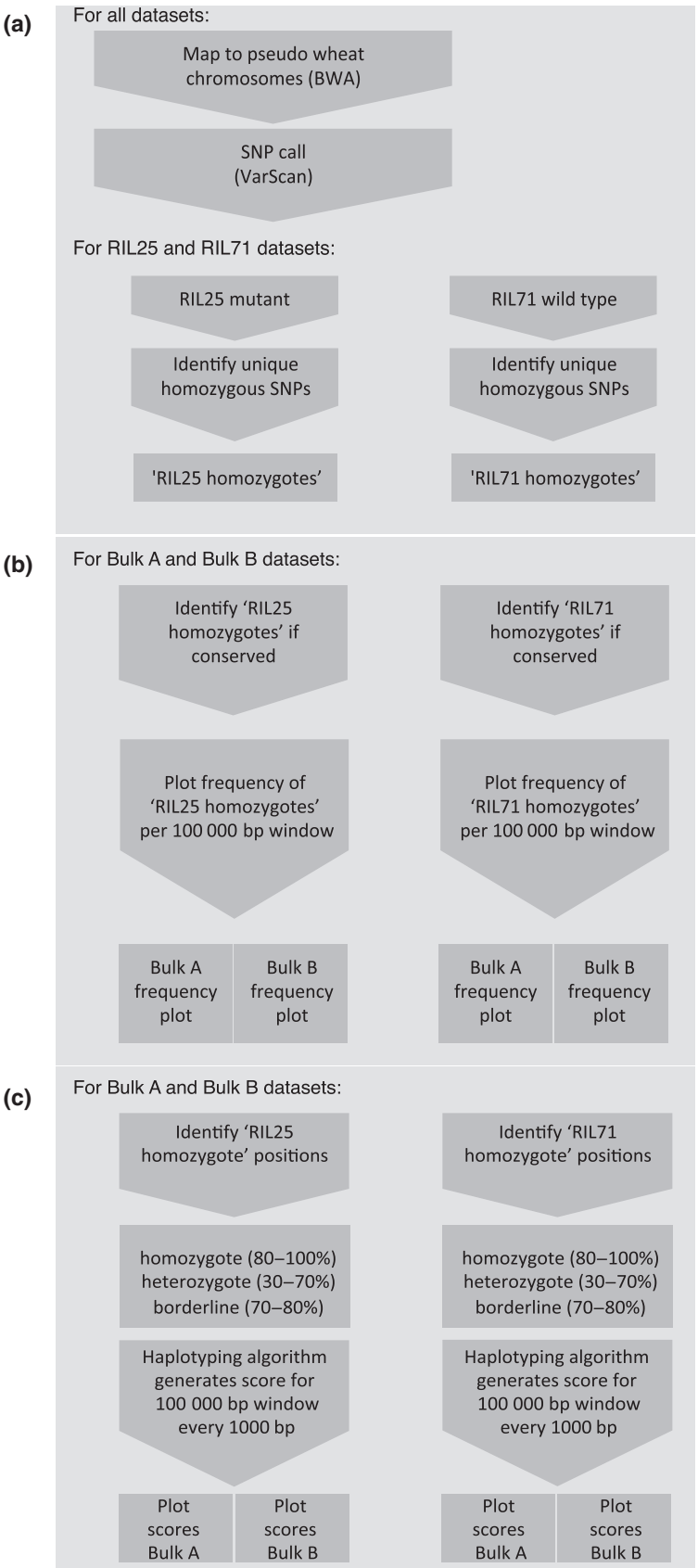

Figure 1. Processing four sets of enriched sequencing data to identify a mapping interval containing the deletion that is inducing the phenotype of interest.
(a) Standard mapping and single nucleotide polymorphism (SNP) calling pipeline to construct 'reference genomes'.
(b) Initial homozygote allele frequency determination method for Bulk-A and Bulk-B samples.
(c) Final allele frequency algorithm or Bulk-A and Bulk-B samples to identify the interval of interest.

bulk contained approximately 250 individual plants and Bulk B contained DNA from F_2 plants that were homozygous mutant-like at the locus, whereas Bulk A contained heterozygous and wild-type plants (Gawroński *et al.*, 2014). The mutation present in Bulk B is expected to be fixed and homozygous because it showed a recessive segregation pattern in an earlier study (Shindo and Sasakuma, 2001). The two bulk segregated populations were enriched using the Nimblegen SeqCap EZ wheat exome capture array in solution, and then sequenced. The sequence data was mapped using BWA, and SNPs scored and filtered as described for the parents, using SAMtools MPILEUP and VARSCAN (see Table S1).

