THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

The wheat pangenome: assembly and analysis

Juan D. Montenegro

BSc Biology

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2017

School of Agriculture and Food Sciences

Abstract

Wheat is one of the most important food crops in the world and its continued breeding is essential to achieve the production goals set by FAO in 2050. Breeding programs benefit from the development of genomic resources that reduce time and costs of selection of suitable varieties. Recent evidence has suggested that an important fraction of crop plant genomes exhibits presence-absence variation and cannot be exploited following the single reference paradigm. Pangenomic studies aim to fill this vacuum by creating a complete catalogue of genes in a species and characterizing them. With the release of the first wheat draft genome for cultivar Chinese Spring, we are able to explore the wheat pangenome. In the first chapter I performed a review of the current status of wheat genomics and pangenomic studies. In the second chapter the public wheat reference is assessed for its suitability as the basis of a pangenomic study. Extensive uncollapsed duplicated sequences and the absence of support for some gene models prompted us to reassemble the genome. Both assemblies were then compared and the new assembly was selected for further study. In the third chapter, eighteen wheat cultivars were used to extend the Chinese Spring reference. A metagenomics assembly approach was employed and 350 Mbp of additional sequence absent from the Chinese Spring reference were assembled. These sequences contained over 20,000 additional genes which were classified into core and variable genes and later characterized. The pangenome size was modelled as a function of the number of genomes and functional enrichment of the variable genes showed that these were enriched with genes involved in response to biotic and abiotic stress. In chapter 4, we use the new pangenome to identify over 34.6 million SNPs and further use these SNPs to characterize core and variable genes, to construct a high density genetic map and to assess the relatedness of the cultivars used in this study. We show that the variable genes have a higher SNP density particularly for non-synonymous SNPs. The results show that the synthetic cultivar W7984 is the most divergent accession alongside Chinese Spring. Finally, in chapter 5, the future of pangenomic studies is evaluated with a critique and suggestions to improve the current wheat pangenome.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

Peer reviewed papers

1. Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C.-K. K., Visendi, P., Lai, K., Doležel, J., Batley, J. and Edwards, D. (2017), The pangenome of hexaploid bread wheat. Plant J, 90: 1007–1013. doi:10.1111/tpj.13515
2. Visendi, P; Berkman, P. J.; Hayashi, S.; Golicz, A. A.; Bayer, P. E.; Ruperao, P.;Hurgobin, B.; Montenegro, J.; Chan, C.K. K.; Staňková, H.; Batley, J.; Šimková, H.; Doležel, J.; Edwards, D. An efficient approach to BAC based assembly of complex Genomes. Plant Methods 2016, 12, 778
3. Ghislain, M., Montenegro, J.D., Juarez, H. et al. Ex-post analysis of landraces sympatric to a commercial variety in the center of origin of the potato failed to detect gene flow Transgenic Res 2015, 24: 519. doi:10.1007/s11248-014-9854-4

Posters

1. Jara Vidalon L., Montenegro J.D., Herrera M. del R. , Ghislain M. TALEN-mediated knock-out of Ry adg candidate genes. Annual Reports 2014, International Potato Centre
2. Jara Vidalon L., Montenegro J.D., Herrera M. del R. , Ghislain M. Plex-assay to select clones with multiple copies of the Ry adg gene conferring resistance to PVY. Annual Reports 2014, International Potato Centre

Oral presentation

1. Montenegro, JD; Golicz, A; Hurgobin, B; Huey Tyng Lee, Chon-Kit Kenneth Chan, Paul Visendi, Philipp Bayer, Jacqueline Batley, David Edwards. Improved Methods for Reassembly and Analysis of the 17 Gb Bread Wheat Genome. International Plant and Animal Genome Conference XXIV, 2016.

Publications included in this thesis

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C.-K. K., Visendi, P., Lai, K., Doležel, J., Batley, J. and Edwards, D. (2017), The pangenome of hexaploid bread wheat. Plant J, 90: 1007–1013. doi:10.1111/tpj.13515

Different parts of this publication are incorporated in Chapters 2, 3 and 4.

Contributor	Statement of contribution
Juan D Montenegro	Developed assembly method (100%) Designed the study (80%) Performed the analysis (90%) Wrote the manuscript (100%)
Agnieszka A Golicz	Designed the study (10%) Edited paper (10%) Performed the analysis (4%)
Philipp E Bayer	Edited paper (10%) Designed the study (5%)
Bhavna Hurgobin	Performed analysis (2%) Critiqued the manuscript (10%)
Huey Tyng Lee	Critiqued the manuscript (10%)
Chon-Kit Kenneth Chan	Performed analysis (2%) Critiqued the manuscript (10%)
Paul Visendi	Performed analysis (4%)
Kaitao Lai	Critiqued the manuscript (10%)
Jaroslav Doležel	Critiqued the manuscript (20%)

	Edited the paper (20%)
Jacqueline Batley	Critiqued the manuscript (20%) Edited the paper (20%)
David Edwards	Designed the study (5%) Critiqued manuscript (20%) Edited the paper (40%)

Contributions by others to the thesis

David Edwards helped with the design of the study. Kaye Basford, Jacqueline Batley and David Edwards help with the revision and editing of the chapters. Paul Visendi contributed in the wheat genome annotation.

Statement of parts of the thesis submitted to qualify for the award of another degree

None

Acknowledgements

I would like to thank my wife Isabel and my son Jose Alejandro for their support, patience and their love. To my mother Carola and my brother Ivan for their constant interest and courage in the distance. To the Edwards' group and the University of Queensland for the stimulating academic environment.

To my supervisors David Edwards, Jacqui Batley and Kaye Basford I would like to thank for your guidance, mentoring, patience and support.

Keywords

Triticum aestivum, next generation sequencing, pangenome, single nucleotide polymorphisms, genetic mapping, bioinformatics, presence-absence variation, genetic diversity

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060102 Bioinformatics, 50%

ANZSRC code: 060408 Genomics, 50%

Fields of Research (FoR) Classification

FoR code: 0601, Biochemistry and Cell Biology, 50%

FoR code: 0604, Genetics, 50%

This thesis is dedicated to my son José Alejandro whose smile makes the impossible possible and to my dear wife Isabel who gives me the peace I need when I need it most.

I would not be here if it were not for you two.

Table of contents

List of figures

List of tables

List of Abbreviations

AFLP: Amplified length polymorphism

BAC: Bacterial artificial chromosome, used for cloning and transforming of DNA in bacteria

BLAST: Basic local alignment search tool

bp: Base pair of DNA or RNA

CDS: Coding sequence

CEGMA: Core eukaryotic genes mapping approach

CNV: Copy number variant

DH: Double haploid

DNA: Deoxyribonucleic acid

EST: Expressed sequence tag

Gbp: Giga base pair = 1,000,000,000 bp

GWAS: Genome wide association study

Kbp: Kilo base pair = 1,000 bp

Mbp: Mega base pair = 1,000,000 bp

ML: Maximum likelihood

MP: mate pair

Mya: Million years ago

NCBI: National centre for biotechnology information

PAV: Presence/absence variant(ion)

PCA: Principal components analysis

PCR: Polymerase chain reaction

PE: Pair end

RFLP: Restriction fragment length polymorphism

RNA: Ribonucleic acid

SAM: Sequence alignment/map format, a standard file format to store genetic read alignments

SMRT: Single molecule real time

SNP: Single nucleotide polymorphism

SSR: Simple sequence repeat

TAIR: The Arabidopsis information resource

TE: Transposable element

TGS: Third generation sequencing

VCF: Variant calling format

1 Chapter 1 Introduction and literature review

1.1 Introduction and objectives

Food production for a growing human population is a challenge in the face of decreasing access to water and land for agriculture, unpredictable changes in weather patterns due to climate change and the constant adaptation of pathogens able to spread disease to newer cultivars. The productivity gains obtained in the past 100 years are curtailed by these challenges and there is growing urgency to address these issues before food shortages and rising food prices hit those who are more vulnerable to such changes.

One way of addressing these challenges is by continuous breeding of crop plants. Crop breeding increasingly benefits from the application of molecular tools such as marker assisted selection (MAS) and the increasing availability of genomic information supports these advanced breeding tools. The decreasing cost of DNA sequencing has accelerated genomics research in recent years. Most sequencing projects have focused on reference genome assembly and the discovery of high density molecular markers like SNPs. Nevertheless, the potential of genome sequencing goes further, because it offers access to novel genetic variants that would be beneficial to successful breeding programs.

With more genomes being released every day, it is now common to perform comparisons between close species and between different individuals of a species. These comparisons have shown that an important fraction of the genome is not present in all the individuals and the genes present in these variable regions help shape the phenotype of their carrier. This discovery led to the realization that a single reference genome cannot possibly represent the entire diversity in a species and, in turn, led to the concept of the pangenome as the entity that encompasses all the genomic sequences in a species.

With the release of the first wheat reference genome, it became possible to reconstruct and explore its pangenome with the addition of sequences from a diverse array of cultivars. The pangenome will be useful for the identification of novel genes that exhibit presence-absence variation in the species, the discovery of hidden genetic variants, the association of such variants to traits of agronomic interest and their eventual introgression into the germplasm of elite cultivars. It has been suggested that variable genes may be involved heterosis; if true their annotation could have dramatic effects in the selection process of parental accessions for breeding programs.

The aims of this thesis are to construct and characterize the wheat pangenome using all publicly available data from diverse wheat cultivars and to show its utility by identifying variable and core genes, annotating intervarietal SNP variants, constructing a high density genetic map and assessing the genetic relatedness of the cultivars.

1.2 Common wheat

1.2.1 Origin and domestication of wheat

Wheat is the common name used to refer to a large and complex group of related species that have been used for human consumption for thousands of years and that was part of the first group of founder crops that were domesticated nearly 12,000 years ago in the Diyarbakir region in South East Turkey (Lev-Yadun et al., 2000, Luo et al., 2007). Wheat comprises diploid (einkorn wheat), tetraploid (durum wheat) and hexaploid (bread wheat) species both domesticated and wild and its evolution has been shaped by recurrent hybridization events with species from the genus *Aegilops* (Tsunewaki, 2009).

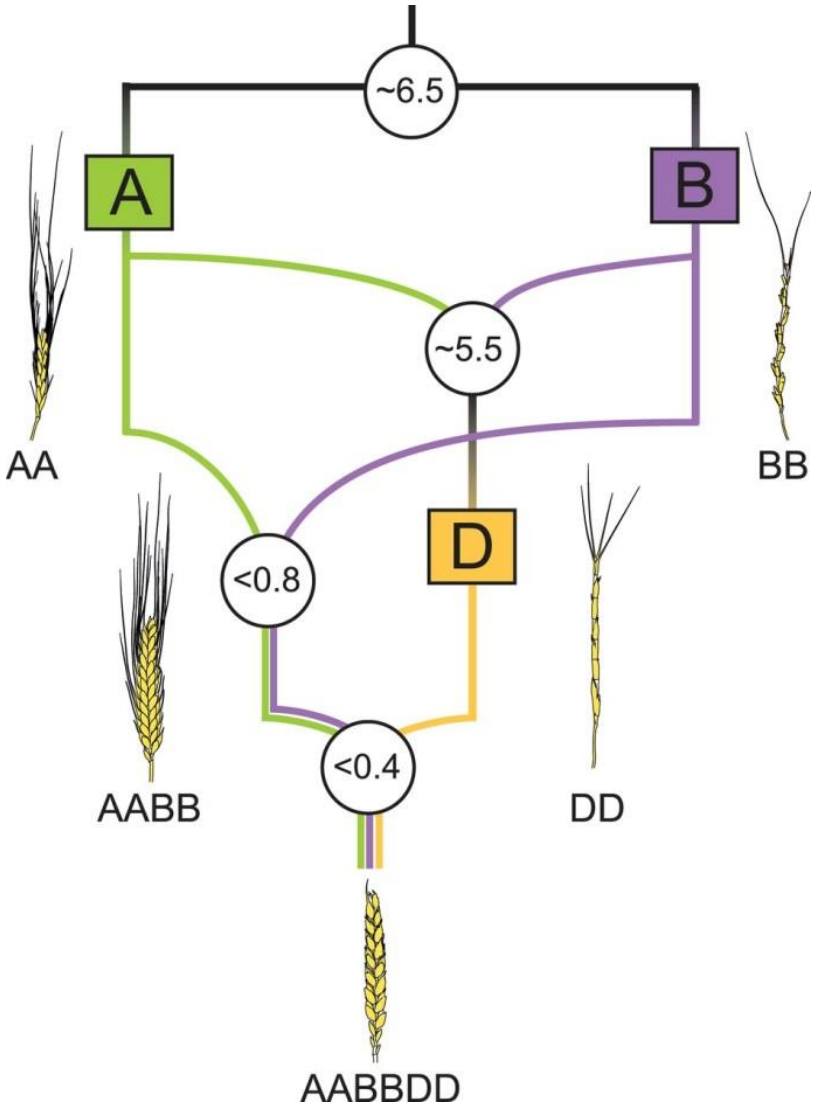
Einkorn wheat (*Triticum monococcum,* A^bA^b genome 2n = 14) is a domesticated diploid species that is a close relative of the wild *T. boeoticum*. Molecular data placed the domestication of einkorn wheat in the Karacadag mountains regions on South East Turkey (Heun et al., 1997). It was part of the "founder" crops alongside rye (*Secale cereale*), barley (*Hordeum vulgare*), lentil (*Lens culinaris*), pea (*Pisum sativum*), chickpea (*Cicer arietinum*), bitter vetch (*Vicia ervilia*), flax (*Linum usitatissimum*) and emmer wheat (*Triticum dicoccum*) (Lev-Yadun et al., 2000). During the last 5000 years, cultivation of einkorn has largely been replaced by tetraploid and hexaploid wheats (Peng et al., 2011). This replacement has decreased the selection pressure on the domesticated einkorn wheat varieties. This is clearly shown by the absence of a domestication bottleneck and the fact that domesticated einkorn has more genetic diversity than the wild *T. boeoticum* race used for domestication (the β race)(Kilian et al., 2007).

The second wild diploid wheat (*T. urartu,* A^uA^u) has not been domesticated, but it has played an essential role in the evolution of wheat by donating the A genome to all tetraploid and hexaploid species (Dvořák et al., 1993). The genome of *T. urartu* has been recently sequenced and assembled, revealing a larger gene content than its counterparts in the tetraploid and hexaploid wheats, an expansion of NBS-LRR type gene family and providing evidence for the role of repeat expansion in genome size enlargement during the evolution of the Triticeae (Ling et al., 2013).

Tetraploid wheats (2n = 28) occur naturally in the near and middle east. Two wild tetraploid species are known *T. turgidum* subsp. *dicoccoides* (wild Emmer with AABB genome) and *T. araraticum* (AAGG genome). The domesticated species *T. turgidum* subsp. *dicoccon* and *T. timopheevii* arose from their wild relatives. Due to its limited cultivation area in the Transcaucasia region, little research has been done on the Timopheevii lineage, which includes the hexaploid *T. zhukovskyi* (2n = 42, $A^mA^mA^uA^uGG$ genome). However, the Turgidum lineage, which includes durum wheat and bread wheat, has been extensively studied because of its economic importance and wide area of cultivation. Both wild tetraploids are thought to have arisen through allopolyploidization events between *T. urartu* (A^uA^u) and a species from the lineage of the wild wheat *Aegilops speltoides* Tausch (Sarkar and Stebbins, 1956, Ogihara and Tsunewaki, 1988). Two populations of domesticated emmer wheat are clearly delimited, one in the Near East (Syria, Israel, Jordan and Lebanon) and the other one on Central—Eastern Asia (Turkey) (Luo et al., 2007, Ozkan et al., 2002). Although there is not enough evidence to support a single domestication site for tetraploid wheats, it is clear that the Central Asian population played a major role in the domestication of emmer wheat. Several different cultivated wheat species derived from the domesticated emmer wheat including the Persian wheat, the Polish wheat, the Khurasan wheat, and the Durum wheat (Damania, 1998).

Durum wheat is the second most cultivated wheat species in the world after bread wheat. It derived from domesticated emmer wheat in the eastern Mediterranean region (Luo et al., 2007, Feldman and Kislev, 2007). Not only does it have a large genome (12Gbp) but it also contains a high number of paralogous genes (Dubcovsky and Dvorak, 2007). Despite the high homology between its homeologous chromosomes, these behave as diploid chromosomes during mitosis and meiosis due to a dominant gene *Ph1* found in chromosome 5B which controls the correct paring of homologous chromosomes and prevents pairing between the homeologous ones (Martinez-Perez et al., 2001). Human mediated expansion of the *T. turgidum* to the northeast of the Fertile Crescent, put it in sympatry with the goatgrass *Aegilops tauschii* (2n = 14, DD genome) which is considered the donor of the D genome in the allohexaploid *T. aestivum*.

