Genome assembly and quality control for non-model organisms

Bernardo J. Clavijo

A thesis submitted to the School of Biological Sciences of the University of East Anglia in partial fulfillment of the requirements for the degree of
Doctor of Philosophy by publication

December 2019

I would like to dedicate this thesis to
my loving wife Fiorella, and my wonderful daughter Clara . . .

Abstract

This thesis presents my work in genome assembly between 2010 and 2019.

Chapter 1 is an introduction to the status of the field, presenting the challenges and opportunities on generating *de novo* genome assemblies. Chapter 2 presents the development of k-mer spectra validation for assembly completeness, from its beginnings as unique sequence coverage analyses, through its implementation in the K-mer Analysis Toolkit, up to its use to assess consensus accuracy on hybrid assemblies. Chapter 3 describes a series of objective guided *de novo* assembly strategies applied to non-model genomes, starting with the assembly of the medicinal plant *C. roseus* to investigate its biosynthesis pathways, continuing with the chromosome-scale assembly of the ash dieback fungus during the UK outbreak, and concluding with my work assembling the hexaploid wheat genome from whole genome shotgun short read data. Chapter 4 describes the creation of haplotype-collapsed assemblies for 16 specimens of *Heliconius* butterflies to enable evolutionary analyses, and presents the Sequence Distance Graph framework to work with genome graphs and multi-technology data integration as a step towards haplotype-specific assemblies. Finally, Chapter 5 discusses this research and its impact in the context of the present and future of the field.

Acknowledgements

A Ph.D. by publication is, by definition, a long winding road. I only made it this far because a lot of people were there for me along the way, and there is no way I can fit everyone's name in here. Still, I would like to specially thank some of them.

First I would like to thank my parents Raúl and Teresa, and my three younger brothers José, and his wife Agustina; Martín; and Javier, his wife Emilia, and their children Inés, Andrés, Cecilia, Ana, Margarita, and María José. Our house was always a place of learning and challenge, and music; and very much the foundation to develop our minds forward.

My journey into science was rather complicated. Thanks to Gonzalo Bellanza for being absolutely crazy and putting up with all kinds of strange working patterns in our company while I started my journey. To Sergio Lew, who realised that a bad student need not be a bad researcher and put me in contact with the INTA crew. To everyone at INTA, but specially Norma Paniego, who was my first mentor in science and still pushes me in the right direction when needed. To all the good people at EI/TGAC, there's too many of you to name you all but you know who you are, and everything we went through together.

I have been lucky enough to work with and receive support and encouragement from some amazing scientists during these years. Thanks Federica Di Palma for providing mentoring and guidance, and for putting up with openly unorthodox ways of doing research. Thanks Manfred Grabherr, James Mallet, Richard Harrison, and James Cuff; for collaborations and conversations that have inspired me to move forward. Thanks Neil Hall and Christine Fosker, for a lot of support at EI. Thanks to my extraordinary research group: Gonza, Jon, Luis, Ben, Cam and Kat; you know how much of my work here would not have been possible without you. And finally thanks Federica Di Palma and Wilfried Haerty for pushing me through this final step of writing my Ph.D. by publication and making it a coherent whole.

Thank you to everyone that gave us personal support through these last 8 years in the UK. Thanks to John and Tineke, Peter and Jackie, and Colin and Linda; you made Norwich a fine city and England a merry country, and will always be our family away from home. Thanks to every member of Slyppery Styxx: Tony, Jon, Rob, Graham, and Jules, I may be a terrible guitar player but I really enjoy playing with you. Thanks to Panos and Elpida, and Nestor; I am rounder, and happier, because of you.

Special thanks to my family in law. Armando and Susi, that spent months with us in Cambridge helping look after their little granddaughter so we could work and study. Pablo and Lucía who were always eagerly at the other side of a Skype call. And Dave and Carolina, who's Dublin flat provided a much needed refuge when halfway through my writing I could not get it going. I owe half of the text in this thesis to that week of family support, peace, and afternoon running.

Last, and mainly, I want to thank my wife Fiorella. She got me interested in biology when I was looking for some science back in my life, and was there to encourage me all the way. Through this year compiling these publications and writing, she selflessly put her own Ph.D. studies second, keeping our small family going on and our little daughter entertained and happy. And Clara, our precious little girl, whose arrival and keenly interested eyes have signaled me it was time to stop procrastinating and start writing my thesis. Mainly, so I could get more guilt-free evenings of endless playtime.

Table of contents

List of figures

List of submitted publications

This thesis is based on the following manuscripts, listed with a succinct description of their scope and my contributions to them. Appendix I contains letter of support from my co-authors, and Appendix II contains the original publications with the kind permission of the relevant journals.

[1] M. Helguera, M. Rivarola, **B. Clavijo**, M. M. Martis, L. S. Vanzetti, S. González, I. Garbus, P. Leroy, H. Šimková, M. Valárik, M. Caccamo, J. Doležel, K. F. X. Mayer, C. Feuillet, G. Tranquilli, N. Paniego, V. Echenique. **New insights into the wheat chromosome 4D structure and virtual gene order, revealed by survey pyrosequencing.** *Plant Science.* 2015;233:200-212. doi:10.1016/j.plantsci.2014.12.004

This publication presented assemblies for both arms of chromosome 4D of *T. aestivum* based on 454 data from flow-sorted libraries and an analysis of their gene content. I contributed: analysis of the raw data, genome assemblies, analysis of content representation and biases, and writing of relevant sections of the manuscript.

[2] D. Mapleson, G. Garcia Accinelli, G. Kettleborough, J. Wright, **B. J. Clavijo**. **KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies.** *Bioinformatics.* 2017;33(4):574-576. doi:10.1093/bioinformatics/btw663

This publication presented software to analyse and compare k-mer spectra. I contributed: design and testing of all methods and software, programming of the early prototypes, and writing of the manuscript, as well as direction of the whole project.

[3] A. V. Zimin, D. Puiu, R. Hall, S. Kingan, **B. J. Clavijo**, S. L. Salzberg. **The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum.** *Gigascience.* 2017;6(11). doi:10.1093/gigascience/gix097

This publication presented an assembly of *T. aestivum* based on Illumina and PacBio data. I contributed: quality control by k-mer spectra comparison to an independet dataset, analysis of consensus quality, and writing of relevant sections of the manuscript.

[4] F. Kellner, J. Kim, **B. J. Clavijo**, J. P. Hamilton, K. L. Childs, B. Vaillancourt, J. Cepela, M. Habermann, B. Steuernagel, L. Clissold, K. McLay, C. R. Buell, S. E. O'Connor. **Genome-guided investigation of plant natural product biosynthesis.** *The Plant Journal.* 2015;82(4):680-692. doi:10.1111/tpj.12827

This publication presented an assembly of *C. roseus* based on Illumina data, and its use to investigate pathways of biosynthesis. I contributed: genome assembly, analyses of transcript reconstruction and copy number variation, and writing of relevant sections of the manuscript.

[5] R. M. Leggett, **B. J. Clavijo**, L. Clissold, M. D. Clark, M. Caccamo. **NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries.** *Bioinformatics.* 2014;30(4):566-568. doi:10.1093/bioinformatics/btt702

This publication presented software to pre-process Nextera long mate pair libraries. I contributed: evaluations of pre-processing effect on genome assembly described in the example results, and writing of relevant sections of the manuscript.

[6] M. McMullan, M. Rafiqi, G. Kaithakottil, **B. J. Clavijo**, L. Bilham, E. Orton, L. Percival-Alwyn, B. J. Ward, A. Edwards, D. G. O. Saunders, G. Garcia Accinelli, J. Wright, W. Verweij, G. Koutsovoulos, K. Yoshida, T. Hosoya, L. Williamson, P. Jennings, R. Ioos, C. Husson, A. M. Hietala, A. Vivian-Smith, H. Solheim, D. MaClean, C. Fosker, N. Hall, J. K. M. Brown, D. Swarbreck, M. Blaxter, J. A. Downie, M. D. Clark. **The ash dieback invasion of Europe was founded by two genetically divergent individuals.** *Nature Ecology & Evolution.* 2018;2(6):1000. doi:10.1038/s41559-018-0548-9

This publication presented an assembly of a british *H. pseudoalbidus* sample based on Illumina paired and long mate paired data, its annotation, and a phylogenetic analysis of European and Japanese varieties suporting a hypothesis for only two divergent individuals as founders of the European population. I contributed: design of the sequencing strategy, raw dataset quality control, genome assemblies and advice using them to evaluate biological hypothesis, and writing of relevant sections of the manuscript.

[7] D. Heavens, G. Garcia Accinelli, **B. Clavijo**, M. D. 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;59(1):42-45. doi:10.2144/000114310

This publication presented improvements to the protocol to produce Nextera long mate pair libraries. I contributed: evaluation of quality and usability of the libraries throughout the development of the protocol, and input on discussions over requirements for genome assembly.

[8] **B. J. Clavijo**, L. Venturini, C. Schudoma, G. Garcia Accinelli, G. Kaithakottil, J. Wright, P. Borrill, G. Kettleborough, D. Heavens, H. Chapman, J. Lipscombe, T. Barker, F.-H. Lu, N. McKenzie, D. Raats, R. H. Ramirez-Gonzalez, A. Coince, N. Peel, L. Percival-Alwyn, O. Duncan, J. Trösch, G. Yu, D. M. Bolser, G. Namaati, A. Kerhornou, M. Spannagl, H. Gundlach, G. Haberer, R. P. Davey, C. Fosker, F. D. Palma, A. L. Phillips, A. H. Millar, P. J. Kersey, C. Uauy, K. V. Krasileva, D. Swarbreck, M. W. Bevan, M. 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 Res.* 2017;27(5):885-896. doi:10.1101/gr.217117.116

This publication presented the first hexaploid wheat genome assembly to effectively recover all the genic space, with a detailed annotation. I contributed: design of the sequencing strategy, raw dataset quality control, design and programming of the assembly pipeline, and genome assemblies, alongside genome assembly quality control and writing of relevant sections of the manuscript.

[9] N. B. Edelman, P. B. Frandsen, M. Miyagi, **B. Clavijo**, J. Davey, R. B. Dikow, G. García-Accinelli, S. M. V. Belleghem, N. Patterson, D. E. Neafsey, R. Challis, S. Kumar, G. R. P. Moreira, C. Salazar, M. Chouteau, B. A. Counterman, R. Papa, M. Blaxter, R. D. Reed, K. K. Dasmahapatra, M. Kronforst, M. Joron, C. D. Jiggins, W. O. McMillan, F. D. Palma, A. J. Blumberg, J. Wakeley, D. Jaffe, J. Mallet. **Genomic architecture and introgression shape a butterfly radiation.** *Science.* 2019;366(6465):594-599. doi:10.1126/science.aaw2090

This publication presented multiple de novo genome sequences, and analyses showing the importance of introgression and selective processes in adaptive radiation. I contributed: raw dataset quality control, primary genome assemblies, haplotype collapsing and re-scaffolding, and writing of relevant sections of the manuscript.

[10] L. Yanes, G. Garcia Accinelli, J. Wright, B. J. Ward, **B. J. Clavijo**. **A Sequence Distance Graph framework for genome assembly and analysis.** *F1000Res.* 2019;8:1490. doi:10.12688/f1000research.20233.1

This publication presented a framework to perform analyses and manipulation of genome graphs, applied to genome assembly and analysis pipelines. I contributed: design of the framework, prototyping, programming, and writing of the manuscript, as well as direction of the whole project.

Chapter 1

Introduction

Genome sequencing changed our understanding of biology. From high-level population dynamics and evolution to the molecular function of single proteins, the data provided by sequencers is shining new light and showing new processes. As with any observational tool, this new understanding of biology is both advanced and constrained by the combination of sequencing technologies and analysis methods. As sequencing technologies rapidly evolve and change, better assembly methods often support new interpretations of the data to generate new knowledge.

1.1 Sequencing and assembly

Computer-based whole genome shotgun assembly [11] brought sequencing and assembly firmly together, shortly after the publication of the 5.4Kbp DNA sequence of bacteriophage ΦX174 [12]. With sequencing reads multiple orders of magnitude shorter than chromosomes, assembling these short fragments, or at least aligning them against a reference, has always been part of any sequencing project. Assemblers are developed to exploit new sequencing data, and sequencing methods devised to produce better data for assembly. The characteristics of sequencing data are the key technical parameter of the assembly problem, alongside the more important, but mainly fixed, parameters of a particular genome's sequence size and complexity. The interplay between these parameters, and the development of clever techniques to manipulate them, underpins the history of the field.

As shown in Figure 1.1, a genome assembly process can be broken down in three main steps: sequencing, contig construction and scaffolding [13]. Multiple times the size of the genome is sequenced in reads, producing overlaps in their sequence that will inform the assembly process. Contigs, contiguous sequences from a set of reads representing the same region of the genome, are constructed first by an assembler by joining reads that share the same sequence. Scaffolds are then constructed using longer range information to orient and order these contigs. This means contigs will have more precise local sequence information, and scaffolds will have longer structural

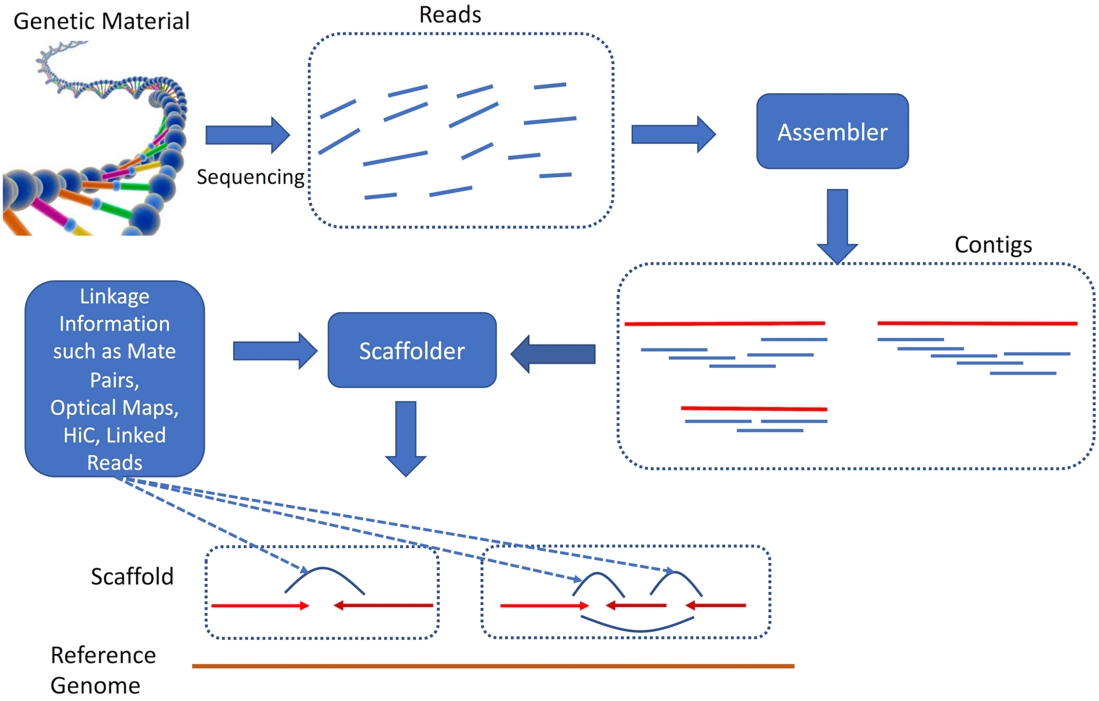


Fig. 1.1 Overview of the genome assembly process. Genetic material is sequenced into reads. An assembler joins reads' sequences via their overlaps into contigs. A scaffolder connects, orients and orders contigs into scaffolds, relying on a variety of linkage information. (Reproduced from Ghurye and Pop [13], under CC-BY 4.0)

information, but may introduce uncertainties at specific points inside them, including gaps of unknown sequence between consecutive contigs. Because of the computational hardness of the problem, and the difficulty in perfectly modelling each type of data, assemblers and scaffolders are essentially heuristic [14].

1.1.1 Sanger sequencing and Overlap Layout Consensus

The first computer-based whole genome shotgun assembly package executed a greedy overlap collapse algorithm, repeatedly finding overlapping sequences in a set and replacing them by their resulting, longer, concatenation [11]. When automatic capillary sequencers increased sequencing throughput, the overlap-layout-consensus (OLC) approach became a standard: find all overlaps between the fragments, arrange the fragments satisfying the overlaps, and produce a consensus from the overlapping fragments at each position [15]. Soon after, the layout problem was formalised as an overlap graph, with sequences as vertices and overlaps as edges, as shown in Figure 1.2B [16]. These formulations underpin all genome assemblers used for Sanger sequences.

To sequence larger genomes, the hierarchical shotgun sequencing reduced the complexity of the assembly problem via a divide-and-conquer approach. Fragments of the target genome were cloned in bacteria to create a Bacterial Artificial Chromosome (BAC) clone library. A physical map of the positions of these these clones was produced and a tiling path of clones, covering the whole of the genome with little overlap between them, was then chosen for shotgun sequencing. Each clone was then assembled separately using OLC contig assemblers such as CAP [17] and PHRAP [18], and their assembled contig sequences were scaffolded together using the map and sequence information. This process left gaps where the restriction maps were not precise enough or the sequencing or assembly failed. These gaps were closed during genome finishing, a laborious process using techniques that went from re-selecting other clones in the region to be sequenced to chromosome walking. Hierarchical shotgun assembled the 12Mbp genome of *S. cerevisiae* in 1996 [19], the 97Mbp genome of *C. elegans* in 1998 [20], and the 125Mbp genome of *A. thaliana* in 2000 [21].

In 1995 the TIGR ASSEMBLER successfully assembled the 1.8Mbp genome of the bacterium *Haemophilus influenzae* from Whole Genome Shotgun (WGS) reads, sequenced randomly from the whole genome without constructing a physical map, by exploiting paired end sequencing to scaffold the sequences [22]. The CELERA assembler used a similar technique to produce the genome of the fruit fly, *Drosophila melanogaster* in 2000 [23].