Using homozygote allele frequency determination to map by sequence the Eps-3A^m mutation

Two mapping-by-sequencing mutant identification pipelines were developed and successfully used for this analysis (Figure 1b,c). The first pipeline was developed as a starting point, based on similar techniques to those demonstrated by SHOREmap, with the intention of future easy adaptation to a polyploid species. This pipeline was used to identify regions with increased homozygous frequency compared with the parent genome. Of the homozygote SNPs that were specific to the RIL25 parent, 34 125 SNPs could be found as conserved homozygous alleles in the Bulk A data, and 46 154 SNPs could be found in the Bulk B data. Of the homozygote SNPs that were specific to the RIL71 parent, 54 376 were conserved as homozygotes in the Bulk A data and 64 700 were conserved as homozygotes in the Bulk B data. Frequencies of these RIL71 and RIL25 SNPs were calculated per 100 000-bp window along each chromosome for Bulk-A and Bulk-B data sets, and displayed graphically. A clear peak of conserved RIL25 homozygous SNP frequency was observed at the end of chromosome 3 in Bulk B (Figure 2a). The same peak was seen for conserved RIL71 homozygous SNPs at the end of chromosome 3 in Bulk A (Figure 2b).

The interval that the peak occurred within translated to the window 10 000 000–10 600 000 bp of the pseudo-wheat chromosome 3. There were 748 probes concatenated to form this region, and these probes aligned

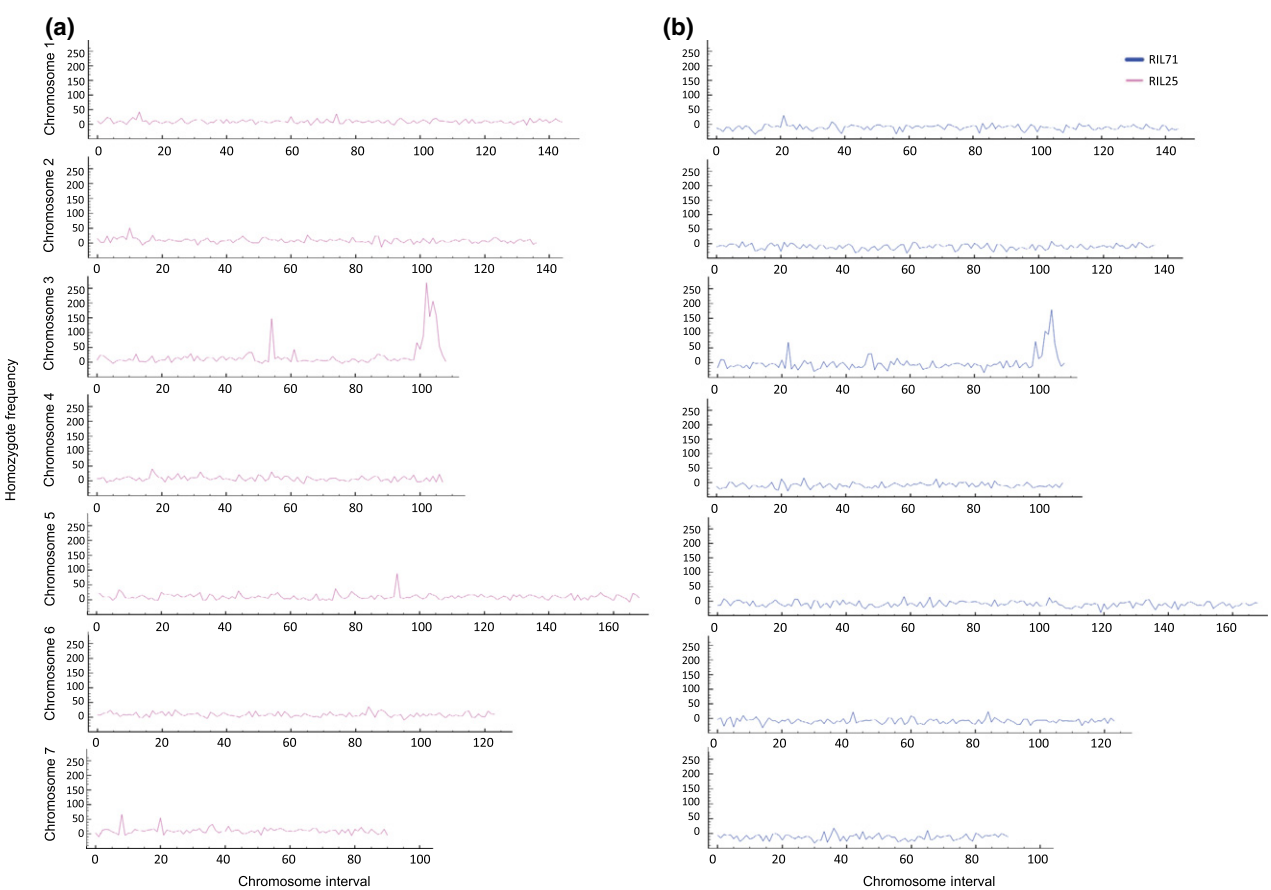

Figure 2. Frequencies of Bulk-A and Bulk-B homozygotes calculated along each pseudo-chromosome.
(a) Frequency of 'RIL25 homozygous' single nucleotide polymorphisms (SNPs) per window: Bulk-B frequency minus Bulk-A frequency per 100 000-bp window.
(b) Frequency of 'RIL71 homozygous' SNPs: Bulk-A frequency minus Bulk-B frequency per 100 000-bp window.