Figure 1-1. Model of the evolution of the *Triticum/Aegilops* complex. (Marcussen et al., 2014)



Bread wheat (*T. aestivum,* 2n = 42, AABBDD genome) is the most important wheat species in the world due to its widespread use across all continents and is considered a staple food for 40% of the human population. No wild relatives of the allohexaploid wheat have been found which supports the hypothesis of its origin by hybridization of a domesticated *T. turgidum* (McFadden and Sears, 1946, Kihara, 1966) with *Ae. tauschii* as far back as 8000 years ago after farming spread from the Fertile Crescent and overlapped the natural distribution of *Ae. tauschii* (Giles and Brown, 2006). Whether allohexaploid wheat is the result of a single hybridization event or several parallel hybridization events is still under debate. By looking at 53 single-copy loci, the *NOR3* rRNA locus and the *Glu1* locus in the D genome of hexaploid wheat and in *Ae. tauschii* genome, Devorak et al

(1998) concluded that the diversity in the D genomes of the wheat varieties analyzed was not enough to support the claim of concurrent hybridization between *Ae. tauschii* and *T. turgidum* (Dvorak et al., 1998). Nevertheless, Dvorak also proposed that human selection and evolution of D genome in hexaploid wheat may have resulted in the loss of genetic diversity and the apparition of new polymorphisms absent from the *Ae. tauschii* genepool as a possible explanation for his results. On the other hand, sequencing of loci *Xwy838* and *Gss* between both species have also provided evidence for multiple polyploidization events as the main drivers of the genepool structure of current hexaploid wheat cultivars (Caldwell et al., 2004). Similarly, a study using microsatellite data found evidence of recurrent hybridization and supported the idea that the D genome of hexaploid wheat is a composite of several sources (Lelley et al., 2000). A different study using a SNP array designed for *Ae. tauschii* (Luo et al., 2013), found that most f the wheat genotypes were more closely related to *Ae. tauschii* lineage 2 (*strangulata* genepool) than to lineage 1 (*tauschii* genepool) which is strange given the extensive opportunity for crossing between tetraploid wheat and *Ae. tauschii* and supporting the monophyletic origin of hexaploid wheat (Wang et al., 2013). Finally, the draft sequence of the wheat genome (IWGSC, 2014) was compared to those of *Ae. tauschii* and *T. urartu* and the topologies of phylogenetic trees of single-copy orthologs genes of the three genomes were compared and analyze to propose a model for wheat evolution. In this model, *Ae. tauschii* originated through homoploid hybrid speciation between the B and the A genomes approximately 1-2 million years after the divergence of the A and B genomes and hexaploid wheat appeared from recurrent hybridization events between *T. turgidum* and *Ae. tauschii* (Marcussen et al., 2014).

1.2.2 Agronomic importance

Common wheat, (*Triticum aestivum*) is one of the most important food crops in the world alongside with maize, rice and potato (FAO, 2016). It is the most widely grown cereal using one sixth of the crop acreage in the world (Gupta et al., 2008). It is estimated that in 2017, nearly 750 million tons of wheat will be produced worldwide (http://www.fao.org/worldfoodsituation/csdb/en/). This product is the main source of protein and calories for 35% of the world population (http://www.idrc.ca/en/ev-31631-201-1-DO_TOPIC.html) and it is estimated to provide one fifth of the total calories consumed by humans (Pfeifer et al., 2014). Between 2010 and 2013 wheat production decreased by nearly 6% mostly due to severe weather events (Asseng et al., 2015, Lobell et al., 2011).

Nevertheless, it is estimated that in the next 50 years wheat production will need to double to keep pace with global demand (Tilman et al., 2011).

1.2.3 Wheat breeding

One way of increasing wheat productivity is by breeding which has proven to be a successful strategy for increasing yield in the past (Reif et al., 2005, Evenson and Gollin, 2003) and will continue to play and important role to overcome the challenges described before. Resistance to hydric stress, high salinity, drought, pathogen infection and low nutrient availability are among the most sought after traits. Landraces and wild relatives are an important source of genetic variants encoding for many of these traits and have been used for gene introgression in wheat (Longin and Reif, 2014, Lopes et al., 2015, Mengistu et al., 2016, Feuillet et al., 2008). Identification of heterotic groups in wheat is another field that promises increases in yield and resistance to biotic and abiotic stress (Mette et al., 2015, Zhao et al., 2015).

Breeding in wheat was focused primarily on yield and thus its production has been on the rise since the late 1960s, mostly due to wide-scale adoption of Green Revolution technologies (Evenson and Gollin, 2003). The semi-dwarf wheat varieties developed at CIMMYT are a prime example of the achievements of wheat breeding programs. These varieties contain the dwarfing genes *Rht8*, *Rht-D1b* and *RhtB1b* which prevented lodging and increased grain yield (Hedden, 2003). This was achieved by diverging nutrients away from the pathways of biomass production, which increased plant size and made it more susceptible to lodging (Tang et al., 2009, Robbins, 2009). By forcing the plant to be smaller, the nutrients were better used in seed production. Successful tests in India and Pakistan in the early sixties led to a revolution that allowed both countries to double the national wheat production in only 4 years passing from net importesr to net exporters of bread wheat (http://maswheat.ucdavis.edu/protocols/Dwarf/index.htm).

Breeding efforts are now focused on different areas including increasing the diversity in elite wheat cultivars by the development of synthetic varieties (Zegeye et al., 2014, Rasheed et al., 2014, Mujeeb-Kazi et al., 2008, del Blanco et al., 2001), increasing radiation use efficiency with the introduction of C4-like traits, increasing the nutrient partitioning to grain yield while maintaining lodging resistance and improving screening, prediction and selection methods (Poland et al., 2012a, Heffner et al., 2010, Crossa et al., 2014)to accumulate complex physiological traits with higher yielding potential (Reynolds et al., 2011, Rebetzke et al., 2009).

For years wheat breeding was curtailed by the lack of large scale genomic resources (Lai, 2015, Muhindira, 2016). Development of these resources for discovery of molecular markers was a priority in the wake of the genomics era and several projects were undertaken to build them up including the sequencing of FOSMID libraries, whole genome shotgun sequencing (Brenchley et al., 2012), sequencing of expressed sequence tags (ESTs) (Yu et al., 2004, Lazo et al., 2004) and the construction and sequencing of bacterial artificial chromosomes (BAC) libraries (Allouis et al., 2003, Šafář et al., 2010, Šafář et al., 2004). These new resources coupled with the extensive cytogenetic resources that have been developed since the beginning of the 20th century have been used to improve our knowledge of the genomic architecture of wheat (Gupta et al., 2005, Stebbins, 1947, McFadden and Sears, 1946, Kihara, 1919, Sakamura, 1918). The ultimate goal was the construction of the complete physical map of hexaploid wheat. The first draft of the wheat genome was published in late 2014 (IWGSC, 2014).

1.2.4 Wheat genetics and genomics

The reconstruction of the wheat genome was extremely difficult in large part due to its size (17Gbp) (Bennett, 1972, Smith and Flavell, 1975), its ploidy (allohexaploid 2n = 6x = 42) (Kihara, 1919, Sakamura, 1918) and the high levels of repetitive sequence estimated at around 80% (Smith and Flavell, 1975, Flavell et al., 1974, Paux et al., 2008). The challenges were such that early assessments considered it infeasible (Gill et al., 2004) and suggestions were made to instead sequence the wild diploid relatives *T. urartu* (AuAu), *Ae. tauschii* (DD) and *Ae. speltoides* (BB). However, due the extensive gene loss during hybridization and polyploidization of *T. turgidum* and *T. aestivum* (Smet et al., 2013, Kashkush et al., 2002, Soltis and Soltis, 2012, Paterson AH, 2012), as well as large genomic rearrangements (Badaeva et al., 2007), the International wheat genome sequencing consortium (IWGSC) preferred to sequence the hexaploid wheat.

Remarkably, despite its high levels of repetitive sequence and numerous orthologous/paralogous loci between homeologous chromosomes, chromosome pairing during cell division occurs correctly. This behaviour is controlled by the Ph1 locus in the B genome and has been responsible for the relative genome stability exhibited by wheat (Martinez-Perez et al., 2001, Griffiths et al., 2006).

Numerous studies have shown that blocks with conserved gene order have remained relatively unchanged in most grasses since the common ancestor (Murat et al., 2010). These syntenic regions can still be found in diverse grasses (International Brachypodium,

2010, Huo et al., 2009, Akhunov et al., 2013) and have been used to order contigs into syntenic blocks and pseudomolecules (Pfeifer et al., 2013, Berkman et al., 2013a, Berkman et al., 2012a, Berkman et al., 2011b, Mayer et al., 2009, Mayer et al., 2011).

Several sources of evidence have shown that the subgenomes of wheat (A, B and D) have very different characteristics. The D genome has lower sequence diversity than the A and B genomes, with B being the most diverse one. This been shown time and again using various lines of evidence. Microsatellite diversity shows that the D genome had fewer alleles than the A or B genomes (Huang et al., 2002b, Prasad et al., 2000, Plaschke et al., 1995). Comparisons of genetic diversity between hexaploid wheat and its wild diploid relatives demonstrated that there was little difference in diversity between the tetraploid wheats (AABB) and hexaploid wheat, but there was a large loss of diversity between Ae. Tauschii and the D genome of wheat (Haudry et al., 2007, Reif et al., 2005). Genetic mapping assays have shown that the D genome consistently has fewer markers and recombination bins (Wu et al., 2015, Li et al., 2015c, Chen et al., 2012, Sorrells et al., 2011, Semagn et al., 2006, Gupta et al., 2005, Peng et al., 2004, Lazo et al., 2004, Kam-Morgan, 1988). It has been suggested that the little diversity found in the D genome is due to the little gene flow between the wild relative *Ae. Tauschii* and the hexaploid varieties compared to more frequent gene flow from tetraploid wheat and common wheat (Reif et al., 2005, Berkman et al., 2013a).

The D genome has also been shown to contain a higher gene density (Wang et al., 2014, Qi et al., 2004, Berkman et al., 2013a). Extensive gene loss has been shown in the early stages after interspecific hybridization and genome duplication (Chen and Ni, 2006, Wendel and Doyle, 2005, Adams and Wendel, 2005, Kashkush et al., 2002, Pestsova et al., 2001, Wendel, 2000, Ramsey and Schemske, 1998) and it has been suggested that the A and B genomes, having experienced two rounds of genome duplication after hybridization compared to one single round for the D genome may be the cause of this difference. However, it has been shown that after polyploidization and rapid gene loss, one genome takes a dominant role in gene expression and is less prone to diversification of the duplicated homeologous genes (Woodhouse et al., 2014, Parkin et al., 2014, Grover et al., 2012, Schnable et al., 2011). The latter theory of genome dominance would also explain the scaling diversity between the three sub-genomes with B being the most divergent, followed by A and finally the D subgenome. This has been explained by assigning the A genome a temporary dominance over the B genome after the formation of tetraploid wheat. This would have increased the selective pressure on the A genome

compared to the B. After the second hybridization between the tetraploid and *Ae. tauschii*, the dominant role was taken by the D genome, which relaxed the constrains on the A genome (Pont et al., 2013).

1.3 Current status of genome sequencing and assembly

The description of the dideoxy chain termination method for sequencing DNA by Sanger et al. (1977) (Sanger et al., 1977) and the publication of the first full genome sequence (phage phiX174) (AIR et al., 1977) laid the foundations for the modern automated sequencing machines that we see today. The dominance of the Sanger method until the early 2000 was then replaced by modern high-throughput sequencing methods like Illumina and 454. More recently, these sequencing platforms are being complemented with third-generation sequencing technologies, which reduce throughput in benefit of longer, low quality reads. Here I will do a short description of these sequencing technologies.

1.3.1 Sanger

It is considered the gold standard of sequencing technologies and has been used to complete several genome sequencing projects until the advent of high throughput sequencing technologies.

This sequencing method relied on PCR to incorporate dideoxy nucleotides that caused early termination of the elongating chain. By preparing 4 reactions, each with a different dideoxy nucleotide, and resolving the PCR fragments in a polyacrylamide gel, it was possible to determine the order of the nucleotides in the DNA chain. The term sequencing by synthesis was coined to refer to those sequencing protocols that depend on the use of DNA polymerase to determine the correct order of the nucleotides.

1.3.2 Roche 454

One of the first high-throughput sequencing technology which relied on sequencing-by-synthesis to determine the DNA sequence. The technology relied on pyrosequencing, which uses luciferase to emit light using the pyrophosphate released after the addition of a nucleotide to the growing chain. The amount of light emitted was directly proportional to the number of nucleotides added in a single cycle. Thus 454 used four consecutive cycles of A, C, G and T used separately and measured the recorded the amount of light emitted in each cycle (Rothberg and Leamon, 2008). This combined with advanced microfluidic

control and massive parallelization allowed higher production of sequences in a short period of time.

This technology produced medium-sized reads from 100 bp to up to 500 bp. Unfortunately, the error rate was greater than Sanger technology and it was particularly prone to insertion-deletion bias, particularly in homopolymeric regions because it was difficult to differentiate between the peak of light produced by five or more identical consecutive nucleotides (Luo et al., 2012a, Mariette et al., 2011). Roche stopped providing support for this technology in 2016.

1.3.3 Illumina

Previously known as Solexa, this is the most popular sequencing platform currently available. It has dominated the sequencing market for the past 10 years in large part due to its high-throughput, which has reduced the price per nucleotide sequenced and the high accuracy of the reads. The technology has prioritized quality and throughput over length of the reads (Cronn et al., 2008, Rougemont et al., 2008). Currently they produce reads as long as 300 bp, although in its beginnings the average read length was 32 bp. The technology also depends on sequencing-by-synthesis and uses high definition cameras to record the addition of nucleotides to several thousands of DNA chains in parallel. These DNA templates are kept in place with the use of beads embedded in the sequencing flowcell. Each bead contains a population of identical templates and thus emits the same fluorescence when a similar nucleotide is added. The use of several clones per bead is used to increase the fluorescence emitted and thus its accuracy (Quail et al., 2012).

The main drawback for this technology is the small size of the reads produced which makes it difficult to resolve long repetitive regions in complex genomes. This problem was somewhat alleviated by the development of paired-end and mate-pair libraries (Leggett et al., 2014). Both libraries allowed the use of longer templates, although the full template will not be sequenced. Instead, reads are generated from the ends of the templates. These library construction techniques increase the long range information stored by the library and can be used to produce scaffolds when enough long-distance evidence from the read pairs supports the connection (van Heesch et al., 2013).

1.3.4 ABI Solid

Solid sequencing was based on the sequencing-by-ligation technology, which used the stringent hybridization of fluorescent–labelled dinucleotides to a template followed by

imaging of the fluorescence emitted to determine the sequence of the DNA template. The use of dinucleotides ensures that every base is read twice and thus increases the accuracy of base calls and improved SNP detection and sequencing error detection (Goodwin et al., 2016, Valouev et al., 2008). The use of dinucleotides forces the system to replace the base calling approach for a colour scheme, where each colour represents 4 possible dinucleotides. This colour-based approach generates coloured-encoded sequences where that needs to be deconvoluted prior to its use in downstream applications. The lack of direct compatibility with downstream analysis tools and the fact that Solid still lags behind in read length with the current platform producing reads of 75bp have made Solid the least used sequencing platform. This sequencing platform is no longer being commercialized.

1.3.5 Pacific Biosciences

This is the most popular third generation sequencing technology commercially available. It is capable of producing reads an order of magnitude larger than current second generation sequencing technologies. These longer reads offer the possibility of producing more contiguous genome assemblies by resolving long repetitive regions that the short reads were unable to complete accurately (Roberts et al., 2013). In prokaryotes, full length chromosome-size contigs are being routinely assembled using only PacBio reads (Uchimura et al., 2016, Korlach et al., 2010). In addition to repeat resolution, the use of longer reads allows the production of phased haplotypes in diploid or polyploid organisms by reducing ambiguity in the assembly graph. PacBio reads are also being actively used in the sequencing of full-length transcripts that can be used to detect complete isoforms and to improve genome annotation pipelines (Ashby et al., 2017, Rhoads and Au, 2015, Abdel-Ghany et al., 2016, Gonzalez-Garay, 2016).

This technology uses the DNA polymerase as the engine of sequencing and allows direct observation of DNA polymerization on real time. The development of the a specialized flowcell with thousands microwells with of zero-mode wavelength is used for the recording of fluorescence emitted by a single nucleotide incorporated at a time (Levene et al., 2003, Korlach et al., 2010, Eid et al., 2009). Despite its many uses, PacBio still lags behind in sequence accuracy and throughput compared to other SGS technologies. PacBio reads often have 15% error rate and currently requires either large amounts of coverage (≥80X) or SGS reads to improve the quality of the raw reads via error correction. Both approaches increase the costs of sequencing. Nevertheless, the

technology has ample space for improvement. Circular consensus sequences (CCS) produce slightly shorter reads (~4 Kbp) with fewer errors (5-10%). With continuous improvements in sequencing accuracy, PacBio is already displacing prior SGS technologies in the *de novo* assembly field.

1.3.6 Oxford Nanopore

Oxford nanopore sequencing technology is another TGS platform with huge potential that has been in development for the past 5 years. The recent launch of the Minion sequencing platform which promises easy real-time sequencing with minor library preparation protocols seems to aim at the personalized genomics field. The technology offers reads longer than those provided by PacBio with some reads reported to be as long as several hundred kilobases long (http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/). As with PacBio, it offers a plethora of applications that are not currently viable with SGS technologies like haplotype phasing, highly contiguous genome assemblies and full length transcripts. However, the error rate of these reads is greater than that of PacBio with an average accuracy of (70%) (Mikheyev and Tin, 2014). Currently, nanopore sequences have been used for the *de novo* assembly of the *Escherichia coli* genome (Loman et al., 2015) or for hybrid de novo approaches (Goodwin et al., 2015).