Two drafts of the Human Genome were published in 2001. One was the hierarchical shotgun draft by the public project of the Human Genome Sequencing Consortium [24], published before the finishing process. The other was a WSG assembly by the private company Celera [25], including data generated from Celera and data from an earlier release of the public project. In 2003, the finished result of the hierarchical shotgun Human Genome Project was unveiled, alongside more detailed analyses and annotation, and it was a clear improvement over both drafts [26]. But, amid the controversy, WGS had been shown to work in the human genome.

The differences in time and cost as sequencing technologies evolved contributed to make WGS a serious contender for genome assembly. Besides the already mentioned TIGR and CELERA assemblers, ARACHNE 2 [27] gained popularity for WGS after its 2002 assembly of the mouse genome [28] with greater emphasis on using the distance information between paired reads. The EULER assembler, using a de Bruijn Graph (DBG) to represent the assembly problem, broke with the OLC paradigm, enabling a much simpler representation for repetitive regions and increased computational performance [29]. Its principles would become the basis of the incoming wave of Next Generation Sequencing assemblers.

1.1.2 Next Generation Sequencing and de Bruijn Graphs

In 2005, Roche's 454 pyrosequencing introduced massively parallel sequencing [31]. These comparatively cheap high throughput sequencers produced less accurate shorter reads, at a fraction of the cost and time, allowing larger coverages. Roche's Newbler OLC assembler was typically used to assemble 454 data, but other assemblers originally developed for Sanger sequences were also popular, such as MIRA [32].

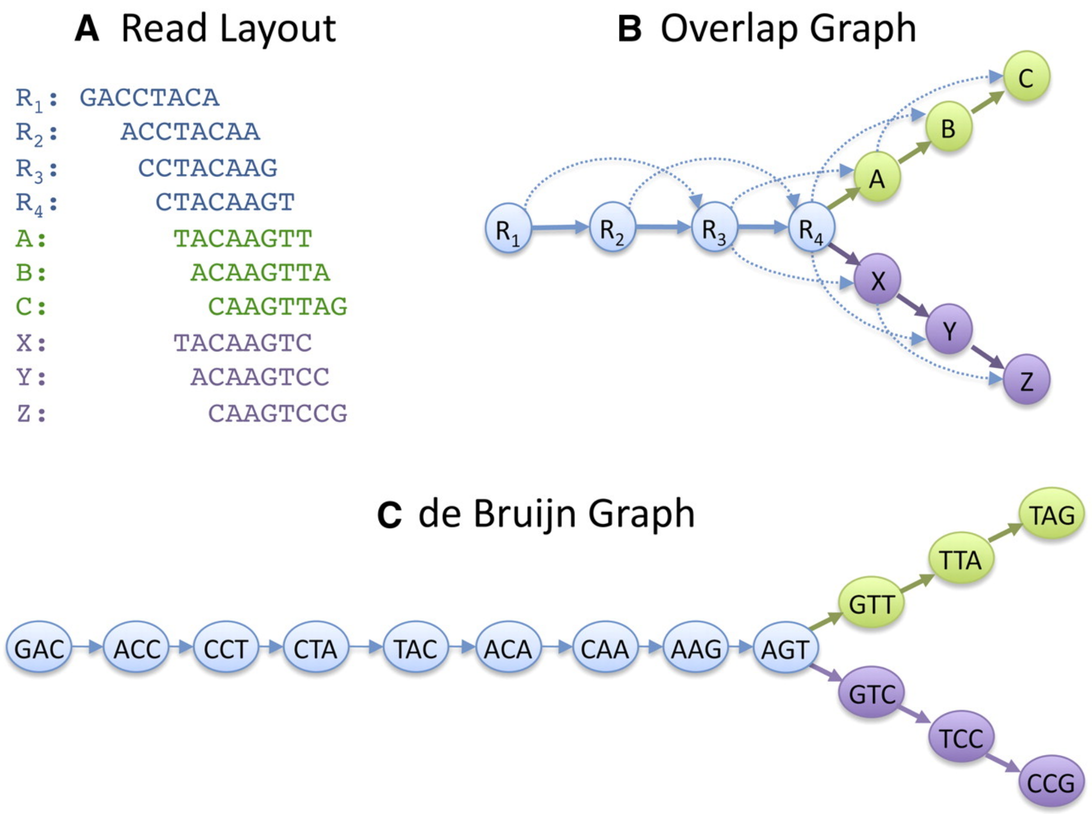


Fig. 1.2 Overlap and de Bruijn Graphs for assembly. From the set of reads (A), we can build an overlap graph (B) where each read is a node and overlaps >5bp are directed edges. Transitive overlaps are shown as dotted edges. In a de Bruin graph (C), a node is created for every k-mer in the reads; here with k = 3. Edges connect successive k-mers in a read, which overlap by k - 1 bases. In both approaches, repeat sequences create a fork. (Reproduced from Schatz, Delcher, and Salzberg [30], under CC-BY 4.0)

The Solexa/Illumina Genome Analyzer, introduced in 2006, produced even smaller 36bp reads, with higher throughput and lower cost [33]. The computational cost of generating *de novo* assemblies from so many reads naturally led to a reintroduction and popularisation of DBG assemblers [34].

In a DBG assembler, as shown in Figure 1.2C, every read is decomposed into k-mers, and consecutive k-mers are connected. This replaces the computationally expensive step of finding all overlaps by simple lookup of each k-mer in the read, but does not keep track of the position of each read in the graph. To reintroduce the linkage information from the reads, DBG assemblers generally remap the reads to the graph either during construction or after graph simplification [34]. Besides contigs and scaffolds, DBG assemblers introduce a shorter, more precisely defined unit of sequence reconstruction: the unitig, formed by a simple chain of k-mers in the graph with no possible forks.

Several DBG assemblers were introduced shortly after Illumina data became available. EULER-SR was a straight optimisation of the pioneering EULER DBG assembler, mostly adapted to work with 454 and Illumina single reads [35]. Velvet, on the other hand, was introduced as a set of algorithms working with a DBG representation of the assembly problem, mainly aimed at Illumina reads, exploiting their short pairs and allowing flexibility of algorithms and parameters [34]. ALLPATHS used a mixture of different pair sizes to extend the single-copy unitigs, effectively bridging simple repeats at their sides [36].

The large number of short reads needed to assemble large genomes with WGS needed specific optimisations. ABySS introduced parallel computing [37]. ALLPATHS-LG specified a sequencing recipe designed for mammalian genomes, using shorter fragment sizes to sequence overlapping reads, effectively constructing longer single reads which allowed larger k values in the DBG construction [38]. SOAPdenovo [39] and SOAPdenovo2 [40] removed all read placement from the DBG construction step, relying on faster remapping of the reads later, while introducing further improvements to scaffolding heuristics.

Since a DBG assembler is effectively representing all perfect overlaps of size k-1, larger values of k make these overlaps longer and the graph simpler by distinguishing similar regions of the genome. But the values of k are limited by the effect of errors, sequencing coverage, and read length [14]. Pre-processing techniques for merging overlapping paired reads such as FLASH [41], and later increases in read size, made it possible to increase the value of K. Since an increased K value presents a trade-off on specificity over sensitivity, multi-K approaches, essentially assembling different parts of the graph at different K values, were implemented on assemblers such as SPAdes [42], SOAPdenovo2 [40]. With the introduction of PCR-free libraries, and read sizes reaching 250bp, DISCOVAR denovo, initially developed as a reference-based local assembler to analyse variants, was developed into a full genome assembler [43]. At present, DISCOVAR denovo produces the more contiguous assemblies with a reasonable tradeoff of accuracy [44].

The string graph [45], a condensed representation of the assembly problem from a set of strings, led to the development of SGA [46]. A string graph can represent all the information in the reads, while a DBG representing every k-mer in the reads would have too many entries and use too much memory, and an OLC assembler could hit both problems with comparing all reads or appropriately identifying the overlaps. Fermi, another application of the string graph, was used as a variant caller. [47].

Hybrid assemblers, allowing combinations of Sanger, 454, and Illumina data, were used to combine the strengths of each technology. Early success to mix 454 data with Sanger data was achieved by simply assembling the 454 with Newbler to create "600bp overlapping pseudoreads" [48]. The CABOG [49] assembler was later developed to effectively combine 454 and Sanger data. The concept of super-reads used in MaSuRCA [50] potentially enables any kind of sequence mapped over a DBG to produce long artificial reads that can be assembled with a modified version of CABOG. This type of hybrid algorithm would dominate the adoption of the next wave of sequencing.

1.1.3 Third Generation Sequencing, Overlap and String Graphs, and polishing

In 2011, PacBio released its RS sequencer, producing longer reads with lower accuracy via single molecule real time sequencing. These low accuracy reads could not be yet assembled on their own, but used to untangle a DBG from Illumina data in ALLPATHS-LG, they produced complete microbial genomes using single-molecule sequences [51].

Error-correcting the long reads prior to assembly holds the promise of simplifying the assembly process and providing a computationally effective way to exploit their information. PBcR [52] the first algorithm to perform this correction, aligned illumina or 454 reads to PacBio reads, distributing reads from repetitive sequences by enforcing even coverage of the long reads. The resulting corrected long reads could then be simply assembled with the OLC Celera assembler. PBcR was updated to use PacBio only data for microbes by generating consensus between groups of long reads [53]. A similar hierarchical approach was later used by HGAP [54] to produce long-read-only assemblies, by choosing the longest reads and aligning shorter reads to them. MaSuRCA mega-reads extended the super-read concept from MaSurCA [55] for the same purposes. FALCON and FALCON-Unzip [56], while continuing with the same idea of hierachical assembly, try to preserve heterozygous differences through the correction steps, and use the corrected reads to generate a string graph which initially fuses the two haplotypes in paths with bubbles, but then phases them using the raw reads placed through their corresponding corrected read.

In 2015, Oxford Nanopore released its MinION sequencer, using nanopore technology to sequence even longer reads. While more experimental than PacBio, its low cost and portability alongside an early access program for researchers and the extremely long reads capable of spanning longer repeats are pushing its rapid adoption. At the same time, PacBio long high-fidelity (HiFi) reads can now achieve as high as 99.8% accuracy, by using adaptors to sequence multiple times over the same molecule to generate a circular consensus, thus sacrificing total length for consensus accuracy [57].

Modern long read assemblers are exploiting the new characteristics of both longer and more precise long reads. Miniasm, an assembler based on minimap, can assemble modern long reads raw, with no correction, although trading off some accuracy in its results [58]. Canu improved on the overlap detection between reads via k-mer weighting, and used a two-step correction process to increase the specificity, and offered an integrated an end-to-end pipeline based in an improved version of CA [59]. Flye introduces a new formulation of the assembly problem, repeat graphs, which should enable better heuristics for repeat resolution while preserving the different instances of a repeat [60] . Wtdbg2 uses a fuzzy DBG, which enables a large performance gain when constructing graph from raw long reads [61].

An unavoidable consequence of the use of long low accuracy reads is the loss of accuracy in the consensus step. Consensus polishing involves re-mapping the raw reads to the assembly, and changing the consensus to better represent the information on the mapped reads. Pilon [62] is very popular to correct with Illumina reads, like Quiver/Arrow [54] for PacBio reads, and Nanopolish [63] for Nanopore reads. Racon [64] was originally designed to rapidly improve the consensus of miniasm assemblies. While polishing is to a certain extent a finishing technique, it is limited not only by the mapping heuristics, but by the resolution achieved by the base assembly, and is itself a highly heuristic and iterative process. In that sense, polishing results are limited by, and will affect the results of, the steps that follow a contig assembly: scaffolding and phasing, specially in ploidy genomes and repetitive regions.

1.2 Scaffolding and phasing

Scaffolding, as shown in Figure 1.1, is the process of linking, orienting and ordering contigs into scaffolds representing larger regions of the genome. Phasing, a related part of the assembly problem, is the reconstruction of the original haplotypes by grouping their variants in the original phase.

Many assemblers perform some scaffolding as part of their heuristics, usually with paired or long reads, and some provide scaffolding modules that can be used with pre-assembled contigs, like ABYSS2 [44] or SOAPdenovo2 [40]. Different kinds of linkage data can contribute to scaffolding and phasing , including: paired reads, long reads, linked reads, optical maps, and contact information. With independent tools for many of them, there is still more work to be done in their integration, especially if we are to tackle more complex genomes [13].

By far the most studied case is that of scaffolding with paired reads, which were credited for the success of WGS assembly from its early years [22]. Paired read scaffolders such as Bambus [65] and SSPACE [66] start from shorter to longer libraries, progressively expanding the scaffolds. Others such as SOPRA [67] and MIP [68] use integer programming to optimise the orientation and ordering of contigs.

Linked reads, a technology initially developed by 10x Genomics, partitions long DNA fragments into droplets and generates tagged reads from them [69]. The 10x Genomics Supernova assembler, forked from DISCOVAR denovo, uses tag information to phase and scaffold haplotype blocks. ARKS [70] provides a 10x module that can be used with LINKS [71] to scaffold pre-assembled contigs, while Scaff10X [72] provides a stand-alone scaffolding pipeline.

Optical maps derived from marking restriction sites in individual DNA molecules, as provided by the Bionano Irys System, can be used to create hybrid scaffolds that combine sequence alongside restriction map data, extending or confirming the sequence scaffolds. There is a limited number of stand-alone scaffolders for Bionano data, besides the software provided by the manufacturer, and more generic optical mapping software as SOMA [73].

Information from Hi-C, a protocol developed initially to capture three dimensional chromosomal conformation, was shown upon the publication of LACHESIS [74] to be able to scaffold genomes to a chromosome-length level. SALSA [75], developed to scaffold long read assemblies, produced higher quality scaffolds without a prior for the number for chromosomes, and introduced correction of the input sequences by breaking putative misjoins. 3D-DNA [76] was later able to scaffold lower contiguity short read assemblies to chromosome level. SALSA2 [77] incorporates information of ambiguous joins from an assembly graph to improve scaffolding. Finally, ALLHIC [78] uses signal density to prune between-haplotype connections, improving reconstruction for polyploid genomes.

Phasing and haplotype resolution are two areas of the assembly problem in active development, and complex cases remain mostly unsolved [79, 13]. Tools for phasing diploid genomes by remapping raw data such as Whatshap [80] and LongRanger [81] try to both recall missed variation from the alternative haplotype discarded by the assembly process and phase it coherently. Diploid-aware assembly, as performed by FALCON-Unzip[56], phases bubbles in the collapsed graph to represent locally alternative haplotypes. When experimentally feasible, the use of a parental trio to bin parental variants in the offspring as implemented by TrioCanu [82] remains the best approach for fully chromosome-resolved phases.

1.3 Non-model organisms and complex genomes

One of the biggest challenges when analysing non-model organisms is the lack of knowledge about how different sequencing recipes and assembly pipelines perform. Given that assembly benchmarks consistently show differences in results between approaches that cannot be easily reconciled in *de novo* scenarios [83–85], standard methods to produce assemblies of equivalent characteristics and quality are key for comparative studies where differences are interpreted biologically.

Also, the characteristics of the genome itself may be particularly challenging. Large and repetitive genomes can make assembly challenging by having too few short unique sequences that would easily linearise a DBG or produce significant overlaps for OLC [3]. Varying levels of heterozygosity can generate complex graphs which are difficult to phase, untangle, and scaffold [82, 13]. Polyploid genomes, like those of many plants, can generate more complex structures in the assembly graphs as exemplified in Figure 1.3 [79]. In general, more complex karyotypes present harder assembly problems that have not yet been solved, and current tools may produce unreliable results [13]. This compounds with the lack of reliability in downstream analysis, which are also affected by not having been designed for the specifics of the non-model-organisms.

This thesis is based on my experience working on non-model organisms. As such, the Chapter 2 deals with quality control, with focus on completeness, the first condition

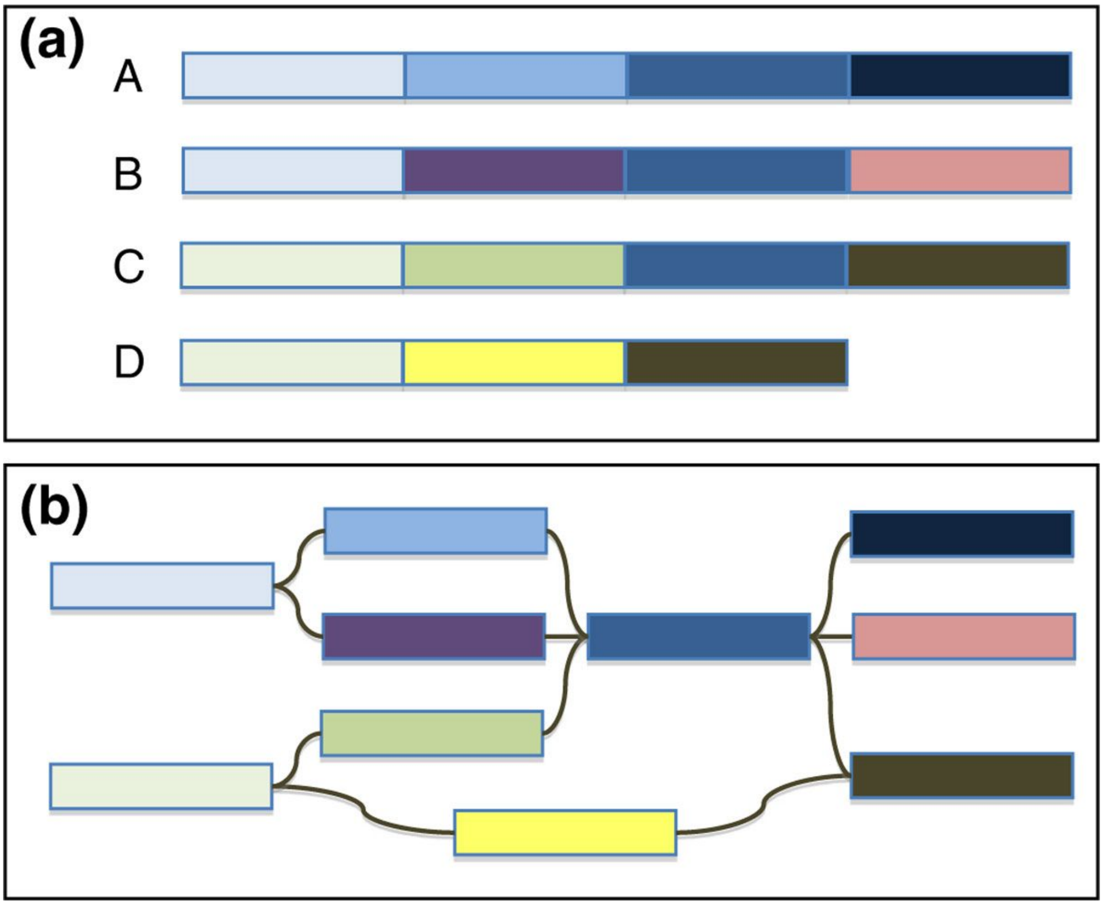


Fig. 1.3 Ploidy, heterozygosity and the assembly graph. (a) Schematic representation of a tetraploid genome, such as apple, cotton or cabbage, consisting of haploid chromosomes A to D with homozygosity/heterozygosity shown as different colored blocks. (b) Even without repeats or sequencing error, the assembly graph of the homozygous and heterozygous segments of the genome branch and intertwine in complex patterns. A plant-specific assembler would need to recognize these branching patterns and attempt to reconstruct the individual sequences for chromosomes A to D. (Reproduced from Schatz, Witkowski, and McCombie [79], under CC-BY 4.0)

of an assembly which remains surprisingly difficult to check in a non-model *de novo* scenario.