with the region 58 063 918–59 004 348 in *Brachypodium* chromosome 2, including the genes Bradi2 g60780–62310 (approximately 160 genes).

Homozygote allele frequency determination algorithm for Bulk-A and Bulk-B samples

The second method that we developed uses an algorithm to score regions of interest by prioritizing long homozygous parental haplotypes: the longer the length and the more homozygous the region, the higher the score generated. Of the homozygote SNPs that were specific to the RIL25 parent, 49 126 of the SNP alleles were conserved in the Bulk-A data, regardless of homozygous or heterozygous status, and 62 388 of the SNP alleles were conserved in the Bulk B data. Similarly, of the homozygote SNPs that were specific to the RIL71 parent, 54 377 of these were found in the Bulk-A data and 83 401 were found in the Bulk-B data. These SNPs were categorized into homozygous, heterozygous or borderline, and a scoring system was developed to calculate a homozygote score per 100 000-bp window along the pseudo chromosomes at 1000-bp intervals (see Figure S2). These frequencies were plotted in Figure 3 to clearly identify a single peak at the end of chromosome 3. Magnification of the peak region confirms that the exact same peak has been defined that was seen in the previous analysis: i.e. the 10 000 000- to 10 600 000-bp window of the pseudo wheat chromosome 3 that aligned to the region 58 063 918–59 004 348 in *Brachypodium* chromosome 2, including the genes Bradi2 g60780–62310 (approximately 160 genes). This method allows increased definition of the interval of interest at the end of chromosome 3 in comparison with our initial method, whilst reducing background noise.

Investigation of the defined region on chromosome 3

The enrichment approach not only allows the identification of SNPs, it also allows the identification of copy-number variation and deletions. Therefore, we scanned the 'reference genome' for deletions. To define a deleted region that potentially induces the mutant phenotype required mapping coverage in the wild-type RIL71 and/or the Bulk-A data sets, but no mapping in the mutant RIL25 and the Bulk-B data sets. Deleted regions were defined that were longer than 100 bp, and that fitted this mapping expectation. Using this approach we identified 163 deleted regions: 11 were within the interval that was identified at the end of chromosome 3.

Although the enrichment array is predicted to contain the majority of the genic sequence of wheat, because only approximately 70% of the reference sequence has been mapped to it, it is possible that the causal deletion could be partially excluded from the analysis. As such, we may