The technology is based on the use of transmembrane proteins able to carry a complete DNA molecule from one side of the membrane to the other without breaking it. With every nucleotide passing the sequencer records the changes in the electric potential of the membrane. Ideally, such change would depend exclusively on the nucleotide that is currently passing through the pore. In reality, up to 6 nucleotides can influence the change in the electric potential of the membrane and that is what complicates base calling. In order to increase the accuracy of base calls from nanopore raw data, sophisticated algorithms based on hidden Markov models and more recently neural networks have been developed to produce more accurate base calls (Boža et al., 2017, David et al., 2017). However, development of accurate base calling algorithm is still an area of active research.

1.3.7 Current algorithms in *de novo* assembly

De novo assembly is the process through which a genome sequence is reconstructed from overlapping fragments of the longer original sequence. Common sense dictates that, in order to reconstruct the original sequence, comparisons between all

fragments must be made and scored, high scoring alignments can be merged into a single consensus sequence and these can, in turn, be extended by lower scoring alignments until all the fragments (reads) are included in the sequence. This naïve approach has been modified and implemented in several public assemblers.

However, despite the simplicity of its formulation, genome assembly is never as straightforward as it seems. The algorithms need to be optimized to deal with the error biases inherent to each sequencing technology. For example, 454 data is more prone to indel errors than Illumina data and PacBio data is more prone to erroneous base calling than Sanger sequencing. Reads' lengths also play a role in the type of implementation that is going to be used for assembly.

On top of that, the existence of large repetitive regions, nearly identical repeats, highly conserved orthologous sequences and different heterozygosity levels increases the challenges that need to be overcome by the assembly algorithms. Thus, different flavours of the same algorithmic approach have been implemented and published. Advances in sequencing technologies will require the fine-tunning of these assembly implementations to deal with the specific error profile of each technology.

1.3.7.1 Overlap-layout-consensus approach

As the name suggests the overlap-layout-consensus algorithm (OLC) has three main steps. In the first step all the reads are aligned to each other in an all vs all fashion and significant overlaps are kept. In the second step, the reads are ordered to form a path, eliminating low quality paths from the graph. In the final stage, ordered aligned reads are used to calculate a consensus sequence. This consensus is usually determined by the sequence quality of the aligned reads and the majority rule, where the consensus nucleotide is the one with the highest frequency in a certain position.

Although this approach tends to produce the most contiguous and better finished genomes, it is usually prohibitively slow, particularly for larger genomes or high coverage datasets. It also uses large amount of memory because all reads must be stored in memory during the overlapping and consensus steps. These drawbacks make it of little use when attempting the reconstruction of large genomes like wheat or barley or when using very high coverage like those obtained with Illumina reads.

This algorithm has been implemented in many assemblers including the wgs-assembler (Myers et al., 2000), PCAP (Huang et al., 2003), Newbler and Phrap (de la

Bastide and McCombie, 2007). This algorithm has also been used in the assembly of high quality reference genomes like Drosophila melanogaster (Myers et al., 2000) and human (Levy et al., 2007). The algorithm is better suited for long reads like the ones produced by Sanger or 454. More recently minor modifications were made to take advantage of PacBio and Oxford nanopore reads. WGS-assembler is routinely used for the *de novo* assembly of bacterial genomes to finished high quality reference genomes using only PacBio reads (Koren and Phillippy, 2015, Scott and Ely, 2015, Liao et al., 2015).

1.3.7.2 De Bruijn graph

The De Bruijn graph algorithm is based in the decomposition of the original reads into smaller kmers of a fixed size. This decomposition is done for two reasons: 1) to reduce memory consumption storing all essential data and getting rid of redundant kmers and 2) to reduce the number of comparisons that need to be made between reads. The decomposition of reads into kmers is usually immediately capitalized by the construction of a de Bruijn graph which connects all kmers with a minimum coverage cut-off through edges of size k-1. The algorithm then travels across the graph and removes low coverage paths, lose ends and resolves bubbles or splits bubbles in the graph. These steps are called graph simplification and are used to remove. The final step is to find an eulerian path across the graph that maximizes the number of nodes visited (Pevzner et al., 2001). This approach is usually faster and less memory intensive than the OLC approach which turns previously unthinkable genomes for de novo assembly into targets to improve the accuracy of the algorithm.

Unfortunately, the decomposition of reads into kmers loses the sequence information from individual reads and can lead to missassemblies caused by repeats longer than the kmer size selected. To overcome this issues, different modifications to the original deBruijn graph have been made including coloured deBruijn graph (Iqbal et al., 2012), rectangle graphs (Vyahhi et al., 2012) or by using several different kmer sizes consecutively and combining the results into a single graph (Bankevich et al., 2012, Peng et al., 2012, Peng et al., 2010). Another improvement was the addition of scaffolding steps to use the information stored in paired-end and mate-paired reads to produce larger contiguous sequences by stitching contigs with enough support from mate-pair or paired-end data and estimated the distance and orientation of the contigs based on the meta data stored in the reads (Boetzer et al., 2011, Pop et al., 2004).

Due to the dominance of short illumina reads in the genomes sequencing market, deBruijn graph assemblers are the most used assemblers available. There are many open source implementations available like Velvet (Seemann, 2012, Zerbino et al., 2009, Zerbino and Birney, 2008), Spades (Bankevich et al., 2012), Soap De novo (Luo et al., 2012b) or Abyss (Simpson et al., 2009) all of which implement slightly optimized versions of the deBruijn graph. Recently, deBruijn graphs have been proposed as efficient structures to facilitate long read error correction (Salmela et al., 2017, Tischler and Myers, 2017).

1.4 Pangenomic studies

The exponential increase in the number of sequenced genomes in the last decades has made evident that large structural variations between individuals of the same species have taken place. This observation has raised concerns that a single reference genome cannot represent the entire sequence diversity present in a population (Saxena et al., 2014, Golicz et al., 2016a). A considerable number of sequences are affected by copy number variations (CNV) (Żmieńko et al., 2014) which are pervasive in all organisms including human (McCarroll and Altshuler, 2007, Iakoubov et al., 2013), maize (Swanson-Wagner et al., 2010) and cyanobacteria (Schirrmeister et al., 2012). An extreme case of CNV variations are the presence-absence variation (PAV). In this type of polymorphisms, a sequence is present in one individual, but absent in another. In wheat the existence of CNVs and PAVs is well documented. A targeted resequencing study of 3,497 genes between two wheat varieties showed that 85 genes exhibited CNV while 10 genes showed PAV (Saintenac et al., 2011). Another study on flowering gene *Ppd-B1* found that cultivars with increased copy numbers of this gene flowered earlier (Díaz et al., 2012). To gain access to those genes and use them for breeding programs a pangenome reference must be constructed and annotated.

The term pangenome was originally coined by Tetellin et al (2005) in his description of comparative genome organization of various strains of *Streptococcus agalactiae* (Tettelin et al., 2005, Medini et al., 2005). The original definition of the pangenome was the sum of all genes present in all individuals in a species. Of course it is impossible to expect that all individuals will be sequenced, but by means of a mathematical extrapolation, Tettelin et al (2005) showed that a plateau in the total number of genes could be reached after modelling the increase as a function of the number of genomes included. He further divided the genes in the pangenome in two groups: those present in all individuals, which

accounted for 80% were the core genes. The other group was the one that exhibited presence-absence variation and was not essential for the survival of any strain. He termed the latter group dispensable genome. The concept of pangenome has grown to include non coding sequencing that could have an effect on the phenotype and ultimately all sequences that exhibit PAV regardless of their biological function (Vernikos et al., 2015).

Since its inception, pangenomic analysis have been performed on several different bacterial species with different levels of resolution, while some aimed to the species level (Schoen et al., 2008), other aimed at the genus level (Jacobsen et al., 2011) or the class level (Collins and Higgs, 2012). In all cases, it was possible to classify the nature of the pangenomes constructed based on the convergence towards a maximum number of genes present in the group analysed. While modelling a pangenome expansion as a function of the number of genomes analysed, it is possible to determine if the addition of more genomes will lead to a convergence in the total number of genes in the pangenome or not. If the total number of genes converges towards a plateau, this means that the pangenome is closed, because the total number of genes is estimated to be finite. However, if the gene count does not stabilize and appears to grow indefinitely, the pangenome is said to be open and the addition of more genomes will not increase the size of the pangenome.

1.4.1 Pangenome analysis

Most pangenomic studies aim to understand the dynamics of gene gain, loss or evolution in a species by estimating the total potential size of the pangenome if all individual were sequenced or by estimating individual contribution of a single individual to the total gene pool of the species. Another interesting focus is on the core and variable genes. As explained above, the core genes are expected to be essential to delimit the identity of the species and include all those genes that make it unique and different from all other. Knowing the functions and the effects on phenotype of these genes would help design better classification tools and set clearer boundaries between species. It may even help in the redefinition of the term “species”. Focus on the variable genes would shed light on the particular adaptations acquired by different individuals and the mechanisms through which these variable genes were acquired or how they originated.

Mathematical regression and modelling of the pangenome expansion are common features of pangenomic analyses. These are used to estimate different metrics that define the characteristics of the pangenome like total gene content, average gene contribution

per individual and average number of unique genes per individual. Several tools have been designed to perform the analysis for bacterial genomes. These tools include BPGA (Chaudhari et al., 2016), PanGP (Zhao et al., 2014) and PanSeq (Laing et al., 2010) and they differ in the approach they take to estimate total gene content of the pangenome. In general they follow three steps to do the estimation: first, subsets with all possible permutations from 2 to X genomes are generated; second, the total number of non-redundant genes and common genes is calculated for each permutation and finally, all points are used to estimate the parameters of a model where X is the number of genomes and Y is either the total number of non-redundant genes (pangenome size) or the total number of common genes (core genome). The main difference between the programs lies in the simplifications taken at every step. These simplifications are designed to reduce the amount of time taken by the analysis, particularly when including hundreds of genomes, while reducing their impact on the estimations. For example, some implementations do not calculate all possible combinations of genomes, but rather choose a random sample of combinations for each value of X (X = number of genomes). Other implementations use different smaller samples of the genomes analysed and estimate the final values for the entire set based on the convergence of the models estimated for the smaller subsets (Laing et al., 2010).

There are many factors that influence the results of a pangenome study. However, given that these studies need to estimate the number of common and total non-redundant genes, accurate gene annotation is of utmost importance to reduce the number of pseudogenes and annotation artefacts that may influence the final estimations. Accurate annotation of genomes relies on the integration of several lines of evidence which include empirical evidence like full-length cDNA, RNA-seq and protein alignment or probabilistic data, like coherence of the gene model, length of the open reading frame (ORF), presence of exon-intron boundaries, 3’ and 5’ untranslated regions, start codon, end codon, polyadenylation signals, codon usage frequency among others (Simão et al., 2015, Campbell et al., 2014, Yandell and Ence, 2012, Holt and Yandell, PGSC, 2011, Cantarel et al., 2008, Elsik et al., 2006, Conesa et al., 2005).

The classification of genes between core and variable is also an important part of the analysis and needs to be taken with caution. The fact that a gene appears to be variable in a species does not mean that its underlying function is also variable. In fact, function redundancy is a common feature of plant genomes (Moore and Purugganan, 2005). So it is important to distinguish core functions and core genes and by extension variable

functions and variable genes. In the face of incomplete functional annotation, orthology may be used for the determination of functional clusters (Li et al., 2003, Gabaldon and Koonin, 2013).

1.4.2 Plant pangenomic studies

As has been mentioned before, pangenomic studies have their foundations in the discovery of structural variants and copy-number variations of which presence-absence variations are an extreme example. Large-scale copy number variations in the human genome have been reported since the mid 2000's (Iafrate et al., 2004, Sebat et al., 2004) and have been linked to cancer (Lee et al., 2007, Yoshihara et al., 2011) and other degenerative diseases (Liao et al., 2012, Pankratz et al., 2011). However, their study in plant genomes is relatively new.

Comparative genome hybridization assays in maize uncovered thousands of examples of structural variation. Springer et al. (2009) designed a CGH array based on the B73 reference genome and discovered over 3,000 CNVs and PAVs. The distribution of these variants was not homogenous along the genome with long stretches of little if any polymorphism between B73 and Mo17 and pockets of high CNV frequency (Pankratz et al., 2011). A closer focus on protein coding regions of the B73 maize and compared with 19 inbred maize cultivars and 14 wild teosinte samples, revealed nearly 4,000 genes with some degree of copy-number variation (Swanson-Wagner et al., 2010). Belo et al (2009) suggested that genes exhibiting PAV may play a role on heterosis (Beló et al., 2009). More recently, the use of NGS has facilitated the discovery and reconstruction of large structural variants between maize genotypes. Resequencing of six elite maize inbred varieties revealed hundreds of complete genes that exhibited presence-absence variation many of which were shared by various cultivars (Lai et al., 2010). Gore et al (2009) used a resequencing approach to compare genomic diversity in 27 inbred maize lines. He found that the B73 genome contain approximately only 70% of the total gene pool available for maize (Gore et al., 2009). The reconstruction of representative transcripts assemblies from 502 diverse maize cultivars, showed that nearly 50% of them did not show evidence of being present in the B73 reference genome (Hirsch et al., 2014). In the same study, it was shown that the maize pan genome was closed and that further sequencing of more cultivars would result in limited gene discovery. SGS was further used to sequence over 14,000 elite inbred maize plants, polymorphic tags were identified and 26 million were identified of which 1.1 million were identified as true PAV tags (Hirsch et al., 2014).

In rice, CGH analysis were used to explore the genomic diversity of rice lines showed that around 700 genes showed some kind of copy number variation with the majority indicating a loss of copies compared to the Nipponbare reference genome. (Yu et al., 2011). Resequencing of 40 inbred lines and 10 wild relatives discovered 1415 new genes that were absent in the Nipponbare genome, 48% of which were present in only one accession. Also nearly 1300 genes were lost in at least one cultivar compared to Nipponbare (Xu et al., 2012). Whole genome comparisons between Nipponbare and var 93-11 showed that at elast 10% of the genes were under PAV or CNV (Ding et al., 2007). These studies were extended by two landmark papers comparing thousands of rice accessions and discovering large amounts of genes under presence absence variation. The first one was done using a low-coverage sequencing of ca. 2,000 rice accessions. The unmapped reads were reassembled using a metagenomics approach. This resulted in the discovery of thousands of new genes that were absent from the Nipponbare reference genome. The authors used linkage'disequilibrium to place 78% of SNP-containing genes into the Nipponbare reference genome. Characterization of these dispensable genes showed that they nearly half of them were transposable elements-related proteins and from the remaining, those with functional annotation revealed an enrichment with disease resistance genes, salt stress response and zinc finger proteins (Yao et al., 2015). Finally, the sequencing of over 3,000 rice varieties from 89 countries resulted in the construction of the first crop plant pangenome. This study found at least 12,000 novel genes that were absent from the Nipponbare reference genome and that have been placed in the rice pangenome. The authors have also generated a graphical representation of the pangenome in the form of a genome browser that is accessible here: http://cgm.sjtu.edu.cn/3kricedb/ (Sun et al., 2017, Li et al., 2014a, The 3000 Genome Project, 2014).

In *Arabidopsis thaliana*, early comparisons of three divergent ecotypes (Col-1, Bur-1 and Tsu-1) using a resequencing approach identified >3.4 Mbp of sequence that was highly dissimilar, deleted or duplicated relative to the Col-1 reference genome (Ossowski et al., 2008). A combination of CGH and whole genome shotgun sequencing to compare 5 different *A. thaliana* ecotypes detected 55,000 medium indels that affected over 1,500 genes in all the ecotypes. Transposable elements were overrepresented in the genes affected by this indels (Santuari et al., 2010). Later, consecutive studies using whole genome shotgun sequencing, identified hundreds of genes under CNV or PAV including 130 genes that were completely lost in the Ler-ecotype (Cao et al., 2011, Lu et al., 2012).

Sequencing of 18 *A. thaliana* genomes identified between 2.1 and 3.7 Mbp of sequence missing from these accessions, but present in the reference Col-1.(Gan et al., 2011). Tan et al. (2012) after resequencing 80 accessions of *A. thaliana* found that around 10% were absent in at least one accession analysed (Tan et al., 2012). A characterization of genes showing PAV found thata in average these were shorter and younger than genes that did not exhibited PAV (Bush et al., 2014).

As sequencing technologies become cheaper, reads longer and more accurate, the number of projects aiming to construct catalogues with the entire gene pools of plant species will become more common. Other pangenomic studies have been performed in soybean (Li et al., 2014c) and *Brassica oleracea* (Golicz et al., 2016b) while exploratory studies have discovered and characterized gene PAV or CNV in important food crops like wheat (Saintenac et al., 2011), potato (Hardigan et al., 2016, Iovene et al., 2013) and sorghum (Zhang et al., 2014).

1.5 Genetic variation

Genetic variation is the result of the natural tendency of DNA sequences to change over time. These variations are usually brought about by mutation and can be fixed in a population by positive selection or removed from it by negative selection. Some mutations are not under any selective pressure so their retention in the population depends on genetic drift. Variations are important because they offer opportunities of survival in the face of a changing environment. Variation can be used by breeders to incorporate a desired trait into an already stablished cultivar or to introduce more diversity to the domesticated stock. However, it has only been after the discovery of DNA, that we were able to access and use the source of all variability. Nowadays, we routinely use molecular markers to identify and exploit the variability found in the genome.

1.5.1 Molecular Markers

Molecular markers are DNA sequences that help discriminate different individuals and that belong to a specific locus in the genome. These sequences are useful because they can be to assess diversity between and within closely related species, to produce genetic linkage maps and to link genotypic information with phenotypic traits that help understand the molecular basis of morphologic characteristics (Edwards and Batley, 2004). Molecular markers have proven useful in breeding programs by facilitating selection

and reducing the time and costs traditionally associated with extensive phenotyping efforts.