Examples of objective guided assemblies in Chapter 3 present particular challenges of non-model genome assembly. A plant genome assembly where full reconstruction of the genes involved in a biosynthetic pathway was complicated by previously unknown duplications. A fungal genome assembly where a complete lack of knowledge about kariotype meant differences between sequencing libraries needed to be analysed via k-mer spectra quality control, and unorthodox combinations of coverage were used to produce an ultimately chromosome-scale assembly. A set of butterfly genomes that needed to be assembled from only short-read paired end data up to a standard that would enable large-scale phylogenetic analyses, while dealing with varying levels of heterozygosity.

The hexaploid wheat genome, one of the most complex ever assembled, is large, full of repeats, and after inbreeding for homozygosity contains three similar haplotypes. At the time I started working on genome assembly it was one of the last standing projects meant to only be solvable by hierarchical shotgun sequencing. My short read WGS assembly not only surpassed all previous WGS assemblies and fulfilled the objective of recovering essentially all its genic content for the first time, but shifted the direction of the community. The hierarchical shotgun project is now finished, with the reference sequence based largely in a WGS assembly enhanced by the hierarchical data.

Chapter 4 first presents an approach to collapse heterozygosity in butterflies, to enable phylogenetic analyses over more contiguous assemblies. While I tackled all of these non-model genomes in an objective specific manner, I completely agree with the assembly community drive towards graph-based tools that can solve even the most complex genomes to a haplotype phased level. The Sequence Distance Graph framework I present as the last publication in this thesis, shows my current work in that direction: to keep untangling complexity, through its simple properties.

Chapter 2

K-mer spectra analyses for quality control

This chapter presents *k-mer* spectra analyses for quality control (QC) of raw reads and genome assemblies. These methods started taking shape when I performed all assemblies and basic quality control for both arms of wheat chromosome 4D [1], while my collaborators performed genic and syntenic analyses. Later on, I designed, prototyped and supervised the implementation of these analyses in KAT, the K-mer Analysis Toolkit [2], with Dan Mapleson eventually taking over the coding under my direction. I later used KAT to assess consensus accuracy of the hybrid PacBio-Illumina hexaploid wheat assemblies by Aleksey Zimin and others from Steven Salzberg's lab at Johns Hopkins University, contributing to improvements and an analysis section in that publication [3].

2.1 Background

Genome assemblies can be evaluated using three properties: *completeness*, or how much of the genome is assembled; *correctness*, or how few errors are introduced; and *contiguity*, or how long the assembled sequences are. These properties should be checked in order, because contiguity can easily be achieved by accepting erroneous joins or discarding discontiguous regions, and errors can easily be avoided by discarding the more complex regions in the genome. In practice, the opposite is true: contiguity, which is the easiest to compute, has become the single go-to metric to describe an assembly [86].

The N50 contiguity metric, defined as "the maximum length L such that 50% of all nucleotides lie in contigs (or scaffolds) of size at least L", was first used in a detailed paragraph alongside insights into the assembly process, followed by contiguity curves [24]. Unfortunately, this contiguity-only score, often out of context, is now accepted as a quality proxy for assemblies. The longstanding prevalence of N50 opened the field for incomplete and incorrect assemblies.

Completeness and correctness are often evaluated in objective-specific manners. In assemblies for differential expression analyses, RNA-seq mapping can measure completeness percentage of reads mapping, and correctness mapping quality. But in assemblies for large-scale structural variation analysis, completeness can be measured via homologous genes found, and correctness via conservation of syntenic blocks.

Many general approaches to measure completeness and correctness rely on sequence mapping. Mapping the assembly to a reference, as implemented in QUAST [87], and QUAST-LG [88]; measures completeness as reference coverage and correctness as identity. Mapping reads to the assembly, as implemented in amosvalidate [89], FRCbam [90, 91], REAPR [92] , ALE [86], and CGAL [93]; measures completeness as proportional to mapped reads, and correctness as coherence of coverage, pair orientation, and fragment size. Mapping core genes to the assembly, as implemented in CEGMA [94, 95] and BUSCO [96, 97]; measures both completeness and correctness in their reconstruction. While all these approaches are useful and work well in many cases, the mapping heuristics themselves introduce biases that affect their applicability.

Mapping methods, being heuristic, are more likely to fail in novel scenarios. Genome characteristics like repetitiveness, size, heterozygosity, and ploidy impose mapping biases. Most tools have been designed and tested only in specific genomes, with complete references, so complex karyotypes are often challenging. Finally, unmapped sequences need to be carefully accounted for and will always impose limitations. Paradoxically, *de novo* validation techniques based on mapping work best when your genome is less informative versus already known genomes, or less novel.

2.2 Coverage and completeness in wheat chromosome 4D

In 2011, I was working on assembling the 4D chromosome of the 17Gbp hexaploid bread wheat genome [1]. With their hierachical shotgun assembly project not yet finished, the international consortium decided to use flowsorted-chromosome-arm shotgun assemblies to obtain partial draft assemblies. The effort I was part of sequenced low coverage Roche 454 data, divided into single reads for contig construction and paired reads for scaffolding. A complementary effort with higher coverage Illumina data was conducted separately by the consortium [98].

This was a typical *de novo* project in a non-model organism: genome complexity limited our prior knowledge and quality control, and the data and assembly tools of the time were not prepared for the analyses. I quantified the biases introduced by flow sorting and amplification followed by low coverage sequencing, over two sets of known features of the wheat genome: insertion site based polymorphism (ISBP) markers, which are unique sequences easily annotated in a draft assembly; and a collection of binned expressed sequence tags (EST), sorted into 4 bins per chromosome arm. These two analyses later provided the main ideas behind KAT's spectra by copy number analysis.

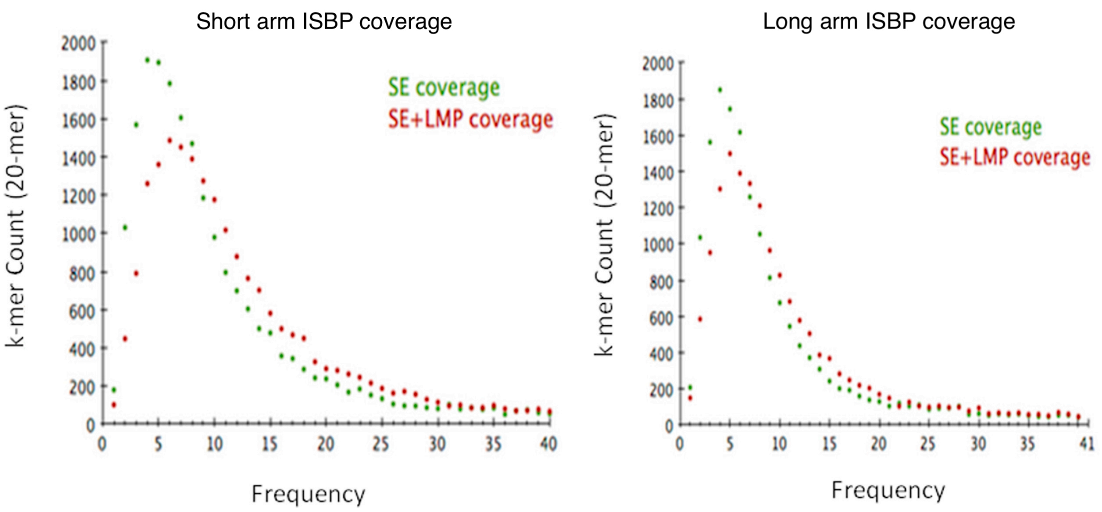


Fig. 2.1 Histograms of reads coverage for each ISBP detected in the wheat 4D assemblies. (Reproduced from Helguera, Rivarola, Clavijo et al. [1])

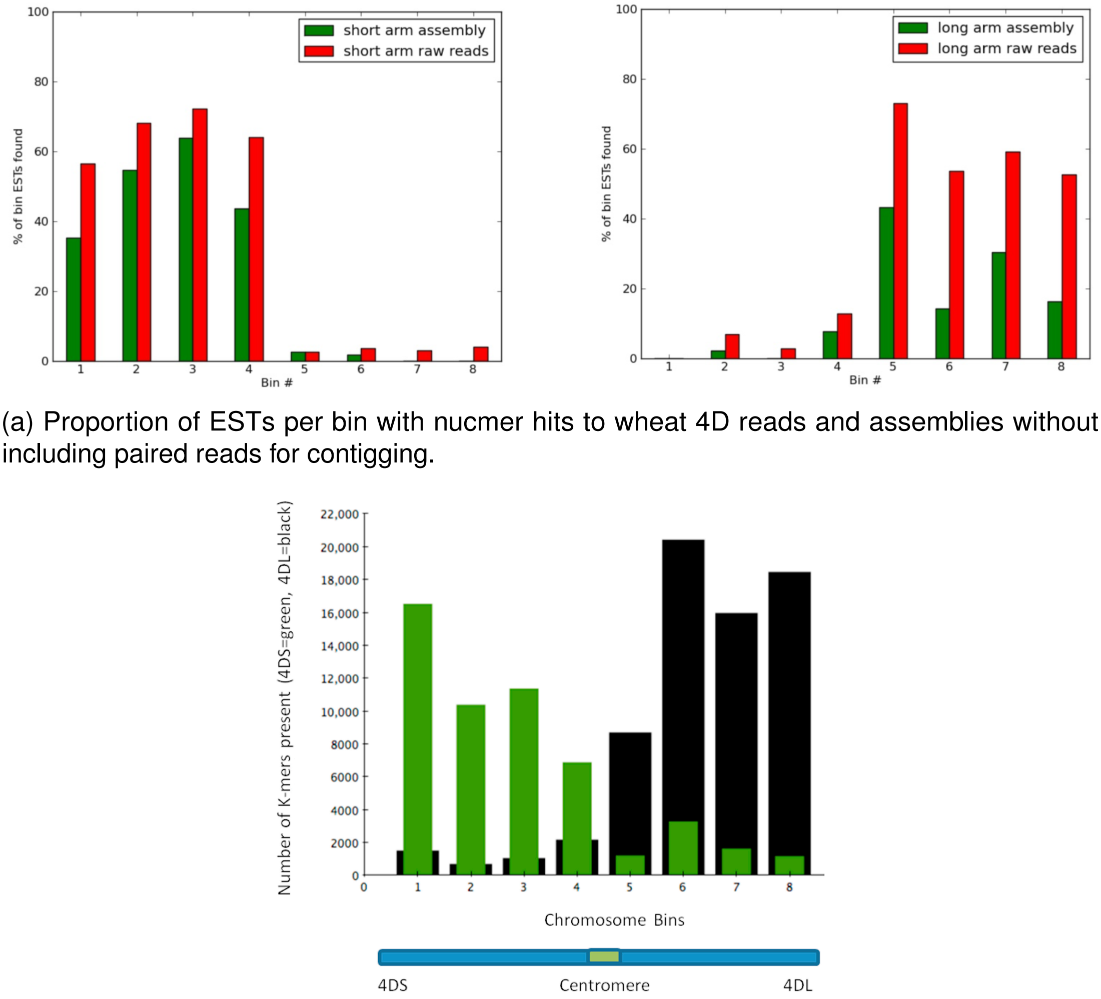
Treating each 20bp ISBP in the assembly as unique sequence, plotted a k-mer coverage distribution for all ISPBs by querying a jellyfish [99] database of the reads' 20-mer counts. The distributions in Figure 2.1 represent *k-mer* coverage of these unique sequences, showing critically low coverage which supported my decision to include paired reads into contig construction.

To assess gene space completeness, I measured coverage of the binned ESTs. My initial analysis, shown in Figure 2.2a and not included in the publication, computed coverage of each EST by nucmer[100] hits in the reads and the assemblies. It showed that, specially in the long arm, the assemblies were failing to recover all the EST content from the reads. I then started measuring completeness as inclusion of k-mers from ESTs into the assemblies, which removed biases from the mapping and increased performance. The final version of this analysis presented as raw k-mer counts rather than ESTs percentages also detected cross-contamination between chromosome arms, as shown in Figure 2.2b.

Later on, evaluating assemblies for the main IWGSC chromosome survey [98], I found the same ISBPs and ESTs were better covered by the less biased and higher coverage Illumina libraries. This was coherent with the use of only one sorting run for the 454 but four or more for the Illumina. Most content in the 454 assembiles was represented in the Illumina data, but parts of it could not be assembled with the very short Illumina reads of the time.

2.3 The K-mer Analysis Toolkit

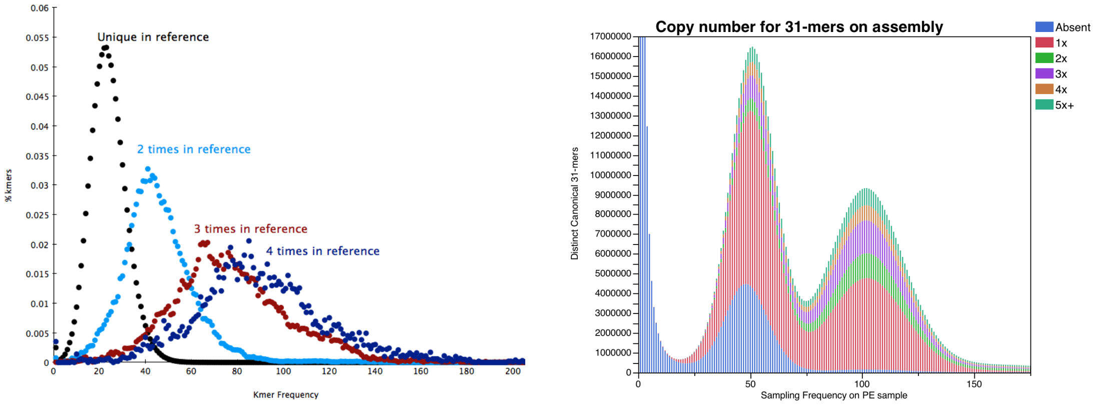
The K-mer Analysis Toolkit (KAT) has its roots in the ISBP and EST analyses presented in the previous section to analyse mostly unique content that should be present in the reads. I expanded the idea by plotting the distributions of coverage for each



(a) Proportion of ESTs per bin with nucmer hits to wheat 4D reads and assemblies without including paired reads for contigging.

(b) EST 20-mers in assemblies for the short arm (green bar) and long arm(black bar) of wheat 4D. (Reproduced from Helguera, Rivarola, Clavijo et al. [1])

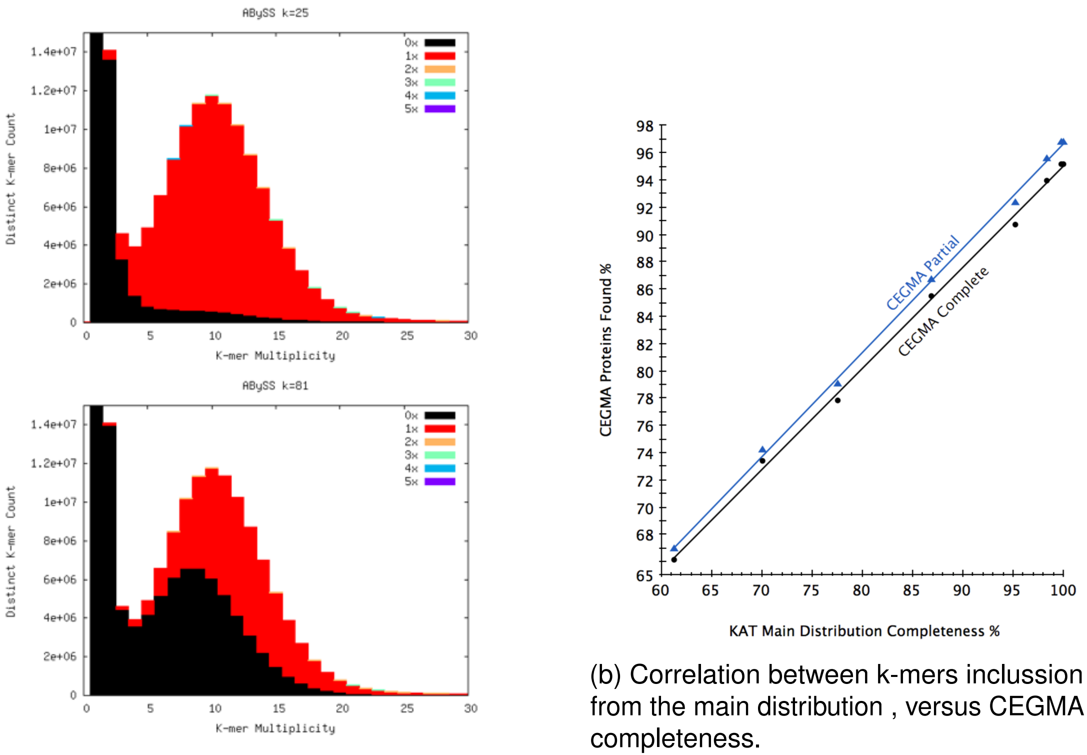
Fig. 2.2 Binned EST content on wheat 4D reads and assemblies.