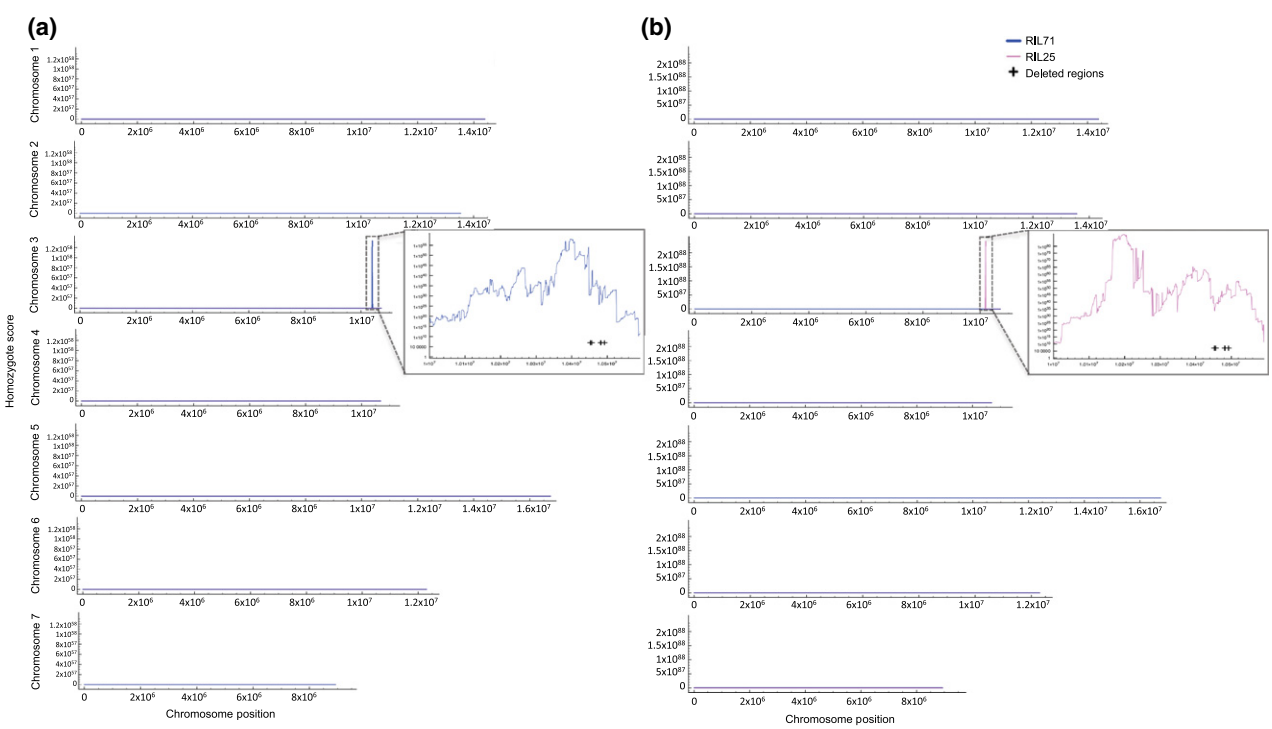

Figure 3. Homozygosity scores calculated for Bulk-A and Bulk-B data sets. Homozygote scores plotted along each pseudo-chromosome. Haplotypes conserved with the RIL25 (magenta line) and RIL71 (blue line) parental lines. Scores calculated per 100 000-bp window and calculated at 1000-bp intervals:
(a) scores for Bulk-A data set;
(b) scores for Bulk-B data set. Peak interval magnified and regions harboring a deletion >100 bp are highlighted.

not see a full segment deletion but a concentrated region of smaller deleted regions using this approach. The 163 deletions were further filtered for regions where two or more deletions could be seen within a 1000-bp window, resulting in 18 deleted regions that are detailed in Table 1, the majority of which lie underneath the defined interval of interest on pseudo-chromosome 3.

Deletions within our peak interval of interest are set in bold type in Table 1, and are also plotted in Figure 3. This highlights sequences covering approximately 40 Kbp, and particularly a wheat gene region (10 481 196–10 482 558 bp) that previously was found to align to the *Brachypodium* gene *Bradi2 g62067*, with similarity to the Arabidopsis *LUX* gene. In a BLASTN alignment of this wheat gene region to the BLAST NR nucleotide database (Altschul *et al.*, 1990), strong similarity was seen with the *Triticum aestivum* cultivar Chinese Spring *LUX* gene (*e*-value $< 1e^{-5}$, length approximately 859 bp and sequence identity 99%). The *LUX* gene is known to affect both the circadian clock and flowering time in Arabidopsis (Hazen *et al.*, 2005), and therefore is a strong candidate for the mutation responsible for the *Eps-3A*^{m} mutation in *T. monococcum* (Gawroński *et al.*, 2014).

DISCUSSION

Here, we have taken a mutant bulk segregant F_2 population that was developed from parental RILs, performed target enrichment for genic regions in the non-model grass *T. monococcum*, with relatively poor genome resources, and identified a region on chromosome 3 that is likely to contain the *Eps-3A*^{m} mutation.

We have been able to identify a region of approximately 600 Kbp within our pseudo chromosomes, and within this region pinpoint a approximately 40-Kbp region based on the identification of deletion hot spots. Finally, by assessing gene annotation, our candidate gene for the phenotype itself could be narrowed to a single capture probe target sequence contig of 3693 bp that had a high deletion frequency and showed a high degree of similarity to the *T. aestivum* cultivar Chinese Spring *LUX* gene. The *LUX* gene is known to affect both the circadian clock and flowering time in Arabidopsis (Hazen *et al.*, 2005).

We have demonstrated the use of a target enrichment strategy using capture probes that have been designed against the hexaploid wheat Chinese Spring to enrich the genic portion of a closely related plant, gaining on average 70X mapping coverage across 70% of the pseudo-chromosome reference sequence. This highlights the possibility for other capture probe sets to be used for close relatives with little or no resources available, for example a similar *Glycine max* (soybean) NimbleGen SeqCap EZ capture probe set could be used to map by sequence other related beans, such as *Cicer arietinum* (chickpea; Freeberg *et al.*, 2012).