1.5.1.1 Restriction fragment length polymorphisms (RFLP)

First developed by Williams (1989), this is a type of marker that finds differences in the migration patterns of bands. It essentially has 4 steps 1): total genomic DNA digestion with a single restriction endonuclease; 2) separation of restriction fragments of different sizes using electrophoresis; 3) transfer of the DNA to a suitable membrane for hybridization and 4) hybridization with labelled probes (Williams, 1989). The underlying principle is that the probes will hybridized only to a handful of well defined sequences in the genome. If the restriction sites in two individuals are different, then the size of the fragments carrying the hybridization targets will also be different and that will be visible as a different migration .pattern of the bands in the membrane.

This was the first DNA fingerprinting technique developed and it was widely used prior to the development of more advanced techniques. The advantage of this technique is that most markers are co-dominant so both alleles in a diploid individual can be observed. The downside of these technique is that it is labour-intensive, requires large amounts of DNA, it is not scalable and requires a long time to prepare.

1.5.1.2 Amplified fragment length polymorphism (AFLP)

Developed by Vos et al. (1995), this technique also searches for differences in the migration patterns of bands on a gel after electrophoresis and relies on the use of restriction enzymes to produce the fragments. The main difference with RFLP lies in the use of polymerase chain reaction (PCR) to obtain observable amount of DNA. It consists of 5 steps: 1) total genomic DNA is digested with two different restriction enzymes; 2) adapters are ligated to the restriction fragments 3) two consecutive selective PCR rounds are performed, each more stringent than the former; 4) the final PCR product is diluted and amplified fragments are separated by electrophoresis; 5) silver staining is used to reveal the banding pattern (Vos et al., 1995).

The main advantage of this technique is that it can reveal hundreds of differences between samples in a single assay. It also requires smaller amounts of starting DNA and does not rely on DNA transfer to a hybridization membrane or the use of labelled probes. Finally, a single library with adapter-ligated fragments can be used in many assays and produce different patterns by changing the selection conditions during PCR. The downside

is that it is not comparable between gels, the large amount of bands makes it difficult of genotype and it is impossible to distinguish the fragments of one chromatid from another so all fragments are treated as dominant.

1.5.1.3 Simple sequence repeats (SSR)

Also called microsatellites, this type of markers was discovered in the early 1980's (Tautz and Renz, 1984), but their use as universal molecular markers for DNA fingerprinting was not realized until the mid 1990's when PCR and the first automatic DNA sequencers were available (Tautz and Schlötterer, 1994). Since then its uses have multiplied and for over a decade they were the preferred method of DNA fingerprinting in many organisms and even today many studies phylogenetic studies use SSR as their preferred markers (Weber and Wong, 1993, Zietkiewicz et al., 1994, Guo et al., 2014, van Belkum et al., 1998, Chen et al., 2012, Singh et al., 2011).

Microsatellites are short sequences composed of tandem repeats of simple or complex motifs flanked by relatively well conserved sequences. They occur naturally in the genomes of all known organisms and they show remarkable intraspecific polymorphism. The difference detected between alleles of a microsatellite is in the number of repeats of the motif, which changes the molecular weight of the fragment containing the repeat. These markers are codominant and all alleles present in a sample can be equally represented. The location of microsatellites in the genome is well conserved in a specie and frequently the markers can be transferred from a species to a close relative (Fan et al., 2013, Barbara et al., 2007, Satya et al., 2016).

1.5.1.4 Single nucleotide polymorphism (SNP)

These are the most abundant type of markers available and their high-throughput detection has only recently become possible thanks to the arrival of next generation sequencing technologies and array genotyping technologies (Edwards et al., 2013). These technologies are able to detect changes in the DNA sequence composition with a resolution of a single base and are thus able to detect changes in a single nucleotide in several hundred or thousand base pairs of sequence. The abundance of this type of markers makes them perfect for the construction of high-density genetic maps (Raman et al., 2014, Iehisa et al., 2014), to improve the detection of quantitative trait loci via genome-wide association studies (GWAS) and genome-wide linkage-disequilibrium analysis (Edwards et al., 2013, Batley and Edwards, 2007).

The development of high quality NGS technologies has made the discovery of SNPs a routine task and millions of SNPs can be accurately discovered and genotyped in a single experiment (Lai et al., 2015b, Lorenc et al., 2012). However, accurate SNP identification faces many challenges. The occurrence of long nearly identical repeats and nested repeats can lead to false-positive SNP identification regardless of the discovery method used. This is even more problematic for polyploid species where the homeologous chromosomes will share a high number of conserved sequences that can confound the discovery algorithm (Lorenc et al., 2012). Access to longer reads provided by third generation sequencing technologies will help solve the challenges posed by repetitive sequences in SNP discovery. Nevertheless, the low base-calling accuracy of current TGS technologies keeps them away from such applications.

2 Chapter 2 Reassembly of the wheat genome

2.1 Introduction

A reference genome is a digital representation of the complete haploid DNA sequence of an organism. They are usually organized in large sequences called pseudomolecules that show the actual physical order of nucleotides in a chromosome. These reference genomes are constructed from small DNA fragments that are sequenced independently and then put together into larger contiguous sequences in a process called genome assembly. In its essence, the assembly process tries to deduce a consensus sequence from a large number of smaller and overlapping sequences. However, this deduction is not straightforward, in part due to the large size of genomes compared to the smaller fragments ("reads") that were sequenced and also due to the intrinsic biological complexity of genomes which include different types and levels of transposable elements, nearly identical repeats spread across the genome and varying levels of heterozygosity. On top of these challenges, the assembly process also needs to deal with the erroneous nature of DNA sequencing technologies which add an extra layer of complexity.

A reference genome is a valuable resource for the development and discovery of other genomic resources that help improve the efficiency of selection in breeding programs by reducing the labour, time and costs needed to maintain and screen large populations (Tester and Langridge, 2010). Modest genomic resources have already been used for marker assisted selection in several plant and animal species (Collard and Mackill, 2008, Xu and Crouch, 2008, Beuzen et al., 2000). Modern genotyping techniques have been widely adopted in breeding programs due to their convenience and high throughput. More recently, advances in sequencing technology have made whole genome sequencing approaches more accessible to the scientific community. These next-generation sequencing (NGS) platforms have been used in resequencing efforts to provide data for variant discovery and marker development. Furthermore, these NGS technologies are being used for genotyping by sequencing (GbS) of large populations ushering in a new era of genomic selection where multiple traits can be screened and selected for in a single assay (Heffner et al., 2010, Jannink et al., 2010, Heslot et al., 2012, Newell and Jannink, 2014). Genotyping by sequencing can be applied to species without a reference genome, but with a higher false-positive variant discovery rate due to undiscovered nearly identical repeats in the target genome. Using a reference genome for genotyping increases the

quality of discovered variants and consequently their utility in marker based crop improvement (Lorenc et al., 2012, Lai et al., 2015b). Thus, the construction of a reference genome is a valuable resource to increase the efficiency of variant discovery, genotyping and ultimately selection by means of marker assisted or genomic selection.

Reference genomes have been produced from individuals of every major clade including many plants (CoGePedia, 2017, Ensembl, 2017, Michael and Jackson, 2013). These reference genomes are useful because they offer an easier and faster analysis pipeline where sequence variations, marker development, gene identification and motif discovery can be done quickly and with high accuracy. Several grass genomes have been sequenced and assembled so far. The first grass genome ever sequenced was rice (Goff et al., 2002, IRGSP, 2005). Its sequence revealed the existence of more than 35,000 genes many of which had orthologs in other grass genomes and shared syntenic regions. The success of the rice genome assembly using the whole genome shotgun approach, was followed by the sequencing of other grasses like *Sorghum bicolor* (Paterson et al., 2009), *Brachypodium dystachion* (IBI, 2010), *Hordeum vulgare* (Mayer et al., 2012) and in recent years wheat and its wild relatives (Ling et al., 2013, Jia et al., 2013). These sequencing projects revealed that the presence of transposable elements and repeat regions is an important factor in genome size variability, but does not necessarily imply larger gene content. Gene content is more closely related to the time since the most recent genome duplication event. Immediately after a genome duplication event, the new polyploid tries to balance the altered gene expression levels caused by the transcription of redundant genes and to stabilize gene networks that may have been altered by it through a process termed “diploidization” (Clarkson et al., 2005, Conant et al., 2014). The longer since the last genome duplication event, the more time the diploidization process has had to reduce the number of redundant genes either by gene loss, sub-functionalization or neo-functionalization.

Accurate reconstruction of the wheat genome is an extremely challenging task mainly due to the very high content of repetitive elements estimated between 60% - 90% of the genome (Smith et al., 1976, Smith and Flavell, 1975, Smith, 1976) and to the presence of three homeologous genomes (A, B and D) (Sakamura, 1918) which were brought together by two consecutive hybridization events (Sarkar and Stebbins, 1956, Dvořák et al., 1993, Peng et al., 2011). Early assessments by different groups concluded that a whole genome *de novo* sequencing of the 17Gb Chinese Spring genome was not feasible given the technology and assembly algorithms available at the time (Gill et al., 2004). Instead, the

wheat scientific community decided to focus on exploring the gene coding space using different sequencing approaches like methylation filtration (Rabinowicz et al., 1999) and C_0t-based cloning and sequencing (Peterson et al., 2002). In a series of follow up studies evaluating both methods, it became evident that neither was effective enough to drastically reduce the content of repetitive elements or to greatly increase gene enrichment. In a methylation-filtration essay, Li et al (2014) found little gene enrichment in the filtrated fraction and the repeat content remained relatively high between 60-70% and did not offer any advantage for wheat genome sequencing (Li et al., 2004). Lamoureux et al. (2005) showed that Cot based filtration resulted in higher gene enrichment (14-fold) and repeat depletion (3-fold), but such enrichment was still not high enough compared to similar essays in maize and rice (Lamoureux et al., 2005). A later study by Šimková et al. (2007) combined Cot based cloning and sequencing (CBCS) with chromosome specific genomic libraries of chromosome arm 1BS, showed that CBCS was too labour-intensive, time consuming and costly to be considered a progressive method in the analysis of large genomes (Šimková et al., 2007). The construction of the physical map of chromosome 3B using a BAC-by-BAC approach (Paux et al., 2008) was a major milestone and provided the proof of concept for the assembly of the wheat genome using a BAC-by-BAC approach, which is, to this date, the gold standard and the main objective for the International wheat genome sequencing consortium (IWGSC). The assembly of orthologous group representatives (OGR) from whole genome 454 shotgun reads, allowed the identification of the majority of wheat genes and the characterization of gene loss during the formation and evolution of hexaploid wheat. This study showed that there was a sharp reduction in the number of genes by comparing the gene family sizes in the hexaploid to the same families in their diploid relatives. The level of gene loss observed, however, was not as pronounced as that seen in maize and *Brassica rapa* (Brenchley et al., 2012). The development of chromosome-specific BAC resources by the Doležel group using flow cytometry of ditelosomic genetic stocks (Šafář et al., 2010) and the subsequent assembly of the chromosome arm 7DS from flow-sorted DNA (Berkman et al., 2011b) which contained 88.5% of 7DS-mapped cDNA sequences laid the foundations for a chromosome-wise approach to sequencing the wheat genome. The successful sequencing and assembly of all group 7 chromosome arms (Berkman et al., 2013a) was then followed by the release of a chromosome-based draft genome of wheat (Mayer et al., 2014).

More recently, newer versions of the Chinese Spring reference have been released using a whole genome shotgun sequencing approach instead of the isolated chromosome arms approach. Using deeper sequencing and a combination of multiple libraries with different insert sizes TGAC and NRgene have been able to assemble 13 GB and 14.5 Gb of the Chinese Spring genome respectively (Kersey et al., 2016, IWGSC, 2016). Although detailed descriptions of the assemblies'' workflows have not been published yet, the TGAC assembly has released a draft manuscript of their method (Clavijo et al., 2016), using a newly developed genome assembler called w2rap-contigger (BIOINFOLOGICS, 2016) based on the Discovar assembler (Weisenfeld et al., 2014, Love et al., 2016) and customized to better deal with the repetitive nature of plant genomes.

In this chapter, a description and analysis of the IWGCS v2 assembly is performed, showing that it contains a high level of sequence duplication and gene annotations that are not supported by raw read data from isolated chromosome arms. Facing these issues, a local *de novo* reassembly and annotation of the Chinese Spring genome was performed using the same public libraries released by the IWGSC, but following a different assembly workflow. This new assembly was compared to the IWGSC v2 reference genome and found to be larger, to contain more genes and fewer duplicated sequences. The coding space of the local reassembly was also assessed for completeness by finding core eukaryotic genes and universal single copy orthologs genes where 98% and 97% of both datasets could be identified. Gene prediction supported by RNA-seq data, green plant ESTs and grass protein sequence homology identified 118,000 gene models in the new Chinese Spring reference genome. A comparison with the TGAC v1 Chinese Spring reference genome showed that the local assembly was highly collinear to it and contained >99% of the genes annotated in it. Furthermore, the local reassembly was able to pinpoint miss-assemblies in the TGAC v1 reference. The assembly and annotation of this reference genome is the first step in the construction and analysis of a wheat pangenome.

2.2 Methods

2.2.1 Reassembly of the wheat genome

2.2.1.1 Raw data

The IWGSC v2 reference genome and transcriptome were downloaded from the URGI repository (https://urgi.versailles.inra.fr/download/iwgsc/Survey_sequence/). The TGAC v1 wheat reference genome and transcriptome were downloaded from the Ensembl Plant genomes ftp (ftp://ftp.ensemblgenomes.org/pub/plants/release-34/fasta/triticum_aestivum/).

Chromosome-sorted raw reads were downloaded from the National Centre for Biotechnology Information (NCBI) Short read archive (SRA) database. Raw 454 reads from whole genome shotgun sequencing of Chinese Spring were also downloaded from the SRA database (Appendix 1). RNA-seq data was downloaded from the URGI repository (https://urgi.versailles.inra.fr/files/RNASeqWheat/).

2.2.1.2 Analysis of the IWGSC v2 wheat genome reference

2.2.1.2.1 Level of sequence duplication

The level of sequence duplication was assessed by aligning every chromosome arm assembly against itself with BlastN v 2.2.30 (Camacho et al., 2009) using standard parameters and an e-threshold of $1e^{-10}$. Blast results were converted to tabular format and filtered based on the following criteria: all Blast alignments where the query and the target were the same sequence were removed from analysis; alignments shorter than 1Kb or with less than 100% sequence identity were also discarded. The total sequence duplication was calculated by adding the total length of high scoring pairs (HSP) remaining after filtering and dividing it by two to compensate for reciprocal alignments (alignments where query and target sequences switched positions).

2.2.1.2.2 Gene content

The sequences of the high confidence gene predictions from the IWGSC v2 annotation were extracted from the genome assembly with the faidx module of the Samtools package (Li et al., 2009a) to obtain the unspliced transcripts and these were used as a reference for the mapping of raw reads from the isolated chromosome arms where they were annotated. Bowtie2 v 2.2.1 (Langmead and Salzberg, 2012) was used for

the alignment of the raw reads with standard parameters. Genes with no reads mapping to them were considered unsupported.

2.2.1.3 De novo assembly

The quality of the raw reads was assessed using FastQC (Andrews) with k=7. Clonal reads were removed using an in-house script (remove_clones.pl), quality trimming and adapter clipping was performed using Trimmomatic (Bolger et al., 2014). All sequences shorter than 73 bp were removed. Velvet (Zerbino and Birney, 2008, Seemann, 2012) was used for assembly using a kmer size of 71 for all chromosome arms except group 7 chromosome arms which had been previously assembled using k=63 (Berkman et al., 2013a) and 3B which was not sequenced by the IWGSC.

2.2.2 Assessment of assembly quality

2.2.2.1 Horizontal and Vertical coverage

Pre-processed reads from each of the isolated chromosome arms were mapped back to the local reassembly using Bowtie2 v 2.2.1 (Langmead and Salzberg, 2012). Also, whole genome shotgun 454 reads were downloaded from NCBI's sequence read archive (SRA) and mapped to the whole wheat genome local assembly. Unfiltered alignments were then used to calculate per base coverage with Samtools (Li et al., 2009a) and the package Sushi (Phanstiel et al., 2014) was used to generate coverage plots. The expected coverage was calculated based on the flow cytometry estimation of each chromosome and the number of reads aligned to the reference. The ratio observed/expected coverage was determined.

2.2.2.2 Gene content

Blast+ v 2.2.30 was used to align high confidence gene models from the IWGSC genome annotation to the local reassembly with a maximum e-value of 1e-5. The genes from the IWGSC were organized in two groups: 1) found and 2) missing. Both groups were further characterized and compared based on length, GC-content, average intron length and number of exons per gene Kb using R (R Core Team, 2014).

2.2.2.3 Assembly completeness

CEGMA (Parra et al., 2007) was used to assess the completeness of both the IWGSC and the local reassembly with standard parameters. The search was performed

chromosome by chromosome and the results were later combined to produce an overall output for the IWGSC and the local reassembly. Also, BUSCO (Simão et al., 2015) was used to determine the presence of universal single copy orthologs in the local assembly annotation using the gene protein sequences for comparisons with the embryophyta odb9 database downloaded from the Busco web page (http://busco.ezlab.org/).