(a) K-mer spectra density for each copy number on the yeast reference, from an illumina sample.

(b) A very early spectra-cn plot, for an abyss assembly of the heterozygous ash tree sample from "Tree 35".

Fig. 2.3 Early versions of the spectra-cn analysis.



(a) Spectra-cn showing assemblies with good completeness (top) and bad completeness (bottom). (Adapted from Mapleson, Garcia Accinelli, Kettleborough et al. [2])

(b) Correlation between k-mers inclussion from the main distribution , versus CEGMA completeness.

Fig. 2.4 Analysing completeness with KAT.

copy number of 31-mers in a reference genome, including copy number zero for 31-mers absent in the reference. The result in Figure 2.3a for a re-sequencing run of *S. cereviseae* S288c showed clear distributions enabling further analysis. Since a perfect assembly for a read set should constitute its correct reference, I soon started analysing assembly results with this method, deciding to use counts rather than densities and stack the histograms, effectively presenting a coloured partition of the reads' spectra given by the copy number in the assembly, as in Figure 2.3b.

KAT started as a simple hack to open two jellyfish hashes in memory and intersect their frequencies, computing the count of k-mers for each pair of frequencies (F_1, F_2). A matrix was then saved to disk with these counts for analysis. Later versions added predefined plots and some simple analysis of the distributions, alongside the *sect* tool to project k-mer coverage from a hash into a set of sequences, and a mode to divide the k-mers of a hash by their GC content. Comparing k-mer frequencies between two read sets can be informative also as QC, and to allow comparison between raw sequences. KAT contains a number of tools for these uses, described in its publication.

The spectra-cn plots from KAT are a good way to assess assembly completeness. Missing k-mers from the main distribution (black in Figure 2.4a) are the main signature of an incomplete assembly, and the percentage of completeness in the first peak has

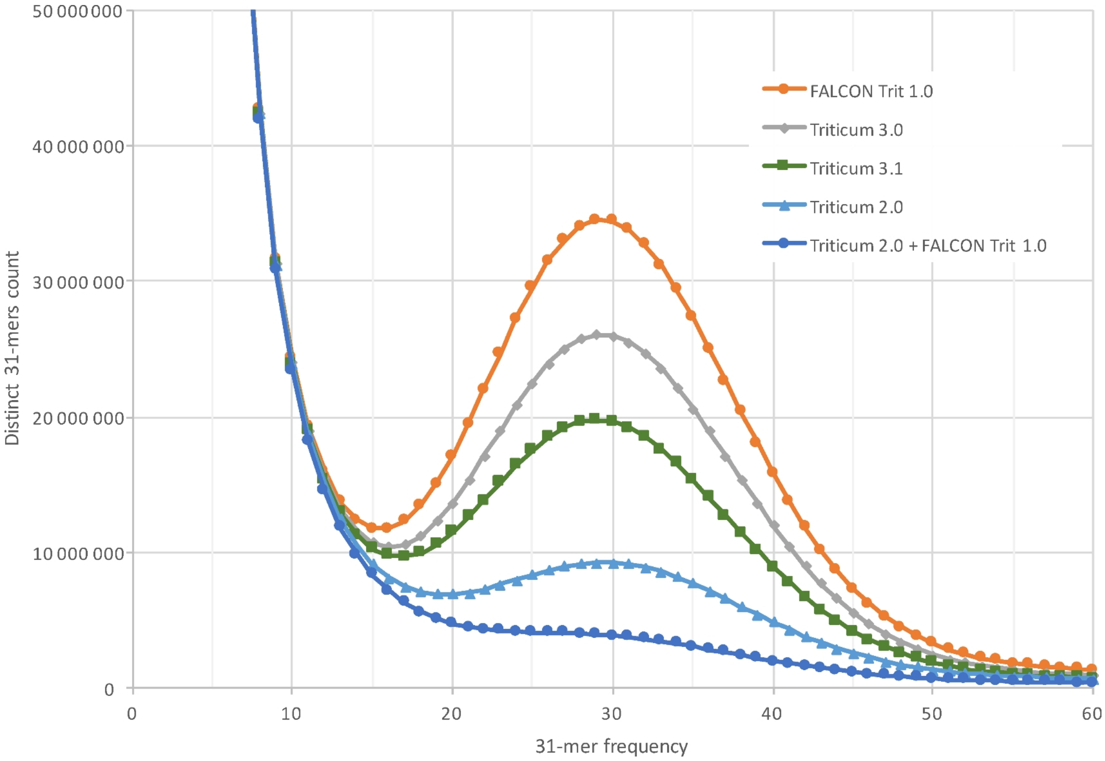


Fig. 2.5 Illumina spectra for 31-mers absent from WGS PacBio and hybrid assemblies using different assembly methods. Triticum 2.0, the MaSuRCA assembly, has the smaller distribution of missing 31-mers of any single assembly. The Triticum 3.1 assembly is an improvement over the Triticum 3.0 hybrid FALCON + MaSuRCA assembly that replaces the less accurate FALCON consensus with the correspongding MaSuRCA consensus for 98% of the assembly. (Reproduced from Zimin, Puiu, Luo et al. [55])

typically good correlation with gene presence scores as in Figure 2.4b. This type of analysis is probably the simplest and more widespread use of KAT.

A perfect spectra-cn plot is a necessary condition for a perfect assembly. It is not a sufficient condition, but given that it includes checks for completion and copy number, it should be the first check of an assembly. If a spectra-cn is perfect, or near enough, most assembly validation tools based on mapping will have minimum non-intrinsic biases.

2.4 Consensus quality on hybrid assemblies

While the structure of the genome is improved by creating assemblies based in long reads or hybrid datasets, the quality of the consensus can be lower than short read assemblies, especially for complex genomes. When a clean k-mer spectra dataset is available, the spectra-cn analysis can provide an insight on consensus quality by highlighting k-mers that are present in the main spectra distribution but have not been included in the consensus, and k-mers included in the consensus with incorrect copy numbers.

In 2017, I collaborated with the John Hopkins University team to analyse the consensus accuracy of their hybrid PacBio/Illumina wheat genome assembly combining results from FALCON [56] and MaSuRCA [3]. KAT spectra-cn analysis showed that MaSuRCA had better content representation. This prompted a change of strategy from merging the MaSuRCA assembly into the FALCON to merging the FALCON assembly into MaSuRCA, improving the overall result. Figure 2.5, shows a large proportion of the missing k-mers from MaSuRCA are included in FALCON, but some k-mers are absent in both. While these assemblies succeeded in recovering a larger structural portion of the wheat genome, with higher contiguity than my previous short read assemblies, no complete recovery of the wheat genome's k-mer spectrum was possible by combining these results.

2.5 Conclusion

Analysing the relation of k-mer frequencies between reads and assembly, sample and model, has provided many insights. It can highlight issues with consensus, and it has also been used to evaluate resolution of heterozygous or ploidy genomes, but most of all it has become a reasonable proxy to assess the first condition for genome assembly: completeness.

Chapter 3

Objective guided non-model genome assembly

This chapter presents genome assemblies with short reads, and the methods I used to deal with the challenges of non-model organism genome assembly, providing the basis for biologically relevant analysis. It starts with my assembly of the medicinal plant *Catharanthus roseus*, and my analyses to detect extra copies of key genes in its biosynthesis pathways. This work is included in a publication [4] where my collaborators annotated the genome and analysed biosynthesis pathways, showing the potential of a draft genome assembly for complex pathway analysis in non-model plants. It then presents my assembly of *Hymenoscyphus fraxineus* as part of the response to the ash dieback outbreak in the UK, using both paired end and NextClip-processed long mate paired (LMP) libraries to construct contigs, followed by scaffolding to chromosome-level pseudomolecules. NextClip was programmed by Richard Legget, and presented in a publication [5] where I contributed an evaluation of the effect of pre-processing on scaffolding. The *Hymenoscyphus fraxineus* assembly is part of a publication [6] where my collaborators annotated the genome and showed phylogenetic evidence for the European population to have been founded by only two divergent haploid individuals. Finally, the chapter concludes with my short read WGS *Triticum aestivum* assembly, the first to robustly capture essentially all the genic space of hexaploid bread wheat, using improved LMP construction and the w2rap assembly method. The improved LMP construction method, by Darren Heavens, was presented in a publication [7] where I contributed QC and feedback on library characteristics and usability. The *Triticum aestivum* assembly was presented in a publication [8] where, alongside my genome assembly and QC, various collaborators contributed annotation, RNA-seq analyses, comparisons to previous assemblies and biological analyses.

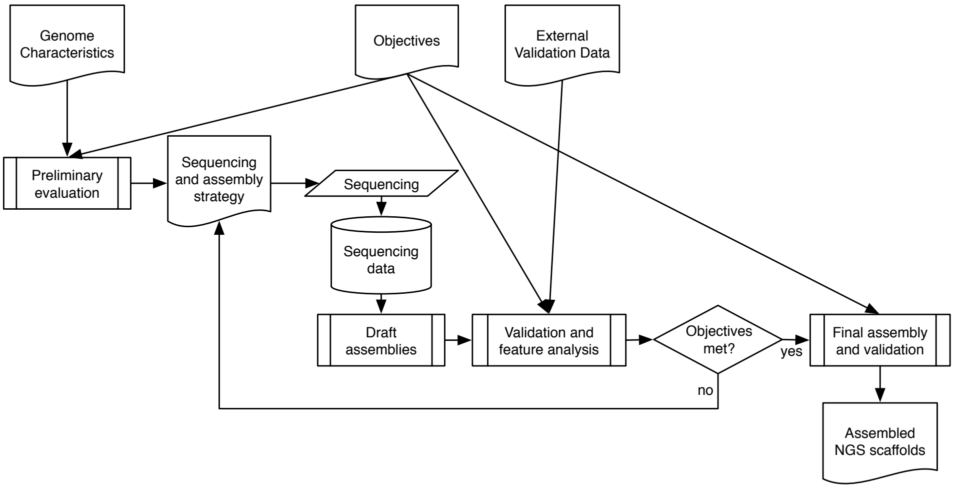


Fig. 3.1 A general process for genome assembly, guided by objectives, with a formal iteration process.

3.1 Background

Genome assemblers may become unreliable when the data and genome characteristics are different to those used to design them, and it is difficult to assess the quality of their output [79, 85, 13]. Defining a clear set of objectives and metrics is then a prerequisite to assemble non-model genomes that reliably support biologically relevant analyses. The two more widespread metrics for assembly are deeply flawed for *de novo* scenarios. Contiguity as measured by N50, is only important if completeness and correctness are satisfied, as discussed in Chapter 3. Gene reconstruction as measured by BUSCO [96, 97] is built on the corpus of known genes and will under-perform on novel datasets, but also the core genes are subjected to different evolutionary constraints than the rest of the genome and won't represent the full genic space appropriately. Objectives and metrics that are significant to the biological analyses and, if possible, make use of organism-specific data, are a better guide for non-model organism genome assembly.

3.2 Objective guided non-model assembly process

The process described in Figure 3.1 is centred around describing a set of objectives and characteristics for the genome assembly and validating them one by one to produce either a success or a well-defined failure. The process starts with an objective, such as reconstructing the genic space as defined by a set of previously known genes, or reconstructing the syntenic order vs. known species, or recovering end-to-end chromosome sequences defined either by synteny or by the presence of telomeric repeats at their ends. Objectives can also be defined as fully expanding a gene family or recovering more of it, recovering sequence around a set of markers, etc. Objectives may change along the way, but at any given time a set of attainable objectives with their associated metrics must guide the trade-off decisions across the whole project. These trade-off decisions will involve dimensions such as consensus accuracy, long-range information recovery (i.e. the true-positive part of contiguity), structural accuracy, and cost. The assumed characteristics for the genome, which dictate the attainability of these objectives need to be checked methodically and as soon as possible, as they are often the source of long-standing problems. Each iterations should confirm genome properties and attainable objectives, or update them.

I normally start by sequencing at least 30x coverage per haplotype in short reads, although it is possible now to use PacBio Hi-Fi reads in a similar way. This initial sequencing produces a k-mer spectrum that can be analysed using KAT or GenomeScope [101] and provides validation for assemblies during all iterations. The k-mer spectrum itself will confirm or update our knowledge about: genome size, GC content, heterozygosity, ploidy or haplotype count, and sequence uniqueness. A first-pass assembly of this data can provide a first measure of how challenging it can be to fulfil the original objectives, alongside sometimes a very good reconstruction of the genic space. Remapping the reads vs. the first-pass assembly can provide extra confirmation or information for many of the characteristics of the genome.

The next step is to improve the objective metrics through iterations. There will be different scenarios of objectives and genome characteristics, but this often requires increasing completeness, dealing with heterozygosity or ploidy, and increasing contiguity through repeat resolution [75]. The three examples presented here had different processes. *C. roseus* met the gene recovery objective in the first iteration, but required manual analyses of some duplicated genes. *H. pseudoalbidus* made good gene recovery on the first iteration, but required extra coverage and LMP data to increase contiguity through repeat resolution. *T. aestivum* failed during contigging on the first iteration but with partial assemblies showed good k-mer spectrum completeness; then produced contigs with good completeness on a second iteration, but its ploidy and repetitions generated fragmentation; and finally, through scaffolding with libraries chosen to match its repeat structure, met the objective of gene recovery.

When the original questions have been addressed, and the assembly strategy has been refined, it is advisable to rerun assembly with the chosen parameters for reproducibility. In some cases, computational requirements or time constraints may make this impractical.

If this process is followed, the main challenges are: to define completeness and a realistic final measurable goal, to choose and adapt a sequencing and assembly technique, to QC and prepare the data, and to validate that the final result is well supported by the data. There is no challenge on parametrising a tool when the parameters' effects can be measured.

3.3 Assembling genes to analyse pathways

The assembly of the *Catharantus roseus* genome to analyse its biosynthethic pathways was a straightforward case of objective-guided assembly. The goal was to recover the full gene set for monoterpene-derived indole alkaloid (MIA) biosynthetic pathway, preferably reconstructing promotors and gene clusters to enable a better study of regulation, with a secondary goal of generating an assembly recovering the genic space well to support further studies. As orthogonal data to test completeness and reconstruction of the pathway, and as proxy for the whole genic space, there were transcript assemblies of genes in the pathway, and partial assemblies of some of the involved loci.

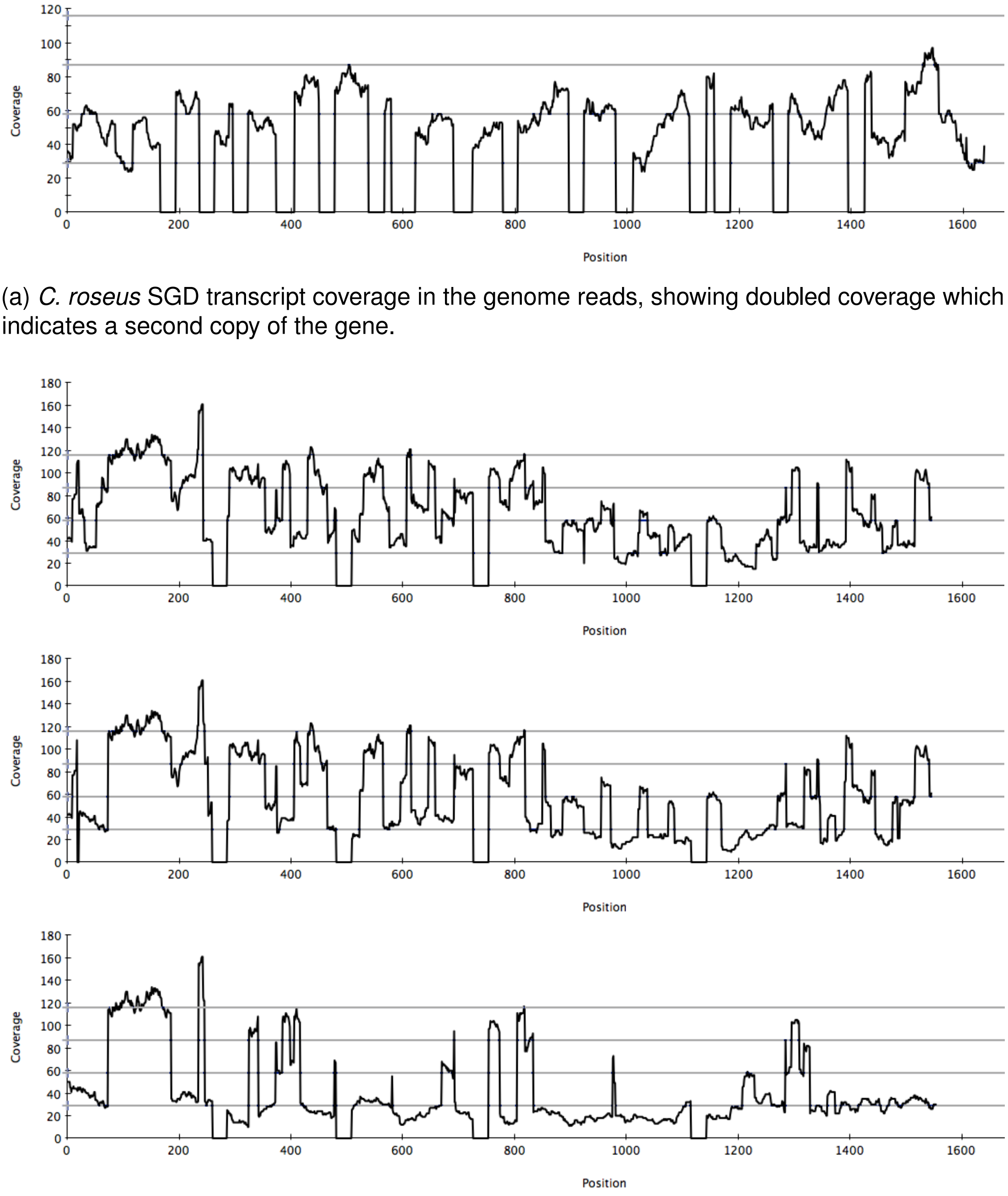
The genome was sequenced and assembled with ABySS [37] with default parameters and k=71bp. This was my typical default for a first-pass assembly because ABySS provided good logging of data characteristics and conservative heuristics with a modest computational cost, and using k=71bp had a good balance of specificity vs. coverage loss to k-mers affected by errors. When locating the genes from the pathways of interest in the assembly, only four transcripts could not be identified easily in the assembly: those of the *strictosidine β-glucosidase* (SGD) gene and the three *secologanin synthase* (SLS) genes. Coverage analysis of the known transcripts with KAT sect is shown in Figure 3.2. It showed the SGD gene had a second copy, leading to a misassembly of both in the result. The SLS genes, in turn, had four copies rather than the three known at the time, which again led to assembly collapse, but explained previously inconsistent results on gene expression. It was decided that the reconstruction of the pathway and genic space was met, and analyses of the contigs from the pathway also recovered promoters and clusters.