This study extends a proof of concept approach where enrichment of a subset of a phenotyped Arabidopsis F_2 mapping population was performed in combination with a mapping-by-synteny approach to order Arabidopsis cDNA into *B. rapa* pseudo-chromosomes, based on synteny. Two mutant intervals were defined in *B. rapa* using allele-frequency analysis at marker positions. This translated to one position in Arabidopsis (Galvao *et al.*, 2012). Here, the full

Table 1 Detailing the pseudo-chromosome regions that lie underneath the peak of interest and harbor potential deletions

Chromosome	Position	Length of hit	Associated gene	Function
3	2 205 655	121	*Bradi2 g50140*	1,3-β-D-glucan synthase activity
3	2 205 890	441	*Bradi2 g50140*	1,3-β-D-glucan synthase activity
3	2 206 382	163	*Bradi2 g50140*	1,3-β-D-glucan synthase activity
3	**10 452 241**	**129**	***Bradi2 g61960***	**DEAD box ATP dependent RNA helicase activity**
3	**10 453 160**	**105**	***Bradi2 g61960***	**DEAD box ATP dependent RNA helicase activity**
3	**10 456 240**	**313**	***Bradi2 g61960***	**DEAD box ATP dependent RNA helicase activity**
3	**10 456 774**	**103**	***Bradi2 g61960***	**DEAD box ATP dependent RNA helicase activity**
3	**10 481 196**	**131**	***Bradi2 g62067***	**Similar to LUX gene G2-like (Myb-like domain)**
3	**10 481 946**	**271**	***Bradi2 g62067***	**Similar to LUX gene G2-like (Myb-like domain)**
3	**10 482 446**	**112**	***Bradi2 g62067***	**Similar to LUX gene G2-like (Myb-like domain)**
3	**10 491 979**	**221**	***Bradi2 g62093***	**(Upstream of) gene contains F-box domain**
3	**10 492 675**	**179**	***Bradi2 g62093***	**(Upstream of) gene contains F-box domain**
4	2 761 361	195	*Bradi1 g69930*	Putative digalactosyldiacylglycerol synthase
4	2 761 813	182	*Bradi1 g69930*	Putative digalactosyldiacylglycerol synthase
5	577 462	132	*Bradi2 g39240*	RNA binding
5	577 604	239	*Bradi2 g39240*	RNA binding
5	4 042 813	110	*Bradi4 g04880*	Protein binding
5	4 042 987	102	*Bradi4 g04880*	Protein binding

Deletions are defined as regions longer than 100 bp that are mapped by the RIL71 data and Bulk-A data, but that are unmapped by the RIL25 data and the Bulk-B data when two or more deleted segments are found within a 1000-bp window. Regions associated with the candidate gene are underlined and regions within our peak interval of interest are set in bold type.

genic sequence of wheat was enriched and assembled into wheat pseudo-chromosomes based on synteny with the closely related *Brachypodium*, to allow sliding window mapping-by-sequencing analyses. The mutant deletion could be identified directly as a position in wheat. To our knowledge, sliding-window analyses have not yet been combined with mapping by synteny when implementing a pseudo genome. By targeting the majority of the genic sequence of wheat the concerns expressed by Galvão *et al.*, that the causal mutation would be unlikely to be targeted with enrichment, are addressed. Here, the likelihood of enrichment of the region of interest is increased. We have not only used a more divergent species to order our fragmented mapping reference, additionally our mapping reference itself and enrichment capture probe set are both divergent from the species under analysis.

This analysis has taken the principles demonstrated by SHOREmap (Schneeberger *et al.*, 2009) to allow the development of a unique homozygote-scoring algorithm to highlight longer homozygous haplotypes shared between the mutant parental line and the bulk segregant mutant F_2 data set. This algorithm was implemented to identify the gene that is likely to contain the phenotype-inducing deletion in the early-flowering diploid einkorn wheat mutant, and the current pipeline is available as a workflow for public use within the iPlant collaborative web portal (see Experimental procedures; The iPlant Collaborative, 2011).

Importantly, it is anticipated that with a simple adjustment of the definitions of a homozygous, heterozygous and borderline SNP within the analysis, this method will be easily adaptable to tetraploid or hexaploid plants using the same pseudo-genome reference. Hexaploid wheat can be treated as three separate diploid genomes (A, B and D) to enable the use of current diploid pipelines; this necessitates mapping sequence data to reference sequences that confidently represent each of the three genomes. It is estimated that homeologous gene copies have high similarity and differ by only 1 in approximately 100 bp (Barker and Edwards, 2009); this increases the difficulty to map sequencing reads to a single genome and ultimately decreases the number of uniquely mapped sequencing reads. We propose that one wheat reference sequence, which is implemented here, representing all three diploid genomes of wheat, is more desirable for mapping analyses in hexaploid wheat and is still applicable to diploid species. As such, it is paramount to confidently detect and prioritize regions of homozygosity in a polyploid data set. The pipeline developed here allows the user to define upper and lower limits for both heterozygote and homozygote SNP calls to identify confident markers for interval detection that can be tailored to polyploid species: i.e. for a hexaploid wheat species we may anticipate a homeologous homozygote SNP that is likely to induce a phenotype to be present in roughly one-third of the sequencing reads, rather than the approximately 100% seen in a diploid organism. Furthermore, the use of a unique scoring algorithm to prioritize blocks of homozygosity will assist in the visualization of a peak interval in a polyploid data set where, because of the presence of SNP alleles at lower frequencies, noise levels from false SNP calls will be significantly higher. Application of this methodology to bread wheat will allow a rapid method to identify markers and potentially genes underlying key phenotypic traits.