2.2.2.4 Comparison with IWGSC v2 assembly

The wheat published reference genome was downloaded from the URGI repository (https://urgi.versailles.inra.fr/download/iwgsc/Survey_sequence/). Assembly statistics (N50, average length, gene content, sequence duplication, completeness) were calculated as described for the local assembly. Additionally, a subset of gene models was evaluated for gene size, GC content and average intron length.

2.2.2.5 Comparison with TGAC v1 assembly

Gene models were aligned to the local assembly with nucleotide Blast+ v 2.2.30 and significant alignments (E ≤ 1e-5, length ≥ 100, and similarity ≥ 99%) were considered valid. Collinearity was measured by aligning the local assembly to the 100 largest scaffolds of the TGAC v1 assembly. First, nucleotide Blast+ v2.2.30 was used to identify the local contigs with highest sequence similarity to the 100 largest TGAC v1 scaffolds. Then the candidate sequences were extracted and realigned using Mummer 3.0 (Delcher et al., 2002).

2.2.3 Genome annotation

RNA-seq reads produced by the IWGSC were downloaded from the URGI repository (Appendix I) and mapped to the newly assembled wheat genome using TopHat2 (Kim et al., 2013) with standard parameters. RepeatMasker (Smit, 2013-2015) was used on the reassembled genome using the repeat consensus library version 20150807 downloaded from RepBase (Jurka et al., 2005) on March 2016. In parallel, Augustus (Stanke and Morgenstern, 2005) was used to predict gene models using external hints obtained from the RNA-seq alignments produced by TopHat2 (Kim et al., 2013, Trapnell et al., 2009). Predicted gene models were first filtered by size (≥ 300bp). To further filter the gene set, gene annotations were intersected with the RNA-seq alignments and repeat-masked region of the genome using Bedops v 2.4.15 (Neph et al., 2012). Genes with no support from RNA-seq alignments or overlapping masked regions were not considered for further analysis.

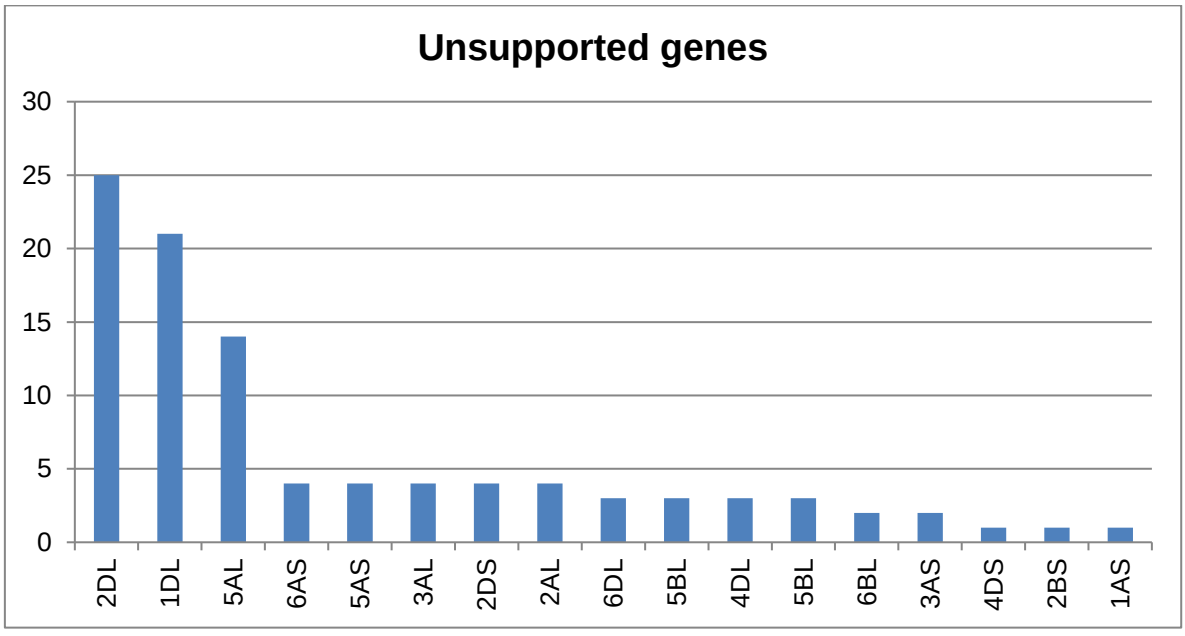
The selected genes were aligned to a dataset of transposable elements-related proteins with NCBI Blast plus (Camacho et al., 2009). Genes with significant alignments (E ≤ 1e-5) to TE-related proteins were discarded from the final dataset. Finally, the remaining genes were also aligned to the proteomes of *Brachypodium dystachion*, *Aegilops tauschii* and *Triticum Urartu* using BlastP. Wheat genes were considered split and were later considered as one if they filled three criteria: 1) both genes were aligned to the same protein, 2) neither alignment was completely contained in the other alignment and 3) all genes were annotated in the same chromosome arm.

2.3 Results

2.3.1 Analysis of the IWGSC v2 wheat genome reference

In order to use a reference wheat genome as the base for the reconstruction of the wheat pangenome it was important to make sure it contains as large a fraction of the Chinese Spring genome as possible. First, the raw reads of the isolated chromosome arms were mapped to the sequences of the raw transcripts and 99 genes were not supported by any read (Figure 2-8). Unsupported genes were found in 17 chromosome arms and ranged from 25 unsupported genes in chromosome arm 2DL to 1 gene in chromosome arms 4DS, 2BS and 1AS.

Figure 2-1 Number of unsupported genes per chromosome arms. Predicted genes in the IWGSC reference genome to which no raw reads could be mapped were considered unsupported. In total 99 genes were not supported by any read



Additionally, the level of sequence duplication per chromosome arm was measured to ensure that each assembly contained either unique sequences or collapsed repeats as expected from deBruijn graph-based assemblies. The results showed that nearly 7% of the IWGSC assembly consisted of identical repeats of 1Kb or larger and some chromosome arms contain over 40% of their total assembly length as duplicated sequences.

2.3.2 *De novo* assemblies

Preliminary tests to define the best sequencing approach were performed on chromosome arm 1DS and the results were compared based on the assembly metrics. Kmer sizes from 61-101 were tested and it was found that K=71 showed the best N50 metrics, which reflects contiguity, and the largest assembly length. Based on those preliminary results, chromosome arms from homeologous groups 1 to 6 of hexaploid wheat (*Triticum aestivum*) cv. Chinese Spring were assembled using the selected kmer size (k=71). Along with the previous assembly of group 7 chromosome arms (Berkman et al., 2013a) and chromosome 3B (Paux et al., 2006) this produced a complete assembly of the wheat genome with a size of 10.7Gb or 67.6% of the size estimated by flow cytometry (17Gb). The average N50 per chromosome arm is 1128 bp and the average number of contigs per chromosome arm is 500,000 (Table 2-1).

Table 2-1. Metrics of the reassembled chromosome arms. The assemblies were performed using Velvet with a kmer size of 71. The average N50 per chromosome is 1,128 bp and the length of their largest contigs is 20 Kbp.

Chromosome Arm	Total bases	Avg size	N50	Biggest Contig
1AL	445579452	346.857	319	21087
1AS	163387547	540.249	722	18962
1BL	465163099	434.370	479	22755
1BS	226719066	451.332	504	21604
1DL	284989697	322.872	300	7732

1DS	128838171	855.715	2443	45844
2AL	306895321	433.166	451	31709
2AS	278022451	416.272	421	23210
2BL	345083249	493.620	591	21016
2BS	318310217	906.993	2803	67657
2DL	212991684	519.075	655	15886
2DS	169562057	577.496	825	15850
3AL	216906018	538.818	694	13810
3AS	177915899	512.939	632	11756
3DL	442156102	396.303	405	16414
3DS	221006526	351.995	331	9175
4AL	343235154	447.318	486	36577
4AS	259037864	448.121	490	29018
4BL	217840717	652.608	1132	22338
4BS	306613646	474.660	540	28632
4DL	292358514	654.402	1035	49582
4DS	130037318	658.267	1191	33545
5AL	277214147	400.895	391	15105
5AS	186149024	555.840	771	17023
5BL	416526338	749.257	1696	59021
5BS	171830241	917.759	2150	29280
5DL	533659458	335.088	307	13345
5DS	140586497	441.125	470	16894
6AL	211206811	607.536	926	20528
6AS	194886487	459.829	497	27709
6BL	432415963	439.674	481	15671
6BS	496350206	436.214	484	19329
6DL	200926319	421.622	421	20837
6DS	127176870	580.351	847	16244

2.3.3 Quality assessment of the reassembly

2.3.3.1 Horizontal and vertical coverage

The pre-processed reads used in the genome reassembly were mapped back to the local assembly to determine the horizontal and vertical coverage of the assembly. The average mapping efficiency was 90%, which suggests that 90% of the sequence present in the raw data is also represented in the assembly. The reads covered 100% of the bases assembled with an average coverage that ranged from 40X to 400X (Figure 2-2). Interestingly, the average coverage was 1.5 fold higher than expected based on the expected chromosome sizes (Šafář et al., 2010). The vertical coverage does not appear to be homogeneous with a notorious difference around 375 Mbp which corresponds to the joining point between the 1DL and 1DS assemblies (Figure 2-2). Also, whole genome shotgun single end reads sequenced using 454 technology (Brenchley et al., 2012) were mapped to the local assembly. The sequencing depth was 5X (85Gb) based on a genome size of 17Gb. The average mapping efficiency was 51.5% for the 137 libraries and resulted in an average vertical coverage of 2X. Approximately 30% of the genome was not covered by reads from this subset.

2.3.3.2 Completeness of the genome

In order to evaluate the completeness of the wheat genome assembly, plant core eukaryotic genes (CEG) were searched in a chromosome wise fashion. Taken together, 245 of the 248 CEGs evaluated (98.8%) were found as either partial or complete genes (Figure 2-3). Similarly, search of universal single copy orthologous genes (USCOs) reported the presence of 98.2% of plant USCOs (72% complete and 25% partial) and only 1.8% of missing genes (Figure 2-4).

Figure 2-2. Comparison of vertical and horizontal coverage of chromosome 1D. A) Whole genome shotgun reads produced by 454 sequencing. B) Chromosome sorted reads produced by Illumina sequencing

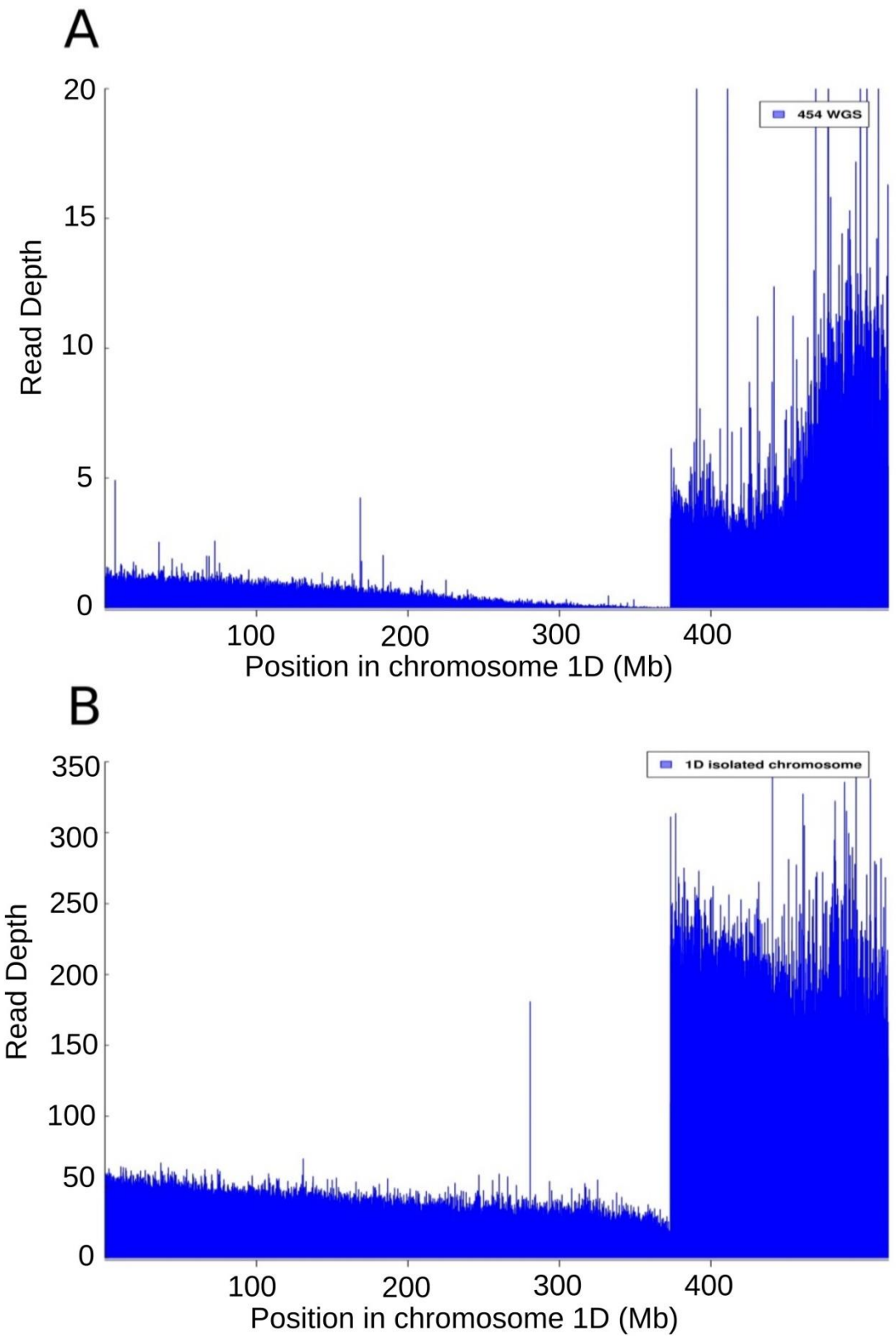


Figure 2-3. Comparison of eukaryotic gene content. CEGMA was used to determine the presence of complete, partial or missing core eukaryotic genes in both wheat assemblies. A) IWGSC v2; B) local reassembly. In total the local reassembly (Australian assembly) contained two more CEGs than the IWGSC v2.

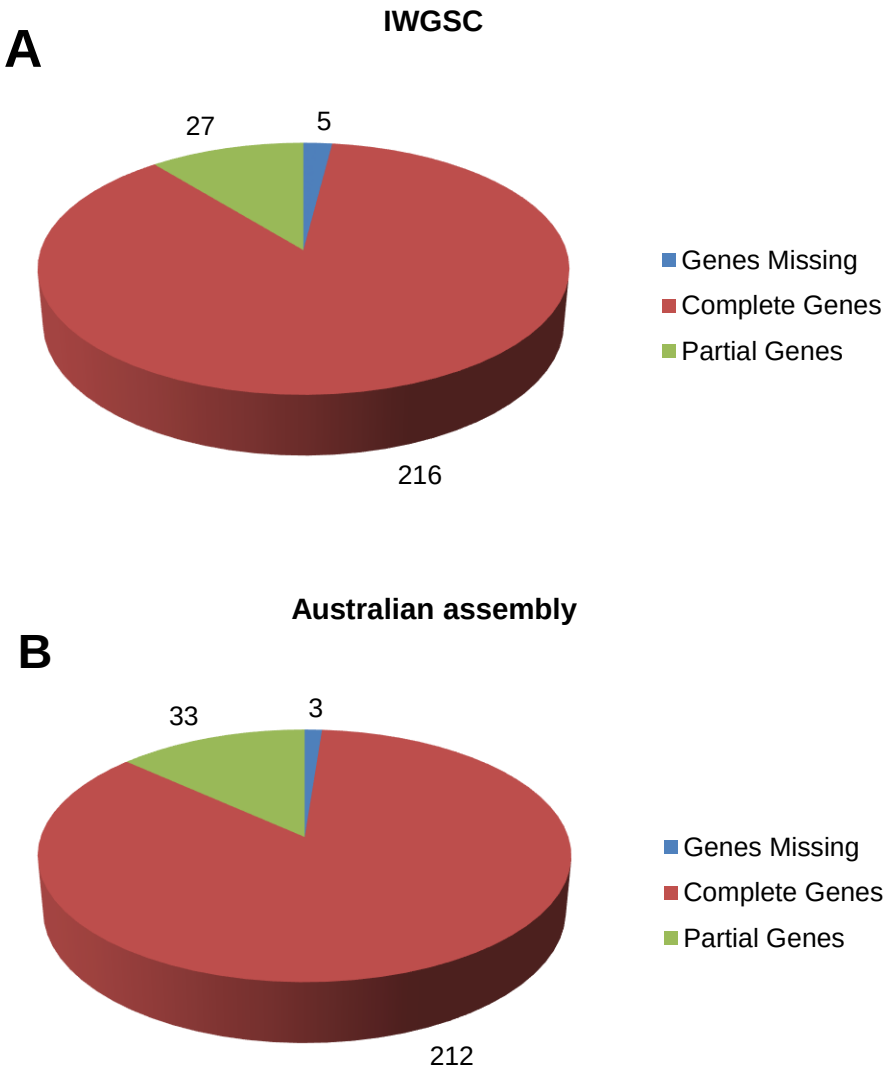
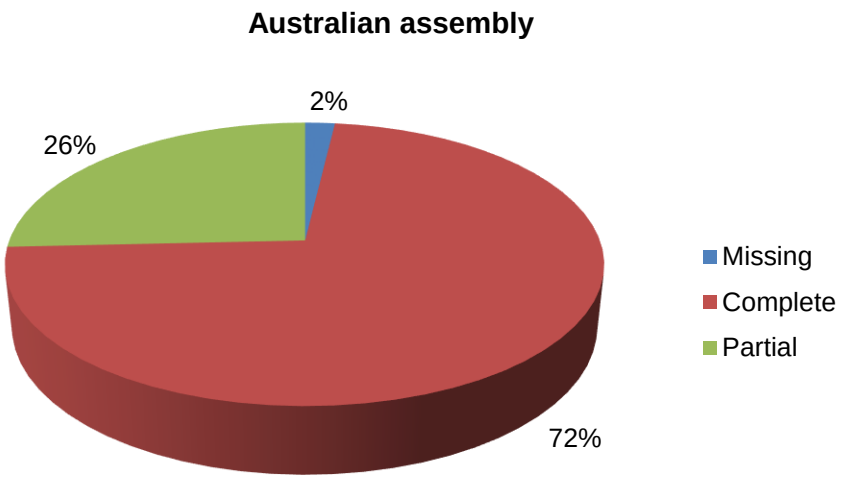


Figure 2-4. Presence of universal single copy orthologs in the local reassembly. BUSCO was used to determine the presence or absence of consensus gene models from the embryophyta odb9 database. Only 2% of the USCOs could not be found in the local reassembly.