In this case, the initial objective was met with the exception of these multi-copy genes, but the reason for the assembly limitations was found and it could be taken into account in subsequent analysis. To the extent needed by the project, the assembly was successful.

3.4 Assembling genes, and long-range structure

The BBSRC emergency response to the Ash Dieback disease in 2012 funded sequencing of ash tree genomes [102] and the *H. pseudoalbidus* pathogen [6]. A local isolate of the pathogen was selected to be used as the reference genome. The main goal of the project was to recover the genic content of the pathogen for expression analyses and variant calling, but longer range structure was also a goal to facilitate the analysis of structural variants and gene placement. We proposed to sequence the isolate using paired end and Nextera Long Mate Pair (LMP) libraries.

The Nextera Long Mate Pair (LMP) was a step forward to produce long-range jumping libraries. I tested and participated in the publication of Nextclip [5], a tool to



(a) *C. roseus* SGD transcript coverage in the genome reads, showing doubled coverage which indicates a second copy of the gene.

(b) SLS transcripts coverage consistent with at least an extra fourth copy of the gene.

Fig. 3.2 *C. roseus* SGD and SLS transcript coverage by 31-mers in genome raw reads for previously available transcript sequences. Horizontal grey lines indicate expected median coverage in the reads for copy numbers 1, 2, 3, and 4, with coverage dropping to zero at splice junctions. Coverage in genome reads indicates extra gene copies for both genes.

filter the read pairs, classifying them into different categories to only use the correct pairs, correctly clipped, for genome scaffolding. In this publication I showed even for the relatively simple genome of *Arabidopsis thaliana*, a dramatic contiguity improvement could be achieved by only using the Nextclip-processed reads.

I started the project with two PE libraries. Of these, a shorter overlapping library was typical at the time, but I also requested a library with slightly larger 450bp fragment size. This larger fragment size library failed QC when a KAT spectra comparison between the two libraries detected substantial biases in it, so an assembly with only the overlapping library was produced using ABySS after joining the pairs with FLASH [41]. This assembly captured all of the k-mer spectrum content and mapped the RNA-seq samples well, so it passed the validation for genic content and provided a resource for rapid response to the disease.

Later, when the LMP library was produced, we used FLASH and Nextclip to process its reads. Since lower quality reads tend to have more errors along all the read, if the Nextera adaptor was found with few errors, the reads were of generally good quality, should contain no more chimeric joins, and we could use them to increase the k-mer coverage of the initial assembly. We decided to error-correct both read sets to be able to use larger K values in the DBG construction and read mapping, but executed the error correction independently to minimise cross-library bias. This final recipe greatly improved contiguity and correct read mapping without any negative effects on content recovery. Later analyses confirmed the genome was resolved into chromosome-scale scaffolds.

3.5 Assembling the wheat whole gene-ome

During the last phase of the wheat chromosome-by-chromosome Illumina sequencing project [103], I used the ISBP and EST coverage analyses described in Chapter 2 to review the biases and help chose the assembly methods and parameters for those datasets. I anticipated that future improvements on Illumina sequencing could enable WGS to surpass the quality of those assemblies. An early Roche 454 WGS assembly [104] and a project generating and anchoring 9Gbp of unique sequence along the wheat genome [105] using WGS Illumina data from various cultivars also showed the promise of WGS.

In 2014, a large strategic BBSRC project originally started in 2012 to provide sequences for 4 flow-sorted chromosomes switched strategy to explore Illumina WGS. After advising on data generation and QC, I joined to lead the assembly of the genome and decided to take advantage of then-recent technological advances. Illumina PCR-free 2x250bp sequencing was producing high-accuracy reads, enabling the use of larger K to untangle more complex genome graphs. DISCOVAR denovo [43], developed at Broad from an earlier variant calling pipeline, was producing good human-sized assemblies by virtue of avoiding short-repeat collapsing. Lastly, our own work on LMP

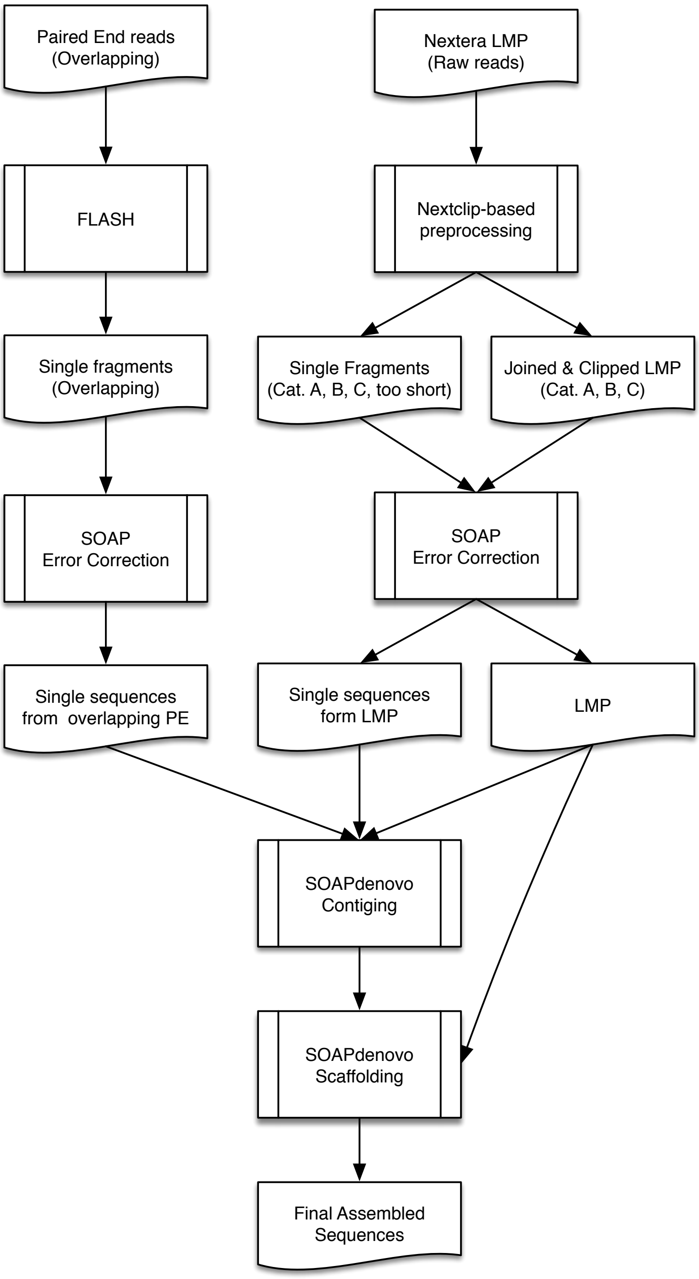
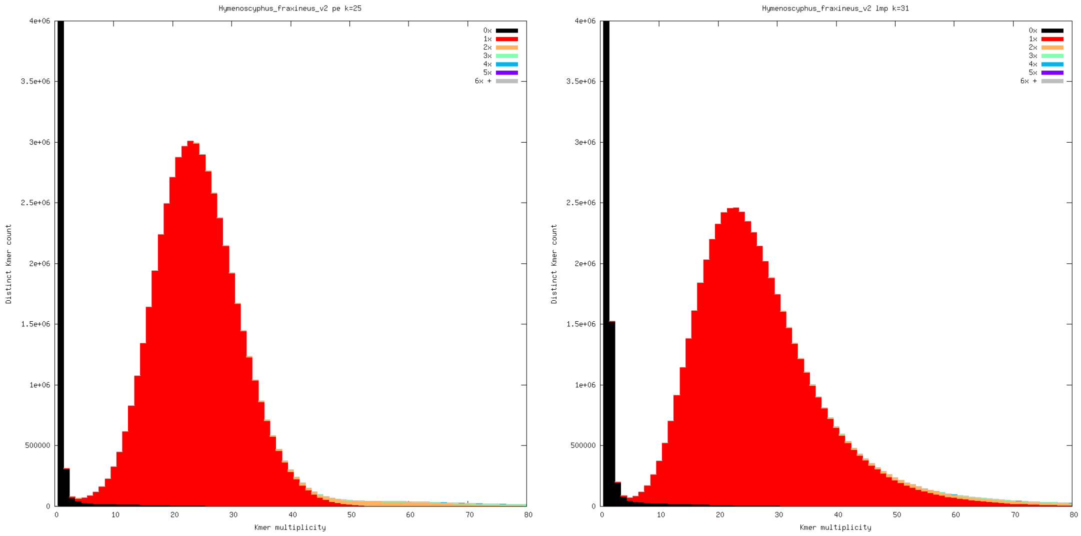

Paired End reads (Overlapping)
Nextera LMP (Raw reads)
FLASH
Nextclip-based preprocessing
Single fragments (Overlapping)
Single Fragments (Cat. A, B, C, too short)
Joined & Clipped LMP (Cat. A, B, C)
SOAP Error Correction
SOAP Error Correction
Single sequences from overlapping PE
Single sequences form LMP
LMP
SOAPdenovo Contiging
SOAPdenovo Scaffolding
Final Assembled Sequences

Fig. 3.3 Final assembly workflow for *H. pseudoalbidus*; using independent error correction on input libraries, PE and LMP data for graph generation, and only LMP for scaffolding. (Reproduced from McMullan, Rafiqi, Kaithakottil et al. [6])



(a) PE after error correction compared to the final assembly.

(b) LMP after FLASH, nextclip and error correction compared to the final assembly.

Fig. 3.4 KAT spectra-cn plots for the *H. pseudoalbidus* processed reads vs. the final assemblies',fig.subcap=c('a. PE after error correction.

library sequencing [7] and preparation [5] was producing precise, long-size libraries. The combination of these three methods should deliver robust assemblies for wheat, with good recovery of the genic space.

Both the goals and the validation of this project were straightforward, helping to guide an extremely challenging assembly. The main aim was to recover essentially all the genic content, while avoiding the creation of cross-subgenome chimeras through over-scaffolding. Besides our usual internal validation of k-mer spectra and read mapping; we were armed with the gold-standard hierarchical shotgun assembly and annotation of Chromosome 3B [106], which we used to assess gene reconstruction and large-scale structure, and the flow-sorted reads from the chromosome-by-chromosome Illumina project [103], which we used to assess cross-subgenome chimeras.

Running DISCOVAR on a complex plant dataset had computational performance challenges, given the sheer size of the dataset, and algorithmic challenges, given the complex repeat structure. Also, DISCOVAR typically used 450bp fragment sizes to span through ALU repeats on mammalian genomes, but when using larger fragment sizes, the "patching" and "repeat resolution" steps needed to deal with more complex graph topologies. After a first partial assembly showed a fully-covered k-mer spectrum and reasonable contiguity, we decided to adapt DISCOVAR to run on larger and more complex genomes, with larger fragment size data. Initially we received help and comments from the Broad team, but soon our efforts started to diverge as we optimised for much larger datasets, and started finding the complexity of the DISCOVAR code a liability. Over the following months we eliminated code for most of the unused heuristics, divided the assembly on its algorithmic steps, and added graph outputs to each part,

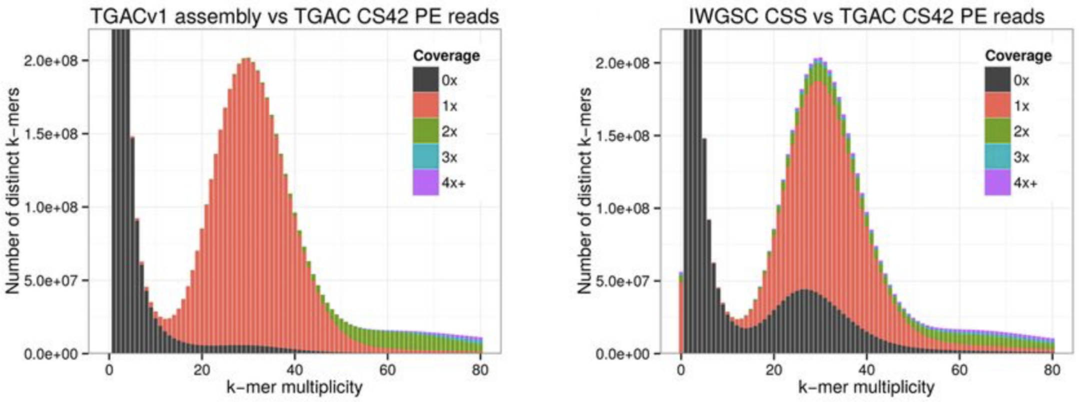


Fig. 3.5 KAT spectra-cn showing the improvement of the w2rap whole genome shotgun assemblies on the left, compared to the flow-sorted assembly from the IWGSC Chromosome Survey Sequence assemblies on the right. The increase in k-mer inclusion, along with better contiguity, resulted in an essentially complete reconstruction of the genic space. (Reproduced from Clavijo, Venturini, Schudoma et al. [8])

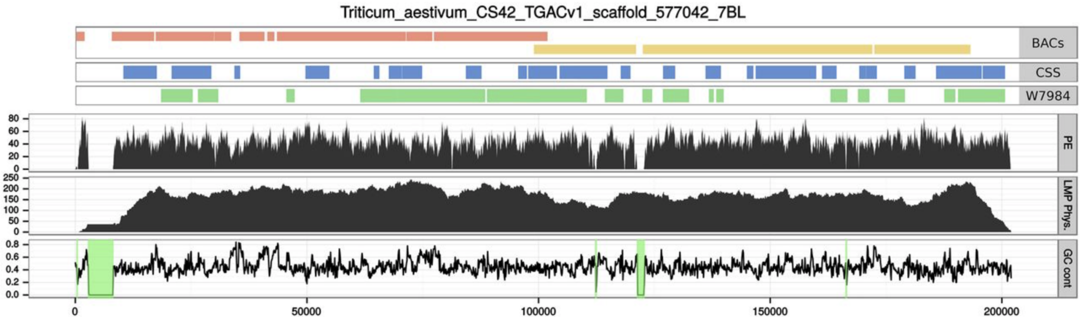


Fig. 3.6 Illustrative example of a TGACv1 scaffold, showing improvement versus all previous assembly attempts on wheat, using the TGACv1 scaffold as reference, with coorinates in the x axis. The top three tracks show fragmentation and incompleteness on the partial sequences recovered by BAC shotgun assembly, flow-sorted chrosomose assembly (CSS), and artificial hybrid WGS assembly (W7984) respectively. The bottom three tracks show support for the scaffold from PE and LMP reads, and GC content, including green boxes marking N srtetches in the scaffold.(Reproduced from Clavijo, Venturini, Schudoma et al. [8])

all in an effort to find the origin of the problem. Finally, we isolated the problems to 3 main causes and fixed them with different strategies.

First, DISCOVAR runs with all reads in memory, indexed by 32 bits read IDs, which we replaced by 64 bits to accomodate bigger readsets. Second, memory usage degrades over time, with fragmentation and leaks due to misusage of Mempool allocator based vectors. As a workaround we divided the execution into steps, which also served as checkpoints, and fixed the more evident fragmentation and leaks. And third, the graph based repeat resolution using read pairs was failing in some complex cases, creating graphs that had different solutions on the forward and reverse complement representations. To fix this last problem, uncovered by the more complex graph and larger fragment sizes, we rewrote part of the repeat resolution heuristics.

Once these fixes were in place, and after performance optimisation, we had a contigger that assembled the wheat genic space correctly with good completeness.

Our DISCOVAR denovo codebase had diverged so much from the Broad's that we agreed with them to keep it forked, and the w2rap-contigger was born.

We had much less trouble with the scaffolding, pretty much our first run of the SOAP-denovo2 [40] scaffolding module worked, and we settled for conservative parameters to avoid overscaffolding. Later on, reviewing the results, we found that the SOAPdenovo2 modules had 2 problems: an old bug on the codebase had an off-by-one condition that meant 1bp every 1024 was not being added to the consensus sequences, and some parameters were being ignored by the scaffolder module, even when they were being set on the command-line. We fixed these bugs on a version of SOAP that we shipped with our w2rap pipeline, and also included code to re-map the N stretches from the DISCOVAR pe-scaffolded final sequences into the final soap-produced scaffolds.