EXPERIMENTAL PROCEDURES

Sequence capture and sequencing protocol for the RIL71, RIL25, Bulk-A and Bulk-B data sets

Initially, genomic DNA was purified using Agencourt AMPure XP beads (Beckman Coulter, http://www.beckmancoulter.com). Samples were quantified using a Qubit double-stranded DNA Broad Range Assay Kit and Qubit fluorometer (Life Technologies, http://www.lifetechnologies.com). A 2.6-μg portion of genomic DNA, in a total volume of 130 μl, was sheared for 3 × 60 sec using a Covaris S2 focused ultrasonicator (duty cycle 10%, intensity 5, 200 cycles per burst using frequency sweeping). The size distribution of the fragmented DNA was assessed on a Bioanalyzer High Sensitivity DNA chip (Agilent, http://www.genomics.agilent.com). In total, 50 μl (approximately 1 μg) of sheared DNA was used as input for library preparation. End repair, 3′-adenylation, and adapter ligation were performed according to the Illumina TruSeq DNA Sample Preparation Guide (Revision B, April 2012), without in-line control DNA and without size selection. Amplification of adapter-ligated DNA (to generate pre-capture libraries), hybridization to custom wheat NimbleGen sequence capture probes, and washing, recovery and amplification of captured DNA were all carried out according to the NimbleGen Illumina Optimised Plant Sequence Capture User's Guide (version 2, March 2012), with the exception that purification steps were carried out using Agencourt AMPure XP beads instead of spin columns. Final libraries were quantified by Qubit double-stranded DNA High Sensitivity Assay, and the size distribution ascertained on a Bioanalyser High Sensitivity DNA chip. The four libraries were then pooled in equimolar quantities based on the aforementioned Qubit and Bioanalyser data. Sequencing was carried out on two lanes of an Illumina HiSeq 2000, using version 3 chemistry, generating 100-bp paired-end reads.

Mapping and SNP identification in the RIL71, RIL25, Bulk A and Bulk B data sets

The resulting sequence data was mapped to the pseudo-chromosome sequences using the bwa-short algorithm in the short-read mapping software BWA 0.6.2 (Li and Durbin, 2009) (Figure 1a). Indexing of the reference sequence involved use of the 'IS' algorithm, and during alignment of reads to the reference using bwa aln, four mismatches were allowed per sequencing read. The mapping seed by default was allowed to have three mismatches within it, the used seed length was 30 and the quality threshold for trimming reads was implemented and set at 20. In anticipation of local re-arrangements in sequence between the diverged species, the 100-bp raw sequencing reads were split into two 50-bp reads and mapped separately using these parameters. All unmapped, non-uniquely mapped and duplicate reads were removed from the analysis. Samtools MPILEUP 0.1.18 (Li *et al.*, 2009)

was implemented on the four data sets and SNP calls were filtered out using VARSCAN 2.2.11 (Koboldt *et al.*, 2012), with the following parameters: discard SNPs covered by 10 or fewer reads; discard sequencing reads with a quality of less than 15; and if the alternate allele has less than two supporting reads passing the quality filter, discard it. For this SNP analysis indels were removed from the VARSCAN output. SNPs in 80% or more of the sequencing reads were identified as homozygous in each of the parental data sets, and homozygous SNPs that were unique to the mutant RIL25 data set and the wild-type RIL71 data set were identified: i.e. RIL25-specific homozygotes and RIL71-specific homozygotes.

Initial homozygote allele frequency determination method for Bulk-A and Bulk-B samples

Homozygous SNPs between the reference and the two bulk segregant data sets individually could be identified from the VARSCAN output using the same 80% cut-off as was used for the parental lines (Figure 1b). If the RIL25-specific homozygotes could be found in the Bulk-A/Bulk-B data as homozygotes, then they were added to produce a Bulk-A and Bulk-B 'RIL25 homozygote' list, respectively. Frequencies of the Bulk-A and Bulk-B 'RIL25 homozygotes' were calculated per 100 000-bp window, along each pseudo-chromosome. The same was done for the homozygous RIL71-specific SNPs to produce a Bulk-A and Bulk-B 'RIL71 homozygote' list, from which frequencies per 100 000-bp window could be calculated.