2.3.3.3 Comparison with IWGSC v2

Overall, the local assembly of the wheat genome is larger and more complete than the public reference (Figure 2-3 and Figure 2-5). In some cases including 1AL, 1BL, 3DL and 3DS, the fraction assembled is larger than that of the IWGSC reference. Chromosome arms 6BS and 5DL showed an assembly size larger than that estimated by flow cytometry (Safar et al., 2010).

The N50 and total number of contigs were compared between the local assembly and version 2 of the IWGSC (Figure 2-6 and Figure 2-7). Both figures highlight the fragmented nature of our assembly compared to the public reference. On the other hand, the local assembly shows very low levels of sequence duplication (0.04%) compared to the IWGSC reference (7%) which in some cases exceeds 40% of intrachromosomal duplications (Figure 2-8). The presence of high levels of sequence duplication in the IWGSC public reference is likely to be an artefact created by the use of the parallel deBruijn graph assembler, AbySS (Simpson et al., 2009).

Figure 2-5. Comparison of total assembly length between the local reassembly and the IWGSC v2. Overall the local reassembly contains a larger fraction of the Chinese Spring genome. The local reassembly is represented by the red bars and the IWGSC v2 assembly by the blue bars.

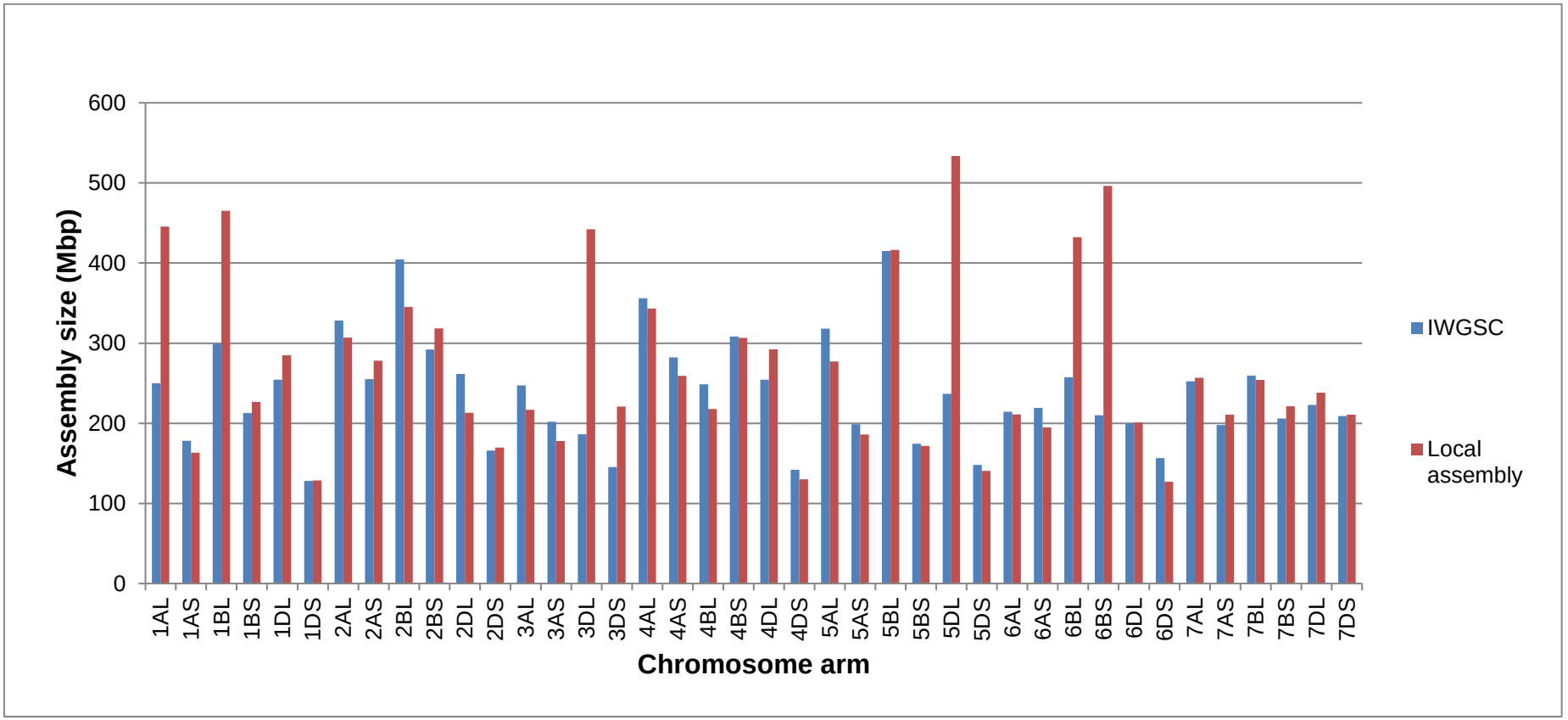


Figure 2-6. Comparison of N50 metrics between the IWGSC v2 assembly and the local reassembly. Overall, the local reassembly had smaller N50 values, suggesting a more fragmented assembly. IWGSC v2 assembly is represented by blue bars and the local reassembly by red

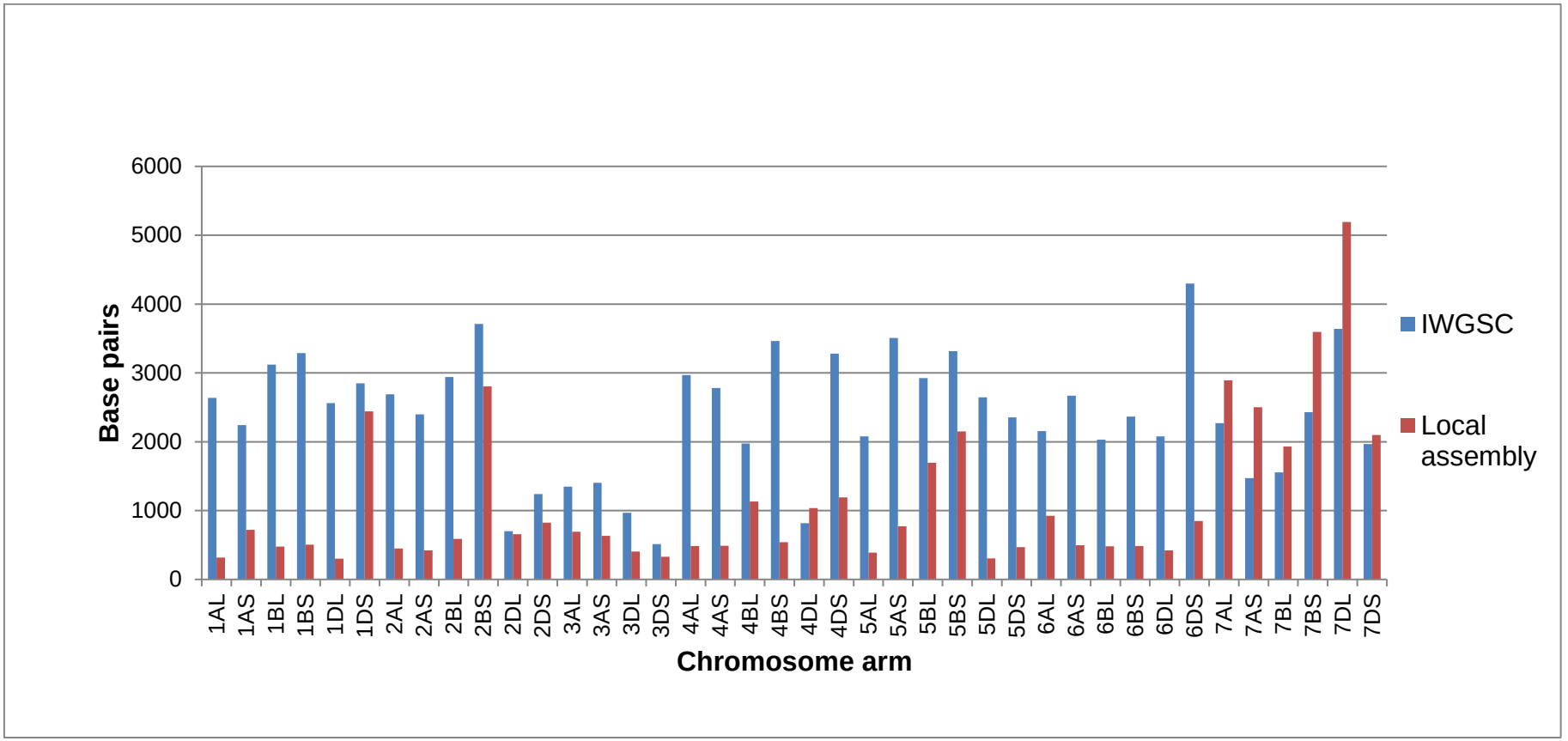


Figure 2-7. Comparison of the total number of contigs per chromosome arm. Overall, the IWGSC 2 assembly contained fewer contigs with the exceptions of 2DL and 4DL. (Blue bars: IWGSC v2, red bars: local reassembly)

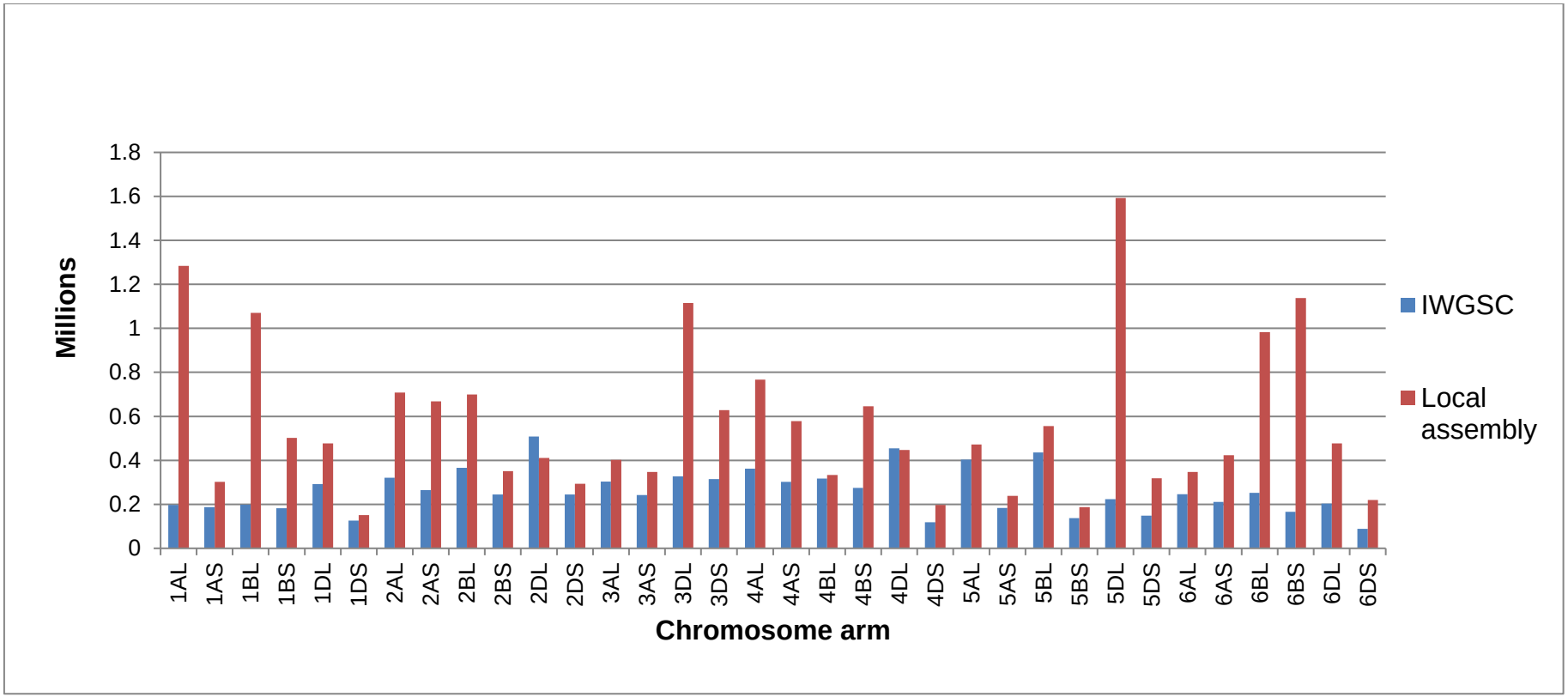
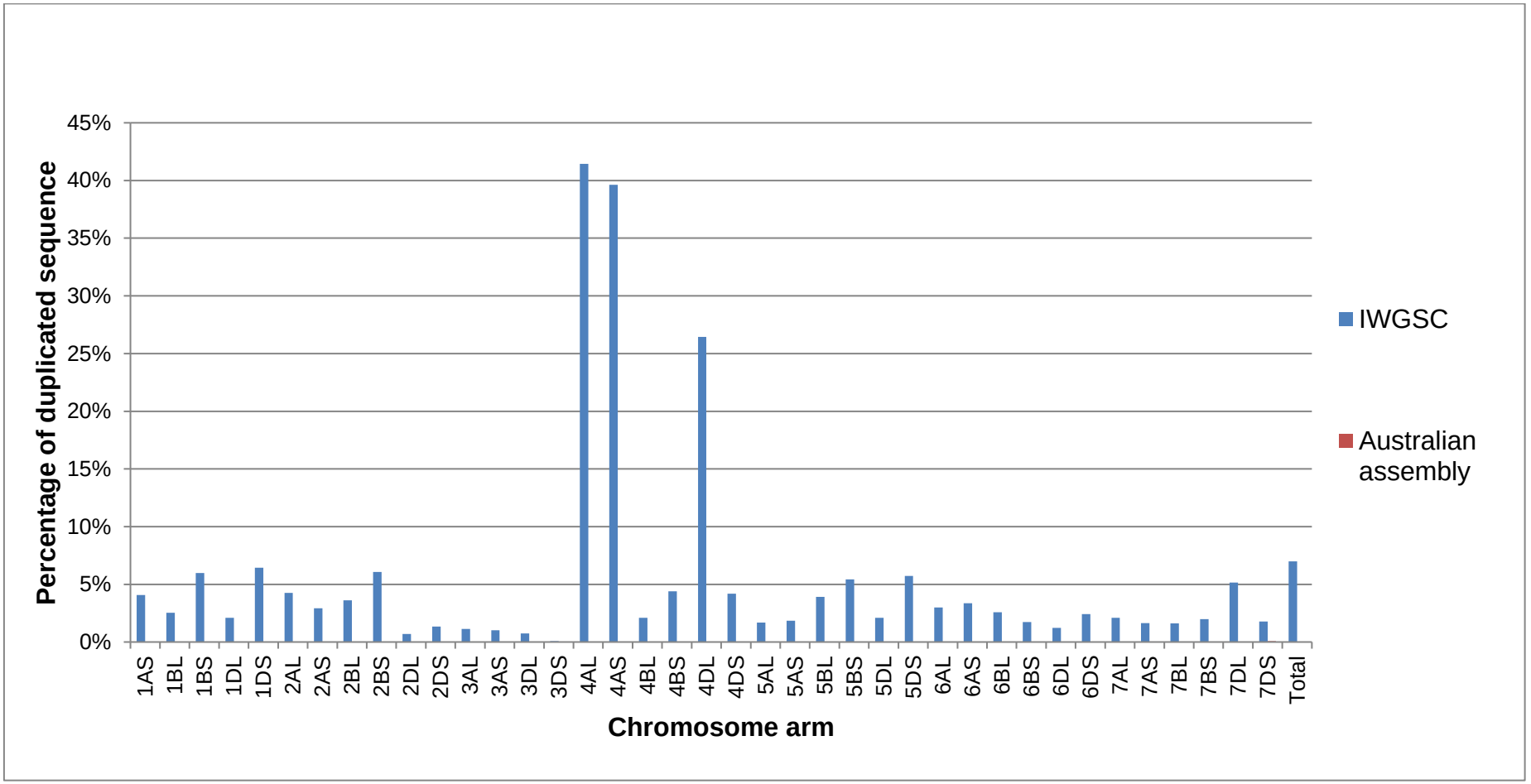


Figure 2-8. Comparison of the level of sequence duplication between the IWGSC v2 and the local reassembly. The local reassembly showed much lower levels of sequence duplication (0.004%) compared to the IWGSC v2 (7%). For some chromosome arms, the level of sequence duplication was close to 40% of the toal assembly size in the IWGSC v2 (4AL and 4AS), whereas for the rest the average duplication was 2% of the total assebmly size.