3.6 Conclusion

Non-model organism genome assembly can take a wide variety of forms, but the principles of defining objectives and metrics to iterate towards a solution providing a biologically relevant result are universally applicable. To this day, in 2019, our wheat TGACv1 assembly, finished in 2015, remains the most accurate assembly without manual curation and correction of the wheat genic space. By 2017 we had assembled and released another 5 wheat sequences for UK elite cultivars, which were immediately put to use by other researchers and breeders. This also showed the added value of our no-manual-curation approach, because these genomes were highly comparable.

Chapter 4

Genome graphs, hybrid datasets, and haplotype resolution

This chapter presents two different approaches to deal with haplotype resolution during genome assembly. It starts with my assembly of single haplotype mosaic assemblies for 16 *Heliconius* butterflies, which were presented as part of a publication [9] where my collaborators analysed their evolutionary history and showed the importance of genomic architecture and introgression in their radiation. It then presents the Sequence Distance Graph framework, a development from my group which I designed, prototyped and partially programmed, using a graph-based representation of an assembly as a basis to integrate hybrid datasets for scaffolding and analysis. This was presented in a publication [10] where members of my group contributed programming and testing, and is being currently used to analyse complex genomes.

4.1 Background

Sequencing techniques, assembly tools, and genome representations are becoming mature enough to deal with haplotypes, with haplotype-phased assemblies driving the development of new methods [75]. Haplotypes are how the genomes exist in reality, from where expression happens, and the units upon which evolution acts. Having access to study haplotypes rather than genomes is an improvement in definition that will revolutionise biological knowledge.

The typical case of haplotypic variation is that of the two haplotypes present in heterozygous diploid genomes. When assembling a genome, small variations between the two haplotypes generate alternative paths or bubbles (in a DBG assembly) or groups of lower identity overlaps (in an overlap assembly). The most common approach to deal with these has been to create a "mosaic" assembly, where one of the two haplotypes is choosen at each location via an arbitrary property, usually read coverage [40]. This "bubble popping" or "variant collapsing", however, can lead to collapse of other regions of high similarity and, even when done perfectly, disregards natural variation that may be fundamental for the understanding of the organism [81].

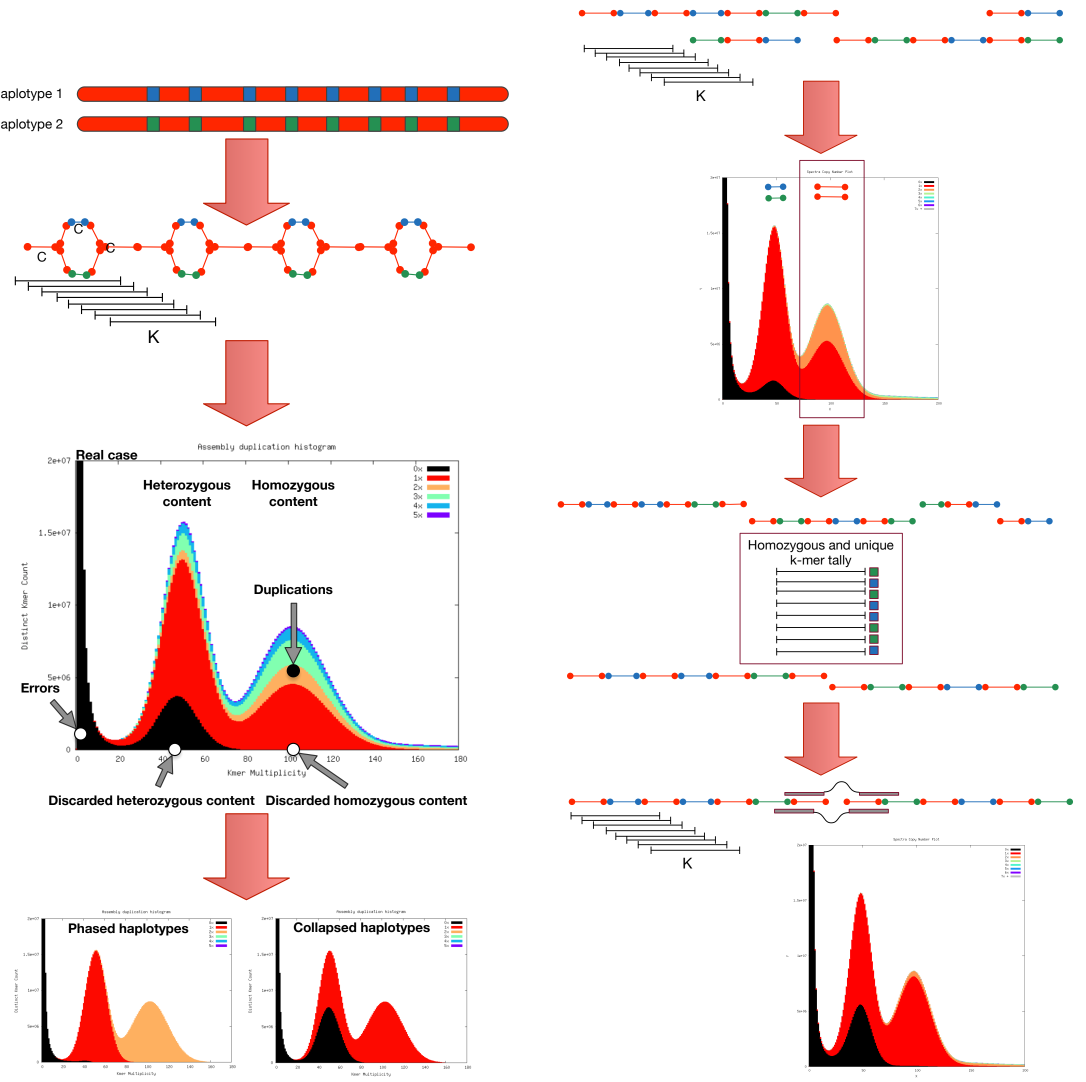
Three main approaches are used to deal with haplotypes during genome assembly in a non-mosaic way. The phasing approach, as in the Supernova assembler for linked reads [107], constructs the assembly graph with minimal cleaning, and then uses read-tag information or allelic variation to decide which sides of consecutive "bubbles" belong to the same haplotype, which are then "separated" by duplicating the shared section in the graph. A more aggressive collapse-and expand phasing approach, as in the FALCON-Unzip [56] algorithm for long reads, tries to find a mosaic-like solution first by collapsing regions of high similarity to linearise the graph, before recovering the alternative haplotype and executing a phasing algorithm (the Unzip stage). Finally, the most accurate approach and current gold standard is to sequence a trio of a sexually reproducing diploid, as in the TrioCanu method [82]. The reads of the child are then classified by similarity to each parent, effectively assembling the haplotypes independently.

Each of these approaches relies heavily on the properties of a diploid heterozygous karyotype, and can be difficult or impossible to apply to more complex karyotypes. Even for diploid assemblies it is often necessary to re-evaluate the collapsing or expanding choices, and they can confound subsequent steps in assembly pipelines. In extreme cases, applying heuristics for diploid assembly to organisms with higher ploidy can produce results that recover two mosaic haplotypes and largely misrepresent the real karyotype [108].

With the need to analyse haplotypic variation, even the format of reference genomes has been found to be problematic, since linear references can only represent a single haplotype out of the whole population, and augmented linear references (like the current human genome), only partially mitigate a minor part of this problem, while complicating alignment and mapping [109]. The graph representation, long used as the assembler's internal representation, is the best alternative to achieve more representative references that can be used efficiently. Tools to combine multiple linear references into reference graphs are maturing and gaining traction, but also in-the-graph analyses that work on graph output from genome assemblers, as well as tools that use the assembly graphs to construct graph references, are now being explored and finding their own niche of application [110]. It is now clear that some analyses will benefit from information integration on the assembly graphs, and haplotype reconstruction needs to be considered as an end-to-end process from sequencing to variation and functional analysis.

4.2 Haplotype collapsing

Heliconius butterflies are a good model to study speciation, since they are relatively easy to breed in laboratory conditions and they have large diversity and mimicry. In 2015 I got



(a) Effect of heterozygosity on assembly graphs and its reflection on k-mer spectra of the final assemblies.

(b) A haplotype-collapsing approach to reconstruct a mosaic representative sequence for a heterozygous region. CAVEAT: crushes structural variation.

Fig. 4.1 Heterozygosity and haploype collapsing in genome assembly.

involved in a project to generate 20 genome assemblies to explore their phylogeny and speciation. The sequencing data, following the DISCOVAR recipe [43] was producing incomplete and highly fragmented assemblies. Upon investigation via KAT spectra-cn I realised some samples needed resequencing, and others were just posing problems for the bubble popping of DISCOVAR denovo. Gonzalo Garcia and I coordinated re-sequencing of the poor samples and re-ran all assemblies with the w2rap-contigger. Because the main objective was to compare the general organisation of the genomes and be able to create phylogenetic trees, I decided a collapsed mosaic assembly would provide my collaborators with enough resolution to answer their biological questions.

The effect of heterozygosity on assembly graphs from short reads, and on the output k-mer spectra of the genome assemblies can be seen in Figure 4.1a: the small differences between haplotypes create bubbles in the graph which then in turn generate local duplications. Typical assemblers collapse part of these bubbles but not all, creating a solution that is neither fully collapsed nor expanded. To cope with this situation I implemented the process in Figure 4.1b. The output contigs from w2rap-contigger are first sorted by size, largest first. For each contig, the set of all its k-mers originating from the homozygous part of the spectrum is created, and then from largest to smallest, they are only added to the assembly if 80% of their homozygous k-mers are not yet included. This produces a set of single-haplotype contigs which are then re-scaffolded with the same paired end data that produced the w2rap contigs. The result is a mosaic assembly which is both more collapsed and more contiguous than its input.

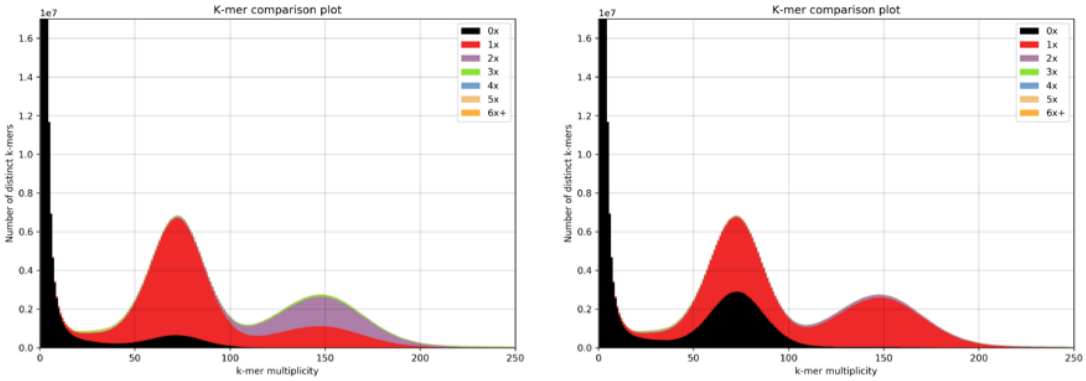


Fig. 4.2 KAT spectra-cn before and after haplotype-collapsing and re-scaffoling of *Agraulis vanillae*.(Reproduced from Edelman, Frandsen, Miyagi et al. [9])

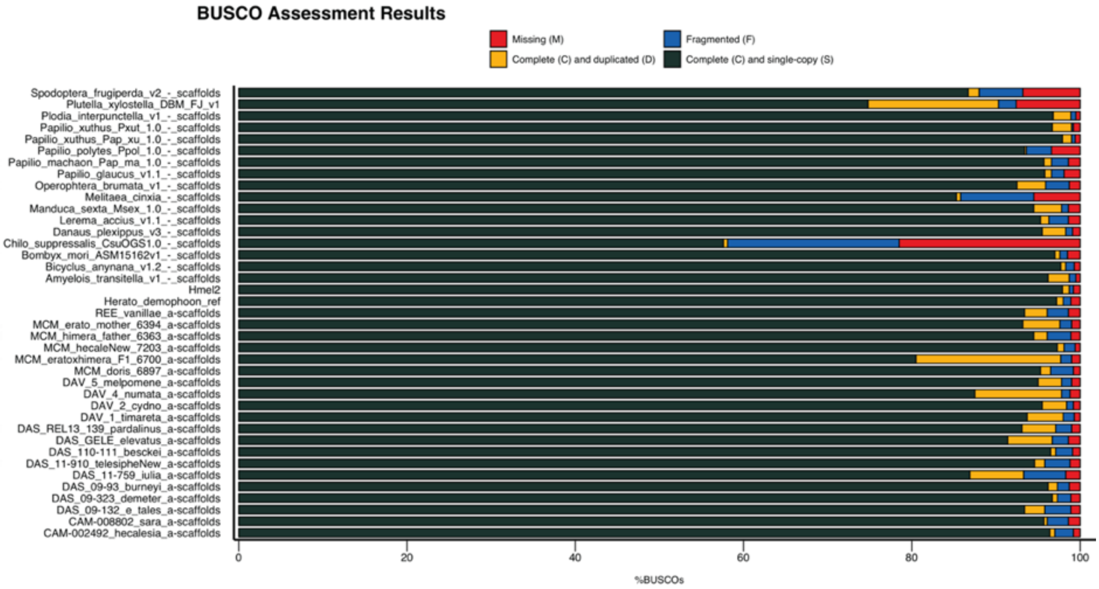


Fig. 4.3 BUSCO results for *Heliconius de novo* genomes, compared to other lepidopteran genomes obtained from lepbase.org. All assemblies labelled "a-scaffolds" are scaffolded with w2rap, haplotype-collapsed and re-scaffolded. Gene content is comparable between the w2rap assemblies and other lepidopteran assemblies, though the percentage of complete genes identified in w2rap genomes is slightly lower, and percent duplicated slightly higher, than the highest quality reference Lepidoptera genomes. (Reproduced from Edelman, Frandsen, Miyagi et al. [9])

The k-mer spectra before and after the collapsing and re-scaffolding process for a representative example can be seen in Figure 4.2. Further QC of these assemblies showed they were fit for the analysis to be done, with BUSCO [96, 97] scores also falling in line with lepbase [111] assemblies as shown in Figure 4.3. These assemblies were used as a basis to construct comparable genome representations between the 20 samples, but also recovered content lost in other lepbase assemblies, which were patched.

The analysis of the collapsed mosaic assemblies improved the understanding of the importance of introgression and selective processes in adaptive radiation, changing some of the assumptions not only about *Heliconius* but about evolution in general. While collapsing is a limited strategy, this project's objectives were well met by our approach.

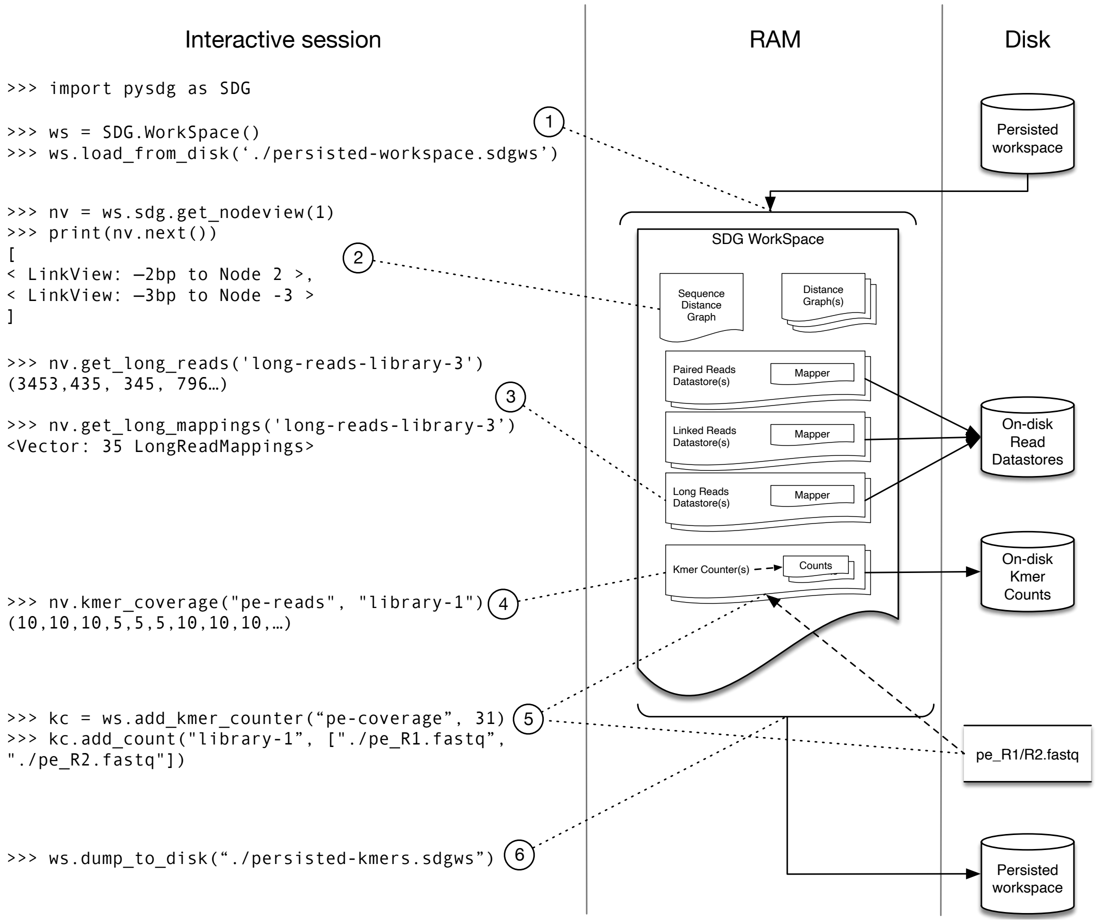
4.3 The Sequence Distance Graph framework

Ever since I started finding spectra problems in assemblies, and heuristic problems in assemblers, thinking about assembly problems and possible genome representations as graphs became a need. Assembly graph representations are powerful tools, but while it is easy to think conceptually about assembly graphs, most implementations are not easy to work with for exploratory analyses. They have been designed with a processing mindset (i.e. designed to facilitate some kind of analysis) rather than with an exploring mindset (i.e. designed to provide access to information about the problem), and are heavily optimised.