Final homozygote allele frequency determination algorithm for Bulk-A and Bulk-B samples

SNP alleles from the RIL25-specific homozygote list that could be found, regardless of homozygous or heterozygous status, in Bulk-A and/or Bulk-B were added to produce a Bulk-A and Bulk-B 'RIL25 homozygote' list respectively (Figure 1c). The same was done for the RIL71-specific homozygotes to produce a Bulk-A and Bulk-B 'RIL71 homozygote' list, respectively. These SNPs were then categorized into homozygous (alternate allele in $\geq$80% sequencing reads), heterozygous (alternate allele in $\geq$30% and $\leq$70% sequencing reads) or borderline (alternate allele in >70 and <80% sequencing reads), and a unique scoring system was used to calculate a homozygote score per 100 000-bp window along the pseudo chromosomes at 1000-bp intervals (see Figure S2).

Implementing the mapping, SNP calling and homozygote haplotyping algorithm in iPlant

Our mapping and SNP calling pipelines are available on iPlant as two workflows within the Discovery Environment: 'Mapping illumina seq data Part 1' and 'SNP calling illumina seq data Part 2'. These workflows will map and SNP call in illumina sequencing data sets, ideally requiring a mutant parental line, a wild-type parental line and a bulk segregant mutant F_2 pool as input, as used within this study. The workflows allow the user to define parameters, but the parameters that were used for this study are implemented by default and the two parental SNP lists generated as output are used as input for the workflow 'Identification of unique homozygous SNPs in mutant' to identify mutant parental-specific SNPs. Finally, the workflow 'Mutant Identification 1' takes this mutant parent-specific SNP list and the bulk segregant mutant F_2 population SNP list as input, finds conserved SNP alleles between the two and implements our homozygote haplotyping algorithm to output a pdf file R plot identifying the mutant interval of interest. The text file of homozygote scores used to plot this is also generated as output. An example of the pdf output generated by iPlant for the Bulk-B mutant pool using the RIL25-specific homozygote list is outlined in Figure S3.

ACKNOWLEDGEMENTS

DNA sequence was generated by The University of Liverpool Centre for Genomic Research (United Kingdom). This project was supported by the BBSRC via a Doctoral Training Grant (L.G.), a BBSRC Career Development Fellowship BB/H022333/1 (A.H.) and a Natural Environment Research Council grant (L.O.). The phenotyping and generation of plant material was carried out by P.G. and T.S. The enrichment and Illumina sequencing was performed by L.O. The project was designed by A.H. with assistance from N.H. The projected was planned and conducted by L.G. and A.H. The paper was written by L.G. with assistance from A.H. All authors read and approved the final manuscript. We are grateful to Alistair Darby for his scientific contribution while car sharing.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Schematic to outline the design of a wheat gene capture array.

Figure S2. Outline of a unique scoring system used to calculate a homozygote score per 100 000-bp window along the pseudo chromosomes at 1000-bp intervals.

Figure S3. Homozygosity scores calculated and plotted using workflows implemented through iPlant for haplotypes conserved between the Bulk-B data set in relation to the RIL25 parental line.

Table S1. Mapping statistics for the four enriched wheat DNA samples in relation to the pseudo-chromosome reference sequence.

REFERENCES

Abe, A., Kosugi, S., Yoshida, K. *et al.* (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat. Biotechnol.* **30**, 174–178.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Austin, R., Vidaurre, D., Stamatiou, G. *et al.* (2011) Mext-generation mapping of Arabidopsis genes. *Plant J.* **67**, 715–725.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A. and Johnson, E.A. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, **3**, e3376.

Barker, G. and Edwards, K. (2009) A genome-wide analysis of single nucleotide polymorphism diversity in the world's major cereal crops. *Plant Biotechnol.* **7**(4), 318–325.

Bossolini, E., Wicker, T., Knobel, P.A. and Keller, B. (2007) Comparison of orthologous loci from small grass genomes Brachypodium and rice: implications for wheat genomics and grass genome annotation. *Plant J.* **49**, 704–717.

Brenchley, R., Spannagl, M., Pfeifer, M. *et al.* (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, **491**, 705–710.

Cavanagh, C., Chao, S., Wang, S. *et al.* (2013) Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc. Natl Acad. Sci. USA*, **110**, 8057–8062.

Choulet, F., Wicker, T., Rustenholz, C. *et al.* (2010) Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell*, **22**, 1686–1701.

Doitsidou, M., Poole, R.J., Sarin, S., Bigelow, H. and Hobert, O. (2010) *C. elegans* mutant identification with a one-step whole-genome-sequencing and SNP mapping strategy. *PLoS One*, **5**, e15435.