To compare the biological content of both assemblies, 93,525 high confidence gene models from the public reference were mapped to the local assembly. This subset of genes excludes those from chromosome 3B which was not assembled in this study and therefore cannot be compared. As shown in Figure 2-9, 2173 gene models could not be found in our assembly. Further characterization of the missing genes showed that they are generally much smaller, showed a higher GC content and contained smaller introns than the rest of genes which suggests that these are in fact pseudogenes (Figure 2-10, Figure 2-11 and Figure 2-12).

Figure 2-9. Distribution of genes annotated in the IWGSC v2 assembly and absent in the local reassembly. All gene models from the IWGSC v2 reference were aligned to the local reassembly in a chromosome-wise fashion. Genes with no significant alignments were considered missing and further characterized.

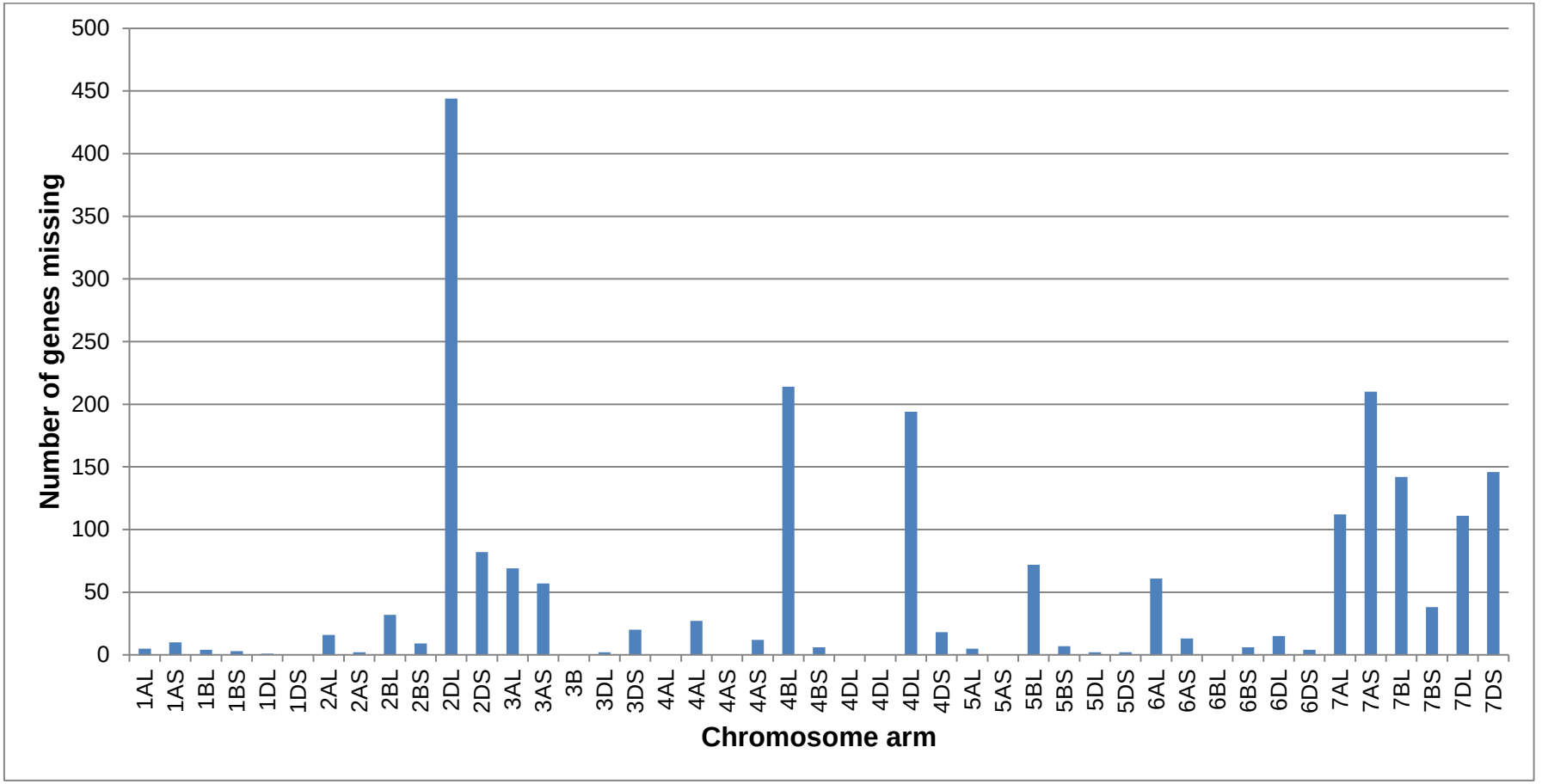


Figure 2-10. Boxplot of gene size distribution in two sets of genes: Missing, were those that could not be found in the local reassembly; Found, all other genes with a significant alignment in the local reassembly. The set of all genes was added as a reference. The genes in the Missing group were significantly smaller than their counterparts.

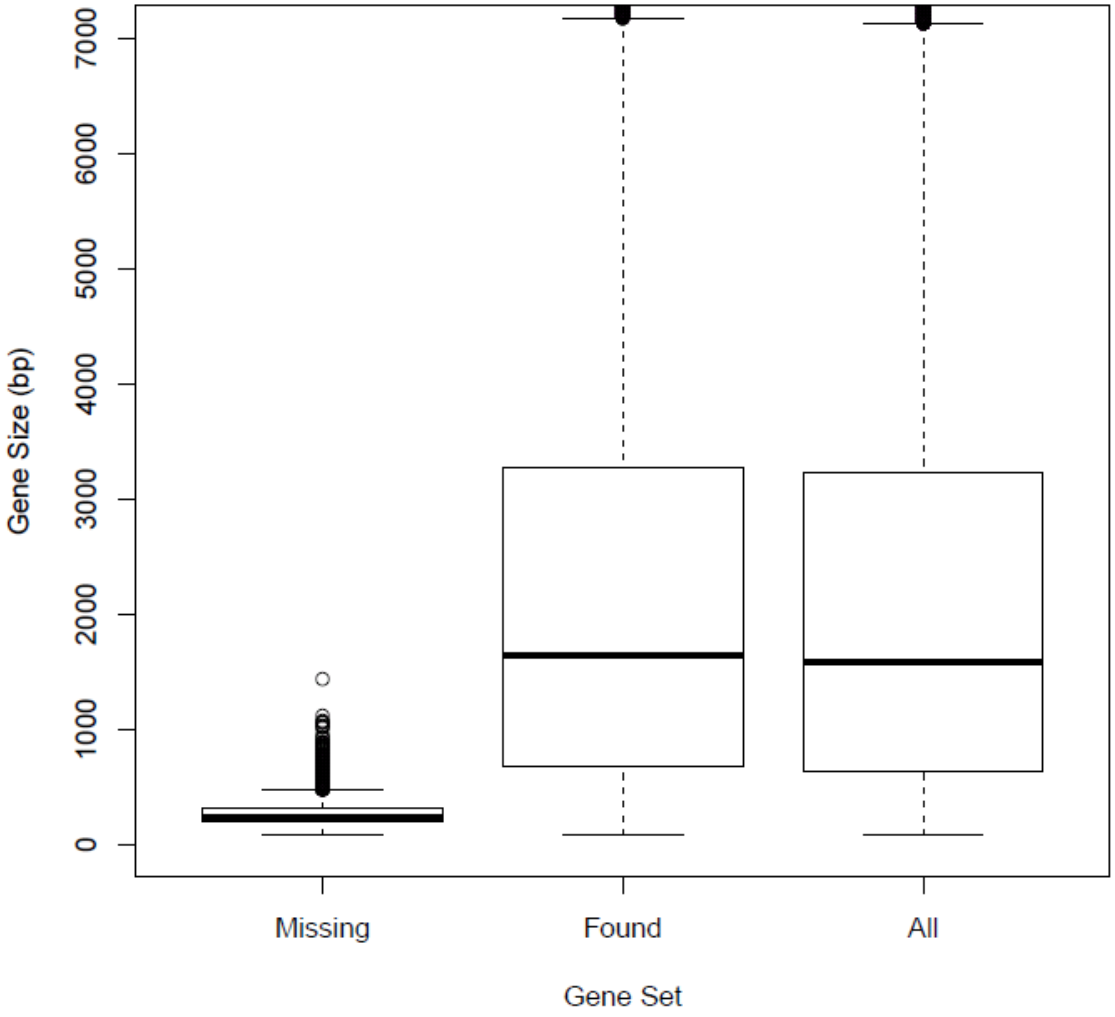


Figure 2-11. Comparison of GC content between 3 groups of genes: Missing greater than 600 bp, Missing and Found. The first group was assessed separate of the total missing genes, due to their closeness in size to the Found group in Figure 10. There is a significant difference in the GC content between both groups of missing genes and the group of genes found.

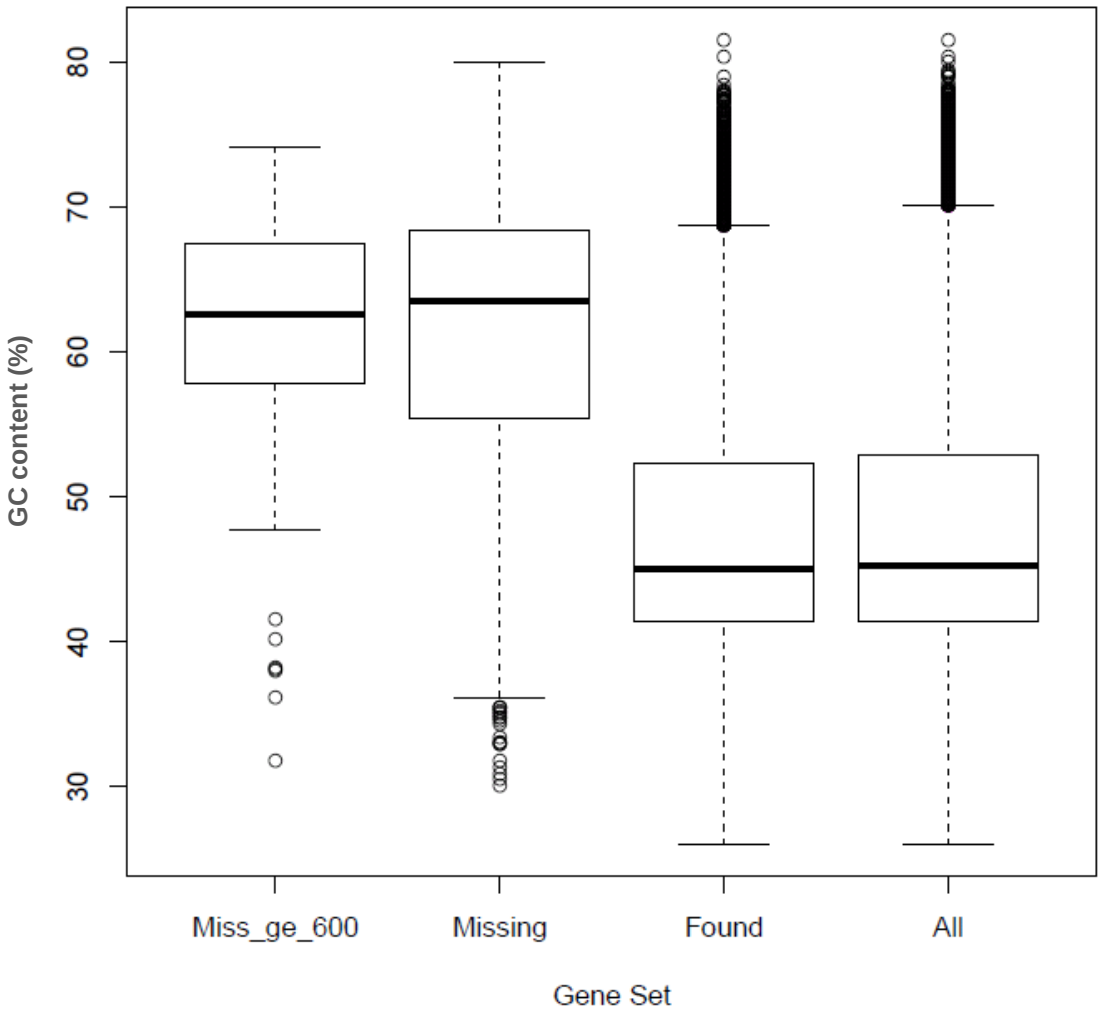
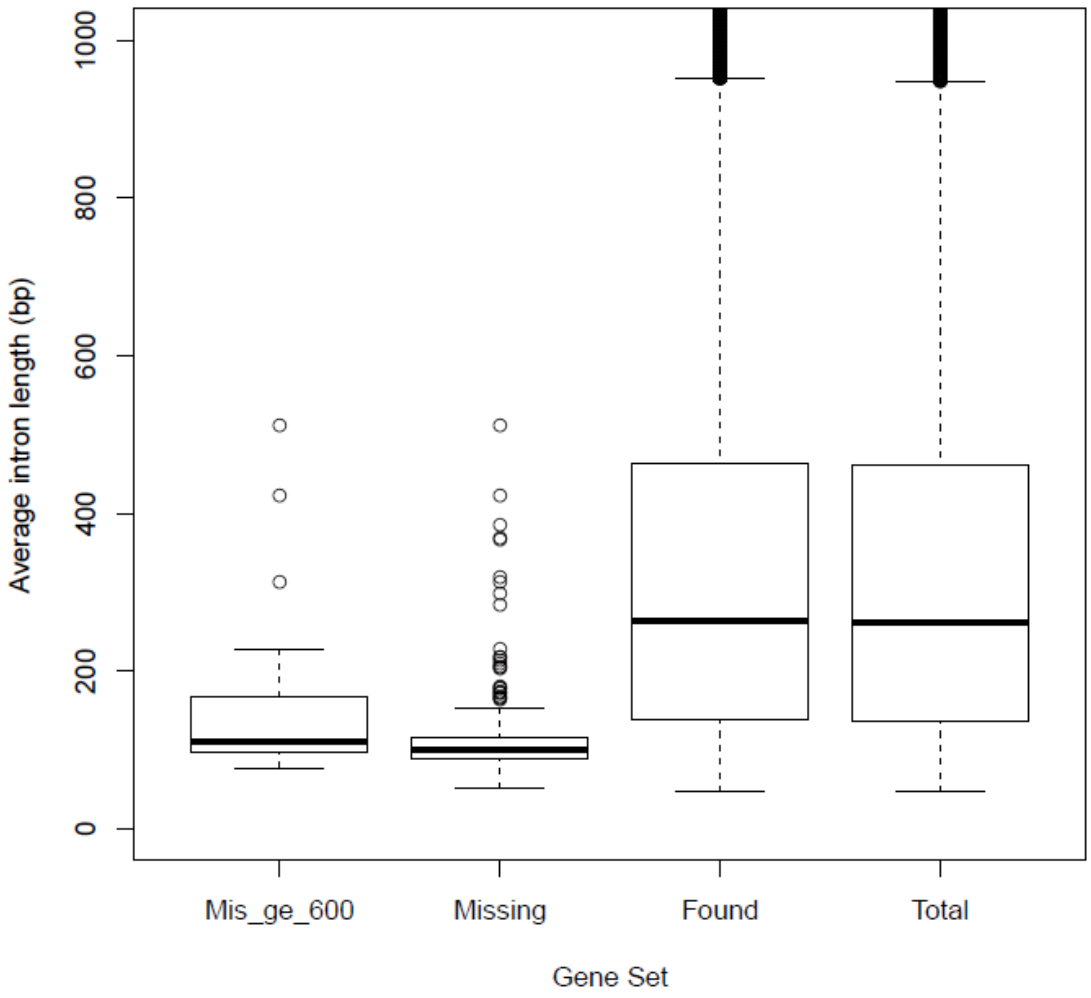


Figure 2-12. Comparison of average intron lengths. Three groups were compared: Missing greater than 600bp, Missing and Found. The complete gene set is added as reference. There is a significant difference between the values in the groups of missing genes compared to those found in the set of Found genes.



2.3.3.4 Comparison with the TGAC v1 Chinese Spring reference and the local reassembly

Collinearity between the local assembly and the TGAC assembly was evaluated by aligning the largest 100 scaffolds of the TGAC assembly to the local assembly. The results show that both assemblies are highly collinear with more than 99% sequence identity and more than 99% of each contig of the local reassembly being completely contained within one scaffold of the TGAC assembly (Figure 2-13). The TGAC scaffolds that had been assigned to a chromosome arm were preferentially aligned to contigs assembled in the same chromosome arm (Figure 2-13 and Figure 2-14).

Figure 2-13. Alignment of contigs from the local reassembly to a single scaffold of the TGAC v1 assembly. The alignments are shown as red lines delimited by red dots. There is one continuous alignment per contig. The scaffold had been placed in the 1AS chromosome arm and only contigs from chromosome arm 1AS were aligned to it. Some regions of the TGAC scaffold are not represented in the local reassembly and therefore are shown as gaps in the figure.