In order to develop our own tools for genome assembly, we started implementing graph modules, eventually creating a relatively unoptimised framework for sequence graph analyses. From the ground up, this framework has been designed to allow haplotype-specific exploratory analysis, as well as programming of tools or pipelines.

The Sequence Distance Graph (SDG) framework implements a SequenceDistanceGraph representation that defines sequences in nodes and their adjacency in links, and an associated Workspace containing raw data and mappings, as shown in Figure 4.4. This provides an integrated working environment to use multiple sources of information to navigate and analyse genome graphs. Datastores allow random access to short, linked, and long read sequences on disk. A mapper in each datastore contains methods to map the reads to the graph and access the mapping data. KmerCounters provide functions to compute k-mer coverage over the graph from sequencing data, enabling coverage analyses. Additional DistanceGraphs, typically representing longer-range information and different linkage levels, define alternative topologies over the SequenceDistanceGraph nodes. Finally, a NodeView abstraction provides a proxy to a node, with methods to navigate the graph and access its mapped data. This comprehensive framework can be used to explore genome graphs interactively or to create processing methods for assembly or downstream analysis.

Working with SDG typically involves two different stages: creating a WorkSpace with the data and mappings (Figure 4.5a), and analysing this WorkSpace (Figure 4.5b-4.5).



```
>>> import pysdg as SDG

>>> ws = SDG.WorkSpace()
>>> ws.load_from_disk(‘./persisted-workspace.sdgws’)

>>> nv = ws.sdg.get_nodeview(1)
>>> print(nv.next())
[
< LinkView: —2bp to Node 2 >,
< LinkView: —3bp to Node -3 >
]

>>> nv.get_long_reads('long-reads-library-3')
(3453,435, 345, 796…)

>>> nv.get_long_mappings('long-reads-library-3’)
<Vector: 35 LongReadMappings>

>>> nv.kmer_coverage("pe-reads", "library-1")
(10,10,10,5,5,5,10,10,10,…)

>>> kc = ws.add_kmer_counter(“pe-coverage”, 31)
>>> kc.add_count("library-1”, ["./pe_R1.fastq”,
"./pe_R2.fastq"])

>>> ws.dump_to_disk(“./persisted-kmers.sdgws”)
```

Fig. 4.4 The SDG WorkSpace holds the information for a project and contains the graphs, the mappers and k-mer counts. From Python, a previously saved WorkSpace is loaded from disk (1). The NodeView object is centred on a specific node and can be used to access node characteristics (ie. size and sequence), graph topology from the perspective of the node you are on (i.e. neighbours in both directions (2)) and can also retrieve information projected onto the selected node (ie. mappings (3) and k-mer coverage (4)). Operations such as adding a KmerCounter to the WorkSpace and adding a count (5) can be performed, and the WorkSpace can be saved back to disk (6). Once loaded, the bulk of the WorkSpace is held in memory for fast access with the raw read data from the DataStores remaining on disk accessible through random access. (Adapte from Yanes, Garcia Accinelli, Wright et al. [10])

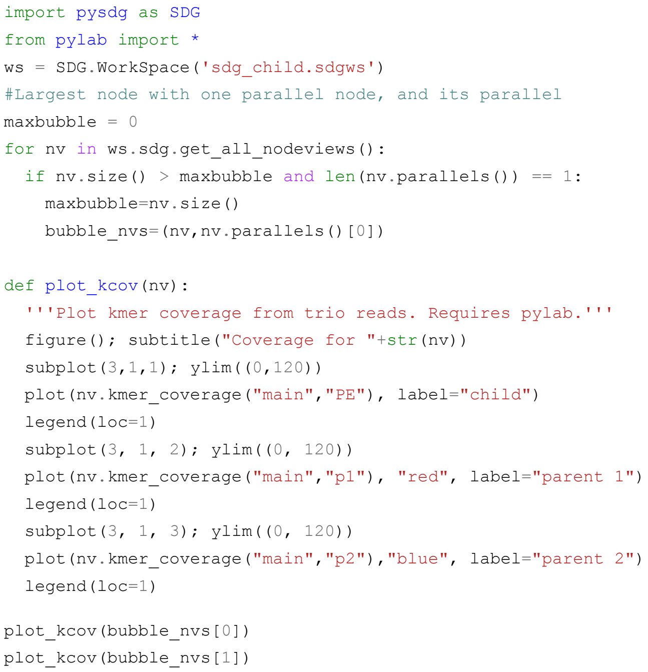
SDG includes command line tools to create DataStores, KmerCounts, and WorkSpaces, and map reads within a WorkSpace.

To explore the graph, the NodeView class, and its associated LinkViews, provide a single-entry point for node-centric analyses. A NodeView from either a DistanceGraph or SequenceDistanceGraph is a wrapper containing a pointer to the graph and a node id, and will provide access to its nodes' previous and next linked nodes, mapped reads, or k-mer coverage. A user with good understanding of the NodeView class should be able to access most information in the WorkSpace through it.

The projection of k-mer coverage and read mappings directly in the graph's nodes enables complex analysis like trio-sequencing node classification to be done very easily with a few lines of code, either for exploration as shown in Figure 4.5 or for automated assembly processes. The original intention behind SDG was to enable exploratory

```
sdg-datastore make -t paired -1 child/pe_R1.fastq.gz -2 child/pe_R1.fastq.gz -o child_pe
sdg-dbg -p child_pe.prseq -o sdg_child
sdg-kmercounter add -c main.sdgkc -n p1 -f p1/pe_R1.fastq.gz -f p1/pe_R2.fastq.gz -o main
sdg-kmercounter add -c main.sdgkc -n p2 -f p2/pe_R1.fastq.gz -f p2/pe_R2.fastq.gz -o main
```

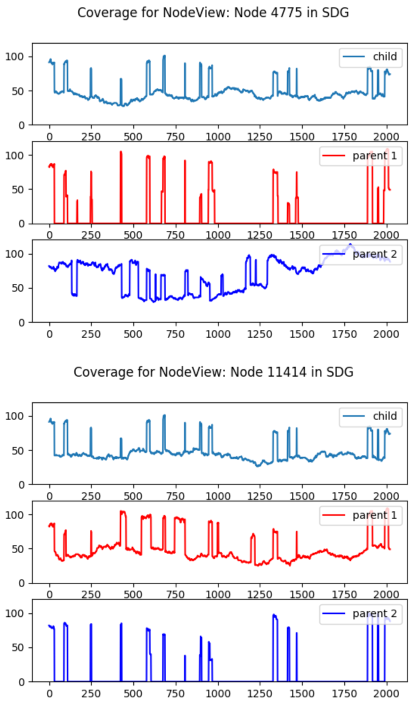
(a) Creation of the WorkSpace with an assembly and read k-mer coverage from the trio.



```
import pysdg as SDG
from pylab import *
ws = SDG.WorkSpace('sdg_child.sdgws')
#Largest node with one parallel node, and its parallel
maxbubble = 0
for nv in ws.sdg.get_all_nodeviews():
  if nv.size() > maxbubble and len(nv.parallels()) == 1:
    maxbubble=nv.size()
    bubble_nvs=(nv,nv.parallels()[0])

def plot_kcov(nv):
  '''Plot kmer coverage from trio reads. Requires pylab.'''
  figure(); subtitle("Coverage for "+str(nv))
  subplot(3,1,1); ylim((0,120))
  plot(nv.kmer_coverage("main","PE"), label="child")
  legend(loc=1)
  subplot(3, 1, 2); ylim((0, 120))
  plot(nv.kmer_coverage("main","p1"), "red", label="parent 1")
  legend(loc=1)
  subplot(3, 1, 3); ylim((0, 120))
  plot(nv.kmer_coverage("main","p2"),"blue", label="parent 2")
  legend(loc=1)

plot_kcov(bubble_nvs[0])
plot_kcov(bubble_nvs[1])
```



(b) Python code to load the workspace, find the longest bubble and plot its coverage.

(c) Coverage plots, showing opposite parent coverage drop to zero on heterozygous k-mers.

Fig. 4.5 Using SDG to analyse coverage along both sides of a heterozygous bubble on the child assembly of a trio dataset (Adapted from Yanes, Garcia Accinelli, Wright et al. [10])

analyses, but we have found that small optimisations to its core enable it to be used for production-level assembly processes while maintaining its simplicity.

4.4 Conclusion

While haplotype-phased assemblies are the goal of many new tools, complex genomes will need new methodological development. Preliminary versions of SDG have already accelerated our internal work with assembly graphs, and my group is developing our new generation of assembly tools based on it. The integration of all data types in a single easy to use Node-centric framework is already helping us conduct more detailed haplotype-specific analyses.

Chapter 5

Discussion

Much of modern biology research depends on genome sequencing and assembly. While the human genome project certainly revolutionised medical research, many efforts are still ongoing to move from a single biased reference to a more complete view of human genomics. However most of the genomic diversity of Earth remains largely unexplored, and holds the keys to understand evolution and biology at both more fundamental level and a larger scale. The Earth Biogenome Project aims to sequence around 1.5 million known eukaryotic species over a 10-year period to revolutionise our understanding of biology and evolution; promote conservation, protection, and restoration of biodiversity; and create new benefits for society and human welfare [112].

This unprecedented level of exploration redefines the concept of *de novo* sequencing. Whole taxa are pretty much unknown, and all kinds of unexpected genomic characteristics are without a doubt awaiting in the tree of life. The UK's Darwin Tree of Life project [113] is already exploring protists, of which many are single-cell organisms that can't be cultured or easily isolated, bringing the complexities of single-cell genomic exploration. While a collection of genomic sequences representing all life on earth is undoubtedly a valuable resource, there is the inescapable reality that life is not static nor discrete, and single individuals are going to be representing populations or whole taxa that may have little in common with their selected representatives.

All of this reinforces the importance of careful *de novo* analyses. Not only are the individual genome assemblies the building blocks of these extraordinary resources, but the view of life we achieve depends on how accurately we represent their content. The simple goals of completeness, correctness and contiguity are very much at interplay here, and as we venture outside the species that have been sequenced over and over again, this trade-off becomes, again, a very real choice to make. Some of the methods I have developed as features in KAT are proving effective to evaluate completeness of the assemblies and haplotype representation, and have been recently adopted by consortia like the Vertebrate Genomes Project and the Darwin Tree of Life.

Genome assemblies are invaluable resources to study non-model organisms. When they are conducted in a careful way, producing robust results supporting biological analyses of interest, they can provide insights or unlock information from population samples. The assembly of *C. roseus* and the analyses performed on it have shown the value of a genic-space assembly for pathway analysis in an scenario where very limited complementary genomic information is available [4]. My first assembly of *H. pseudoalbidus*, produced in immediate response to the ash dieback invasion, enabled analyses of pathogenic genes and molecular mechanisms of the disease; the second, near-chromosome scale, assembly made phylogenetic analyses easier [6]. In all these cases the use of objective-led assembly, combined with a drive towards completeness based on k-mer spectra, and a focus on avoiding errors rather than maximising contiguity, led to assemblies that accelerated research.

My short read WGS wheat assembly approach was focused on clean data generation, completeness, and stringent parameters for automated assembly [8]. It used relatively low coverage of PCR-free paired end data and long mate paired libraries to produce good reconstructions of the genic space of hexaploid wheat from short reads. This genome assembly, alongside the annotation produced by my co-authors showing better gene models, effectively ended hierarchical shotgun efforts for wheat, marking the final adoption of straightforward WGS on the most complex genome attempted by both methods. The gene models and annotation of this essentially complete genic assembly greatly impacted breeding and agricultural applications. Later, the hybrid assembly by Zimin et al. I helped validate and improve showed further structural resolution from long read WGS [3]. The focus has now quickly shifted towards generating pan-genomes using relatively cheap WGS approaches, which are enabling the exploration of the large pool of diversity in wheat [114]. Perhaps the main focus on my wheat assembly approach was to produce assemblies that were ready to be used straight out of the pipeline, and to that effect I rejected any improvements that implicated manual curation or correction. This made it very easy for the reference genome, carefully put together combining the best elements of WGS and hierarchical shotgun approaches, to surpass the metrics of my simple WGS pipeline; but at the same time, a focus on not requiring human intervention in the assembly pipeline will eventually enable faster discovery and easier result integration between assemblies. The impacts of this work are still ongoing, and it is possible we will have a version of the same discussion over the creation of pan-genomes: at the end of the day, I still believe in the importance of a model for simple and statistically powerful data analysis and comparison that can be automated and replicated without the extensive use of human decision making.

Haplotype separation on eukaryotes has come a long way in the the last decade. With the increase in accuracy of NGS reads and the sampling depth they provide, divergent sections of the haplotypes were being generated, but the typical approach was to discard every alternative but one, thus assembling a "representative haplotype mosaic". This was being done at the level of the DBG topology by "popping bubbles" or in the overlap graph by adjusting the overlap constraints and linearising the resulting graph. More sophisticated approaches such as the k-mer spectra inclusion check I designed for the *Heliconius* genomes succeed in maintaining appropriate representation of the genomes, and have been extremely useful to support genome-scale analyses [9]. But these haplotype-aware approaches, which are based on a strong prior about haplotype homology and sequence repetition, are not perfect, and will introduce unwanted biases in cases where the karyotypes are unknown. I have co-supervised a masters student who used the k-mer spectra analyses implemented on the Sequence Distance Graph (SDG) [10] to uncover a third haplotype signature, most likely arising from triploidy, on a diatom genome previously assembled as a diploid [108]. These type of enforcement of inappropriate priors lead not only to assembly errors, but to completely inappropriate biological interpretations at all levels from individual genome composition to population evolution. With new graph-based approaches gaining traction in the genomics community, I hope SDG and its design as not only a framework for method development but a workspace for users to analyse a complete dataset will become not an isolated effort but a typical example of a new generation of tools for genomic analyses.

In summary, the publications presented here advanced genomic research and provided scientific impact, moving within a highly changing technological field, and sometimes contributing to its momentum. In the future of genome sequencing and assembly, haplotype specificity will become the norm rather than the exception, pan-genomics a reality rather than an aspiration, and truly *de novo* analyses will provide amazing new findings as we venture further away through the branches of the tree of life. I hope genomic tools will become more robust; and analyses simpler yet more powerful, significant, and accessible; bringing further understanding of life upon earth.

References

[1] M. Helguera, M. Rivarola, B. Clavijo, et al. New insights into the wheat chromosome 4D structure and virtual gene order, revealed by survey pyrosequencing. *Plant Science*, 233:200–212, 2015. ISSN 0168-9452. doi: 10.1016/j.plantsci.2014.12.004.

[2] D. Mapleson, G. Garcia Accinelli, G. Kettleborough, et al. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*, 33(4):574–576, 2017. ISSN 1367-4803. doi:10.1093/bioinformatics/btw663.

[3] A. V. Zimin, D. Puiu, R. Hall, et al. The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum. *GigaScience*, 6(11):1–7, 2017. ISSN 2047217X. doi:10.1093/gigascience/gix097.

[4] F. Kellner, J. Kim, B. J. Clavijo, et al. Genome-guided investigation of plant natural product biosynthesis. *The Plant Journal*, 82(4):680–692, 2015. ISSN 1365-313X. doi:10.1111/tpj.12827.

[5] R. M. Leggett, B. J. Clavijo, L. Clissold, et al. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics*, 30(4):566–568, 2014. ISSN 1367-4803. doi:10.1093/bioinformatics/btt702.

[6] M. McMullan, M. Rafiqi, G. Kaithakottil, et al. The ash dieback invasion of Europe was founded by two genetically divergent individuals. *Nature Ecology & Evolution*, 2(6):1000, 2018. ISSN 2397-334X. doi:10.1038/s41559-018-0548-9.

[7] D. Heavens, G. G. Accinelli, B. Clavijo, and M. D. 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*, 59(1):42–45, 2015. ISSN 0736-6205. doi:10.2144/000114310.

[8] B. J. Clavijo, L. Venturini, C. Schudoma, 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(5):885–896, 2017. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.217117.116.

[9] N. B. Edelman, P. B. Frandsen, M. Miyagi, et al. Genomic architecture and introgression shape a butterfly radiation. *Science*, 366(6465):594–599, 2019. ISSN 0036-8075, 1095-9203. doi:10.1126/science.aaw2090.

[10] L. Yanes, G. Garcia Accinelli, J. Wright, et al. A Sequence Distance Graph framework for genome assembly and analysis. *F1000Research*, 8:1490, 2019. ISSN 2046-1402. doi:10.12688/f1000research.20233.1.

[11] R. Staden. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, 6(7):2601–2610, 1979. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/6.7.2601.

[12] F. Sanger, G. M. Air, B. G. Barrell, et al. Nucleotide sequence of bacteriophage. 265:9, 1977.

[13] J. Ghurye and M. Pop. Modern technologies and algorithms for scaffolding assembled genomes. *PLOS Computational Biology*, 15(6):e1006994, 2019. ISSN 1553-7358. doi:10.1371/journal.pcbi.1006994.

[14] N. Nagarajan and M. Pop. Parametric Complexity of Sequence Assembly: Theory and Applications to Next Generation Sequencing. *Journal of Computational Biology*, 16(7):897–908, 2009. doi:10.1089/cmb.2009.0005.

[15] H. Peltola, H. Söderlund, and E. Ukkonen. SEQAID: a DNA sequence assembling program based on a mathematical model. *Nucleic Acids Research*, 12(1 Pt 1):307–321, 1984. ISSN 0305-1048.