Freeberg, L., Gerhardt, D., Burgess, D., Millard, T., Green, D. and Jeddeloh, J. (2012) Soybean Exome Capture: Roche NimbleGen Inc. http://www.nimblegen.com/products/lit/06703526001_SeqCapEZ_SoybeanTechNote_0512.pdf.

Galvao, V.C., Nordstrom, K., Lanz, C., Sulz, P., Mathieu, J., Pose, D., Schmid, M., Weigel, D. and Schneeberger, K. (2012) Synteny-based mapping-by-sequencing enabled by targeted enrichment. *Plant J.* **71**, 517–526.

Gawroński, P., Ariyadasa, R., Himmelbach, A. *et al.* (2014) A distorted circadian clock causes early flowering and temperature-dependent variation in spike development in the *Eps-3A* mutant of einkorn wheat. *Genetics*, **196**, 1253–1261.

Gu, Y.Q., Coleman-Derr, D., Kong, X. and Anderson, O.D. (2004) Rapid genome evolution revealed by comparative sequence analysis of orthologous regions from four Triticeae genomes. *Plant Physiol.* **135**, 459–470.

Hartwig, B., James, G.V., Konrad, K., Schneeberger, K. and Turck, F. (2012) Fast isogenic mapping-by-sequencing of ethyl methanesulfonate-induced mutant bulks. *Plant Physiol.* **160**, 591–600.

Hazen, S.P., Schultz, T.F., Pruneda-Paz, J.L., Borevitz, J.O., Ecker, J.R. and Kay, S.A. (2005) *LUX ARRHYTHMO* encodes a Myb domain protein essential for circadian rhythms. *Proc. Natl Acad. Sci. USA*, **102**, 10387–10392.

Huang, S., Sirikhachornkit, A., Su, X.J., Faris, J., Gill, B., Haselkorn, R. and Gornicki, P. (2002) Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. *Proc. Natl Acad. Sci. USA*, **99**, 8133–8138.

Koboldt, D., Zhang, Q., Larson, D., Shen, D., McLellan, M., Lin, L., Miller, C., Mardis, E., Ding, L. and Wilson, R. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R.; **1000 Genome Project Data Processing Subgroup** (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lukowitz, W., Gillmor, C.S. and Scheible, W.R. (2000) Positional cloning in Arabidopsis. Why it feels good to have a genome initiative working for you. *Plant Physiol.* **123**, 795–805.

Michelmore, R.W., Paran, I. and Kesseli, R.V. (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Genetics*, **88**, 9828–9832.

Minevich, G., Park, D., Blankenberg, D., Poole, R.J. and Hobert, O. (2012) CloudMap: a cloud-based pipeline for analysis of mutant genome sequences. *Genetics*, **192**, 1249–1269.

Nordstrom, K.J.V., Albani, M.C., James, G.V., Gutjahr, C., Hartwig, B., Turck, F., Paszkowski, U., Coupland, G. and Schneeberger, K. (2013) Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. *Nature Biotech.* **31**, 325–331.

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A.H., Nielsen, K.L., Jørgensen, J., Weigel, D. and Andersen, S.U. (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods*, **6**, 500–501.

Shindo, C. and Sasakuma, T. (2001) Early heading mutants of *T. monococcum* and *Ae squarrosa*, A- and D-Genome ancestral species of hexaploid wheat. *Breed. Sci.* **51**, 95–98.

The iPlant Collaborative. The iPlant Collaborative: Cyberinfrastructure for Plant Biology (2011). https://www.iplantcollaborative.org.

Trick, M., Adamski, N., Mugford, S., Jiang, C., Febrer, M. and Uauy, C. (2012) Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat. *BMC Plant Biol.* **12**, 14.

Wicker, T., Yahiaoui, N., Guyot, R., Schlagenhauf, E., Liu, Z.D., Dubcovsky, J. and Keller, B. (2003) Rapid genome divergence at orthologous low molecular weight Glutenin loci of he a and am genomes of wheat. *Plant Cell*, **15**, 1186–1197.

Wilkinson, P.A., Winfield, M.O., Barker, G.L.A., Allen, A.M., Burridge, A., Coghill, J.A. and Edwards, K.J. (2012) CerealsDB 2.0: an integrated resource for plant breeders and scientists. *BMC Bioinformatics*, **13**, 219.

Winfield, M.O., Wilkinson, P.A., Allen, A.M. *et al.* (2012) Targeted re-sequencing of the allohexaploid wheat exome. *Plant Biotechnol. J.* **10**, 733–742.

Yang, Y.W., Lai, K.N., Tai, P.Y. and Li, W.H. (1999) Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. *J. Mol. Evol.* **48**, 597–604.