[16] E. W. Myers. Toward Simplifying and Accurately Formulating Fragment Assembly. *Journal of Computational Biology*, 2(2):275–290, 1995. ISSN 1066-5277, 1557-8666. doi:10.1089/cmb.1995.2.275.

[17] X. Huang. A contig assembly program based on sensitive detection of fragment overlaps. *Genomics*, 14(1):18–25, 1992. ISSN 0888-7543. doi:10.1016/S0888-7543(05)80277-0.

[18] P. Green. Documentation for PHRAP. *Genome Center, University of Washington, Seattle*, 1996.

[19] A. Goffeau, B. G. Barrell, H. Bussey, et al. Life with 6000 Genes. *Science*, 274(5287):546–567, 1996. ISSN 0036-8075, 1095-9203. doi:10.1126/science.274.5287.546.

[20] T. C. e. S. Consortium*. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science*, 282(5396):2012–2018, 1998. ISSN 0036-8075, 1095-9203. doi:10.1126/science.282.5396.2012.

[21] Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814):796–815, 2000. ISSN 1476-4687. doi:10.1038/35048692.

[22] R. Fleischmann, M. Adams, O. White, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223):496–512, 1995. ISSN 0036-8075, 1095-9203. doi:10.1126/science.7542800.

[23] E. W. Myers, G. G. Sutton, A. L. Delcher, et al. A Whole-Genome Assembly of Drosophila. *Science*, 287(5461):2196–2204, 2000. ISSN 0036-8075, 1095-9203. doi:10.1126/science.287.5461.2196.

[24] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860, 2001. ISSN 1476-4687. doi:10.1038/35057062.

[25] J. C. Venter, M. D. Adams, E. W. Myers, et al. The Sequence of the Human Genome. *Science*, 291(5507):1304–1351, 2001. ISSN 0036-8075, 1095-9203. doi:10.1126/science.1058040.

[26] Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004. ISSN 1476-4687. doi:10.1038/nature03001.

[27] D. B. Jaffe, J. Butler, S. Gnerre, et al. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Research*, 13(1):91–96, 2003. ISSN 1088-9051. doi:10.1101/gr.828403.

[28] Mouse Genome Sequencing Consortium, R. H. Waterston, K. Lindblad-Toh, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002. ISSN 0028-0836. doi:10.1038/nature01262.

[29] P. A. Pevzner, H. Tang, and M. S. Waterman. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753, 2001. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.171285098.

[30] M. C. Schatz, A. L. Delcher, and S. L. Salzberg. Assembly of large genomes using second-generation sequencing. *Genome Research*, 20(9):1165–1173, 2010. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.101360.109.

[31] M. Margulies, M. Egholm, W. E. Altman, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005. ISSN 1476-4687. doi:10.1038/nature03959.

[32] B. Chevreux, T. Pfisterer, B. Drescher, et al. Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. *Genome Research*, 14(6):1147–1159, 2004. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.1917404.

[33] A. Lakdawalla and H. Vansteenhouse. Illumina Genome Analyzer II System. In M. Janitz, editor, *Next Generation Genome Sequencing*, pages 13–28. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2008. ISBN 978-3-527-62513-0 978-3-527-32090-5. doi:10.1002/9783527625130.ch2.

[34] D. R. Zerbino and E. Birney. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829, 2008. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.074492.107.

[35] M. J. Chaisson and P. A. Pevzner. Short read fragment assembly of bacterial genomes. *Genome Research*, 18(2):324–330, 2008. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.7088808.

[36] J. Butler, I. MacCallum, M. Kleber, et al. ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Research*, 18(5):810–820, 2008. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.7337908.

[37] J. T. Simpson, K. Wong, S. D. Jackman, et al. ABySS: A parallel assembler for short read sequence data. *Genome Research*, 19(6):1117–1123, 2009. ISSN 10889051. doi:10.1101/gr.089532.108.

[38] S. Gnerre, I. MacCallum, D. Przybylski, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, 2010. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.1017351108.

[39] R. Li, H. Zhu, J. Ruan, et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20(2):265–272, 2010. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.097261.109.

[40] R. Luo, B. Liu, Y. Xie, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1):18, 2012. ISSN 2047-217X. doi:10.1186/2047-217X-1-18.

[41] T. Magoč and S. L. Salzberg. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963, 2011. ISSN 1367-4803. doi:10.1093/bioinformatics/btr507.

[42] A. Bankevich, S. Nurk, D. Antipov, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5):455–477, 2012. doi:10.1089/cmb.2012.0021.

[43] R. R. Love, N. I. Weisenfeld, D. B. Jaffe, et al. Evaluation of DISCOVAR de novo using a mosquito sample for cost-effective short-read genome assembly. *BMC Genomics*, 17(1):187, 2016. ISSN 1471-2164. doi:10.1186/s12864-016-2531-7.

[44] S. D. Jackman, B. P. Vandervalk, H. Mohamadi, et al. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Research*, 27(5):768–777, 2017. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.214346.116.

[45] E. W. Myers. The fragment assembly string graph. *Bioinformatics*, 21(Suppl 2):ii79–ii85, 2005. ISSN 1367-4803, 1460-2059. doi:10.1093/bioinformatics/bti1114.

[46] J. T. Simpson and R. Durbin. Efficient de novo assembly of large genomes using compressed data structures. *Genome Research*, 22(3):549–556, 2012. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.126953.111.

[47] H. Li. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics*, 28(14):1838–1844, 2012. ISSN 1367-4803. doi:10.1093/bioinformatics/bts280.

[48] S. M. D. Goldberg, J. Johnson, D. Busam, et al. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proceedings of the National Academy of Sciences*, 103(30):11240–11245, 2006. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.0604351103.

[49] J. R. Miller, A. L. Delcher, S. Koren, et al. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, 24(24):2818–2824, 2008. ISSN 1367-4803. doi:10.1093/bioinformatics/btn548.

[50] A. V. Zimin, G. Marçais, D. Puiu, et al. The MaSuRCA genome assembler. *Bioinformatics*, 29(21):2669–2677, 2013. ISSN 1367-4803. doi:10.1093/bioinformatics/btt476.

[51] F. J. Ribeiro, D. Przybylski, S. Yin, et al. Finished bacterial genomes from shotgun sequence data. *Genome Research*, 22(11):2270–2277, 2012. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.141515.112.

[52] S. Koren, M. C. Schatz, B. P. Walenz, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, 30(7):693–700, 2012. ISSN 1087-0156, 1546-1696. doi:10.1038/nbt.2280.

[53] S. Koren, G. P. Harhay, T. P. Smith, et al. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology*, 14(9):R101, 2013. ISSN 1474-760X. doi:10.1186/gb-2013-14-9-r101.

[54] C.-S. Chin, D. H. Alexander, P. Marks, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, 10(6):563–569, 2013. ISSN 1548-7091, 1548-7105. doi:10.1038/nmeth.2474.

[55] A. V. Zimin, D. Puiu, M.-C. Luo, 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 Research*, 27(5):787–792, 2017. ISSN 1088-9051. doi:10.1101/gr.213405.116.

[56] C.-S. Chin, P. Peluso, F. J. Sedlazeck, et al. Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing. *Nature methods*, 13(12):1050–1054, 2016. ISSN 1548-7091. doi:10.1038/nmeth.4035.

[57] A. M. Wenger, P. Peluso, W. J. Rowell, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10):1155–1162, 2019. ISSN 1546-1696. doi:10.1038/s41587-019-0217-9.

[58] H. Li. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, 32(14):2103–2110, 2016. ISSN 1367-4803. doi:10.1093/bioinformatics/btw152.

[59] S. Koren, B. P. Walenz, K. Berlin, et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5):722–736, 2017. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.215087.116.

[60] M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 2019. ISSN 1087-0156, 1546-1696. doi:10.1038/s41587-019-0072-8.

[61] J. Ruan and H. Li. Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, pages 1–4, 2019. ISSN 1548-7105. doi:10.1038/s41592-019-0669-3.

[62] B. J. Walker, T. Abeel, T. Shea, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE*, 9(11):e112963, 2014. ISSN 1932-6203. doi:10.1371/journal.pone.0112963.

[63] N. J. Loman, J. Quick, and J. T. Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, 12(8):733–735, 2015. ISSN 1548-7105. doi:10.1038/nmeth.3444.

[64] R. Vaser, I. Sovic, N. Nagarajan, and M. Sikic. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, page gr.214270.116, 2017. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.214270.116.

[65] M. Pop, D. S. Kosack, and S. L. Salzberg. Hierarchical Scaffolding With Bambus. *Genome Research*, 14(1):149–159, 2004. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.1536204.

[66] M. Boetzer, C. V. Henkel, H. J. Jansen, et al. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, 27(4):578–579, 2011. ISSN 1367-4803. doi:10.1093/bioinformatics/btq683.

[67] A. Dayarian, T. P. Michael, and A. M. Sengupta. SOPRA: Scaffolding algorithm for paired reads via statistical optimization. *BMC Bioinformatics*, 11(1):345, 2010. ISSN 1471-2105. doi:10.1186/1471-2105-11-345.

[68] L. Salmela, V. Mäkinen, N. Välimäki, et al. Fast scaffolding with small independent mixed integer programs. *Bioinformatics*, 27(23):3259–3265, 2011. ISSN 1367-4803. doi:10.1093/bioinformatics/btr562.

[69] Y. Mostovoy, M. Levy-Sakin, J. Lam, et al. A hybrid approach for de novo human genome sequence assembly and phasing. *Nature Methods*, 13(7):587–590, 2016. ISSN 1548-7105. doi:10.1038/nmeth.3865.

[70] L. Coombe, J. Zhang, B. P. Vandervalk, et al. ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers. *BMC Bioinformatics*, 19(1):234, 2018. ISSN 1471-2105. doi:10.1186/s12859-018-2243-x.

[71] R. L. Warren, C. Yang, B. P. Vandervalk, et al. LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience*, 4(1):35, 2015. ISSN 2047-217X. doi:10.1186/s13742-015-0076-3.

[72] Scaff10X https://github.com/wtsi-hpag/Scaff10X. .

[73] N. Nagarajan, T. D. Read, and M. Pop. Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics*, 24(10):1229–1235, 2008. ISSN 1367-4803. doi:10.1093/bioinformatics/btn102.

[74] J. N. Burton, A. Adey, R. P. Patwardhan, et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, 31(12):1119–1125, 2013. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.2727.

[75] J. Ghurye, M. Pop, S. Koren, et al. Scaffolding of long read assemblies using long range contact information. *BMC Genomics*, 18(1):527, 2017. ISSN 1471-2164. doi:10.1186/s12864-017-3879-z.

[76] O. Dudchenko, S. S. Batra, A. D. Omer, et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*, 356(6333):92–95, 2017. ISSN 0036-8075, 1095-9203. doi:10.1126/science.aal3327.

[77] J. Ghurye, A. Rhie, B. P. Walenz, et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLOS Computational Biology*, 15(8):e1007273, 2019. ISSN 1553-7358. doi:10.1371/journal.pcbi.1007273.

[78] X. Zhang, S. Zhang, Q. Zhao, et al. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature Plants*, 5(8):833–845, 2019. ISSN 2055-0278. doi:10.1038/s41477-019-0487-8.

[79] M. C. Schatz, J. Witkowski, and W. R. McCombie. Current challenges in de novo plant genome sequencing and assembly. *Genome Biology*, 13(4), 2012. ISSN 1474760X. doi:10.1186/gb-2012-13-4-243.

[80] M. Patterson, T. Marschall, N. Pisanti, et al. WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *Journal of Computational Biology*, 22(6):498–509, 2015. doi:10.1089/cmb.2014.0157.

[81] P. Marks, S. Garcia, A. M. Barrio, et al. Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Research*, 29(4):635–645, 2019. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.234443.118.

[82] S. Koren, A. Rhie, B. P. Walenz, et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology*, 36(12), 2018. ISSN 1087-0156. doi:10.1109/BHI.2014.6864426.

[83] S. L. Salzberg, A. M. Phillippy, A. Zimin, et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, 22(3):557–567, 2012. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.131383.111.

[84] D. Earl, K. Bradnam, J. S. John, et al. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, 21(12):2224–2241, 2011. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.126599.111.

[85] K. R. Bradnam, J. N. Fass, A. Alexandrov, et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2(1):10, 2013. ISSN 2047-217X. doi:10.1186/2047-217X-2-10.

[86] S. C. Clark, R. Egan, P. I. Frazier, and Z. Wang. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics*, 29(4):435–443, 2013. ISSN 1367-4803. doi:10.1093/bioinformatics/bts723.

[87] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, 2013. ISSN 1367-4803. doi:10.1093/bioinformatics/btt086.

[88] A. Mikheenko, A. Prjibelski, V. Saveliev, et al. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*, 34(13):i142–i150, 2018. ISSN 1367-4803. doi:10.1093/bioinformatics/bty266.

[89] A. M. Phillippy, M. C. Schatz, and M. Pop. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biology*, 9(3):R55, 2008. ISSN 1474-760X. doi:10.1186/gb-2008-9-3-r55.

[90] G. Narzisi and B. Mishra. Comparing De Novo Genome Assembly: The Long and Short of It. *PLOS ONE*, 6(4):e19175, 2011. ISSN 1932-6203. doi:10.1371/journal.pone.0019175.

[91] F. Vezzi, G. Narzisi, and B. Mishra. Reevaluating Assembly Evaluations with Feature Response Curves: GAGE and Assemblathons. *PLoS ONE*, 7(12):e52210, 2012. ISSN 1932-6203. doi:10.1371/journal.pone.0052210.

[92] M. Hunt, T. Kikuchi, M. Sanders, et al. REAPR: a universal tool for genome assembly evaluation. *Genome Biology*, 14(5):R47, 2013. ISSN 1465-6906. doi:10.1186/gb-2013-14-5-r47.

[93] A. Rahman and L. Pachter. CGAL: computing genome assembly likelihoods. *Genome Biology*, 14(1):R8, 2013. ISSN 1474-760X. doi:10.1186/gb-2013-14-1-r8.

[94] G. Parra, K. Bradnam, and I. Korf. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23(9):1061–1067, 2007. ISSN 1367-4803. doi:10.1093/bioinformatics/btm071.

[95] G. Parra, K. Bradnam, Z. Ning, et al. Assessing the gene space in draft genomes. *Nucleic Acids Research*, 37(1):289–297, 2009. ISSN 0305-1048. doi:10.1093/nar/gkn916.

[96] F. A. Simão, R. M. Waterhouse, P. Ioannidis, et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, 2015. ISSN 1367-4803. doi:10.1093/bioinformatics/btv351.

[97] R. M. Waterhouse, M. Seppey, F. A. Simão, et al. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution*, 35(3):543–548, 2018. ISSN 0737-4038. doi:10.1093/molbev/msx319.

[98] T. I. W. G. S. Consortium (IWGSC). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science*, 345(6194):1251788, 2014. ISSN 0036-8075, 1095-9203. doi:10.1126/science.1251788.

[99] G. Marçais and C. Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, 2011. ISSN 1460-2059, 1367-4803. doi:10.1093/bioinformatics/btr011.

[100] S. Kurtz, A. Phillippy, A. L. Delcher, et al. Versatile and open software for comparing large genomes. *Genome Biology*, page 9, 2004.

[101] G. W. Vurture, F. J. Sedlazeck, M. Nattestad, et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, 33(14):2202–2204, 2017. ISSN 1367-4803. doi:10.1093/bioinformatics/btx153.

[102] E. S. A. Sollars, A. L. Harper, L. J. Kelly, et al. Genome sequence and genetic diversity of European ash trees. *Nature*, 541(7636):212–216, 2017. ISSN 1476-4687. doi:10.1038/nature20786.

[103] T. I. W. G. S. C. (IWGSC), T. Marcussen, S. R. Sandve, et al. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome Ancient hybridizations among the ancestral genomes of bread wheat Genome interplay in the grain transcriptome of hexaploid bread wheat Structural and functional pa. *Science (New York, N.Y.)*, 345(6194):1251788, 2014. ISSN 1095-9203. doi:10.1126/science.1251788.

[104] R. Brenchley, M. Spannagl, M. Pfeifer, et al. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, 491(7426):705–710, 2012. ISSN 1476-4687. doi:10.1038/nature11650.

[105] J. A. Chapman, M. Mascher, A. Buluç, et al. A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biology*, 16(1):26, 2015. ISSN 1465-6906. doi:10.1186/s13059-015-0582-8.

[106] F. Choulet, A. Alberti, S. Theil, et al. Structural and functional partitioning of bread wheat chromosome 3B. *Science*, 345(6194), 2014. ISSN 0036-8075, 1095-9203. doi:10.1126/science.1249721.

[107] N. I. Weisenfeld, V. Kumar, P. Shah, et al. Direct determination of diploid genome sequences. *Genome Research*, 27(5):757–767, 2017. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.214874.116.

[108] K. Hodgkinson. *Short Read Sequencing Reveals Sub-Genome Structure of the Polyploid Pennate Diatom Fragilariopsis cylindrus*. Master's thesis, University of East Anglia, 2018.

[109] B. Paten, A. M. Novak, J. M. Eizenga, and E. Garrison. Genome graphs and the evolution of genome inference. *Genome Research*, 27(5):665–676, 2017. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.214155.116.

[110] E. Garrison, J. Sirén, A. M. Novak, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9):875–879, 2018. ISSN 1546-1696. doi:10.1038/nbt.4227.

[111] R. J. Challi, S. Kumar, K. K. Dasmahapatra, et al. Lepbase: the Lepidopteran genome database. *bioRxiv*, page 056994, 2016. doi:10.1101/056994.

[112] H. A. Lewin, G. E. Robinson, W. J. Kress, et al. Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, 115(17):4325–4333, 2018. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.1720115115.

[113] Darwin Tree Of Life website: https://www.darwintreeoflife.org/. .

[114] The 10+ Wheat Genomes Project website http://www.10wheatgenomes.com/. .