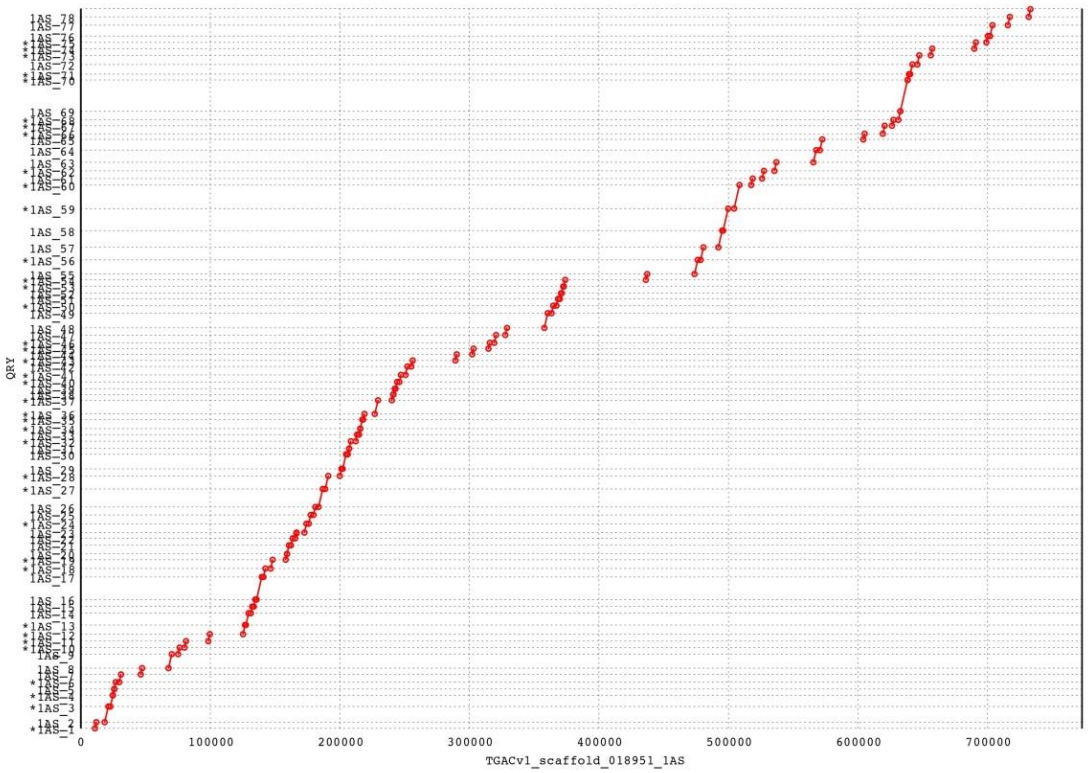
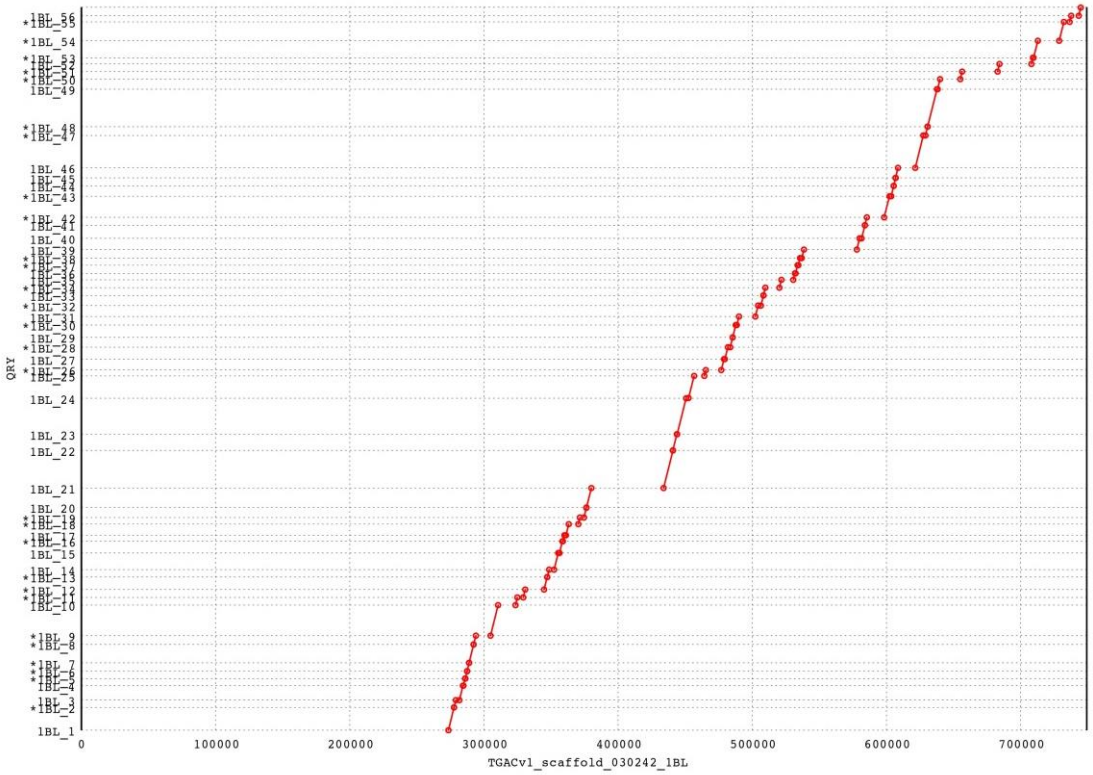
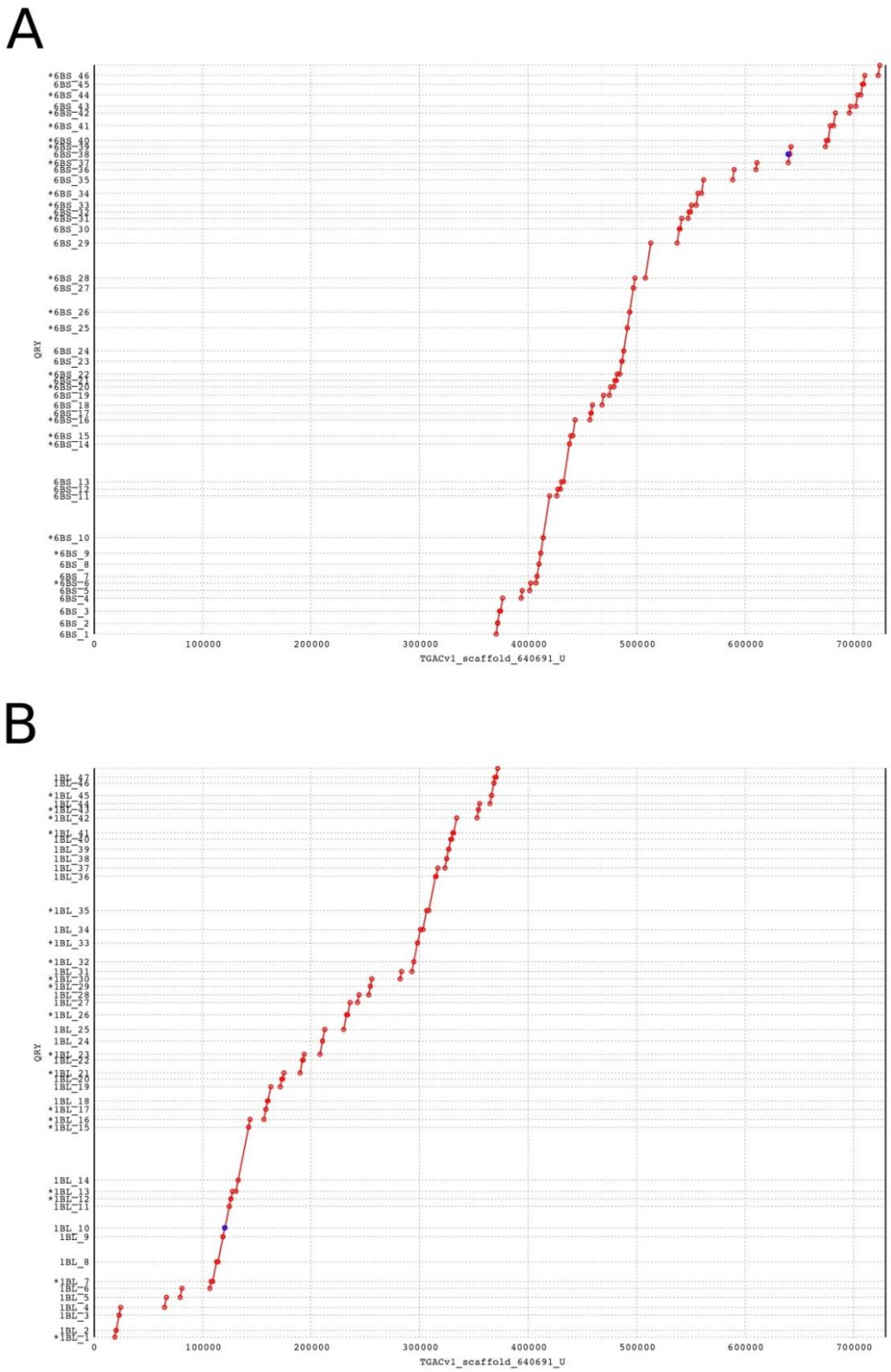


Figure 2-14. Alignment of contigs from the local reassembly to a scaffold from the TGAC v1 assembly placed in chromosome arm 1BL. As in Figure 13, only contigs from chromosme arm 1BL were aligned to this scaffold, showing concordance in the position assigned within the genome. Similarly, every contig was completely contained in the sequence of the TGAC scaffold and no two contigs overlapped each other. There is a significant gap in the initial 270 Kbp of the TGAC scaffold that corresponds to sequences that were not present in the local reassembly.



In some cases, TGAC scaffolds that had not been assigned to a position in the reference genome were aligned to two different sets of contigs from different chromosome arms. As shown in Figure 2-15, an unassigned scaffold was aligned to contigs from chromosome arms 6BS and 1BL. The alignments clearly highlight the position of the missassembly with an apparent small overlapping between contigs from the different chromosome arms near the middle of the scaffold.

Figure 2-15. Missassembly in the TGAC assembly detected by the differential enrichment of contigs from different chromosome arms to different loci of the TGAC scaffold. Contigs from chromosome arm 1BL are preferentially aligned to the 5' end of the scaffold, whereas the 3' end is aligned to contigs from chromosome arm 6BS.



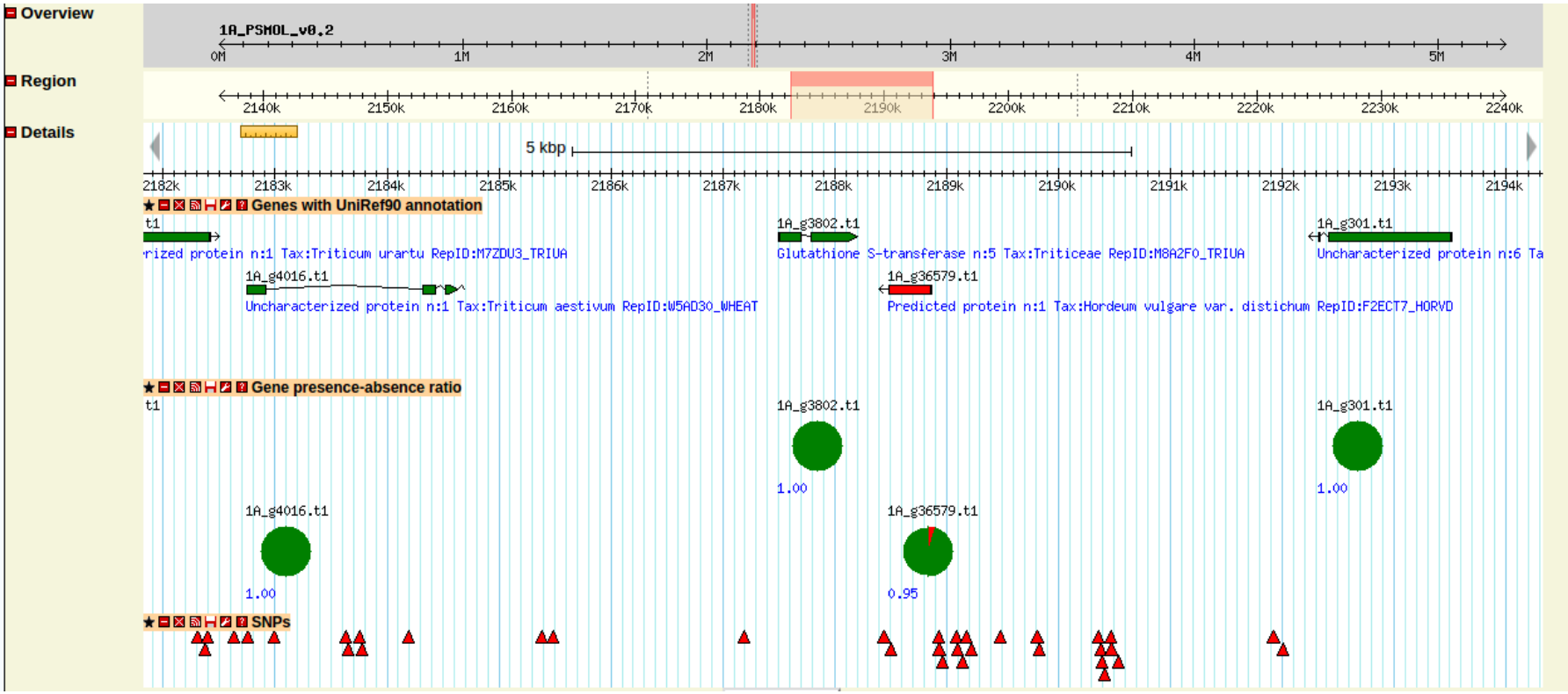
To compare the gene space of both assemblies, 155,080 protein coding gene models from the TGAC wheat genome assembly were aligned to the local assembly. In total, 151,040 genes (97.4%) were found in the local assembly. On average, every gene was found in 2.69 chromosome arms in the local assembly. In a single chromosome arm, the number of contigs aligned to any gene ranged from 1 to 10 and the non-overlapping alignments per gene ranged from 1 to 4 contigs.

2.3.4 Gene Annotation

Gene annotation was performed using all available expression data including RNA-seq produced by the IWGSC, flcDNAs and public EST data. RNA-seq data was mapped to the local wheat assembly with a mean efficiency of 75%. The accepted alignments were used to inform the prediction software about the presence and location of CDS features and exon-intron junctions. Green plant ESTs, wheat flcDNA and proteins were used as external evidence for the validation of the annotations. After filtering out genes predicted in repeat masked regions, genes with high similarity to TE-related proteins and genes without any external support (RNA-seq, EST sequence similarity, flcDNA alignment or protein similarity) a total of 139,246 genes were kept in the final annotation. The average gene size was 985 bp with an average of 2.6 exons per gene (Figure 2-16).

Closer comparison of the peptide sequences of the local gene models and the proteins encoded by close relatives (*T. urartu*, *B. distachyon* and *Ae. tauschii*) revealed that many of the putative genes in the local assembly were in fact smaller pieces of larger genes that had been annotated separately because they were in different contigs. In total, 35,339 putative genes were merged into 14,556 merged genes based on their alignments to *T. urartu* protein dataset to produce a final gene number of 118,463 of which 14,556 (12.3%) are split across different contigs.

Figure 2-16. Example of the annotation of the local assembly of the Chinese Spring genome. The graph shows pseudochromosome 1A as the first track, followed by the region selected delimited by a pink shadow. The following racks contain the genes annotated, their presence-absence status in 19 elite wheat cultivars and the position of homozygous intervarietal SNPs. This image was extracted from the wheat pangenome gbrowse (http://appliedbioinformatics.com.au/cgi-bin/gb2/gbrowse/WheatPan/).



2.4 Discussion

2.4.1 *De novo* assemblies

In this chapter, the IWGSC v2 Chinese Spring reference genome was evaluated to assess its usefulness as the basis for the construction of a wheat pangenome. Initial analysis showed that 99 genes that had been annotated in the IWGSC v2 assembly were not actually supported by raw read data from the chromosome arms where they had been found. This raised questions about the accuracy of the genome assembly and annotation. Secondly, the level of sequence duplication was measured by aligning each chromosome arm assembly to itself. The results showed that a large part of the assemblies contained long duplicated sequences of 1Kb or more which is unlikely to occur given the assembly methodology and raw sequence data used for assembly.

DeBruijn assemblers like Velvet (Zerbino and Birney, 2008, Zerbino et al., 2009) or ABySS (Simpson et al., 2009) build an assembly graph based on a fixed kmer-size, therefore, early deBruijn assemblers lost information about kmers in paired-end reads and were limited to solve repeats that were smaller than the kmer size selected (Li et al., 2010, Zerbino and Birney, 2008). Extensions to these initial algorithms managed to preserve the connections between kmers of paired reads, but these improvements offered little help for solving repeats longer than the template size of the paired-end libraries (Simpson et al., 2009, Zerbino and Birney, 2008).

Although the use of longer insert size libraries like mate-pair libraries, offer a way to offset the limitations of small insert size libraries, these were not used in the original IWGSC v2 assembly (Mayer et al., 2014). The average insert size of the libraries used was 500 bp and any repeat longer than that would have been difficult to solve if at all possible and the repeats would have collapsed into high coverage repeat contigs whose ends could not be uniquely identified. The presence of large identical sequences within a single chromosome arm assembly suggests that, at some point in the assembly, unsolved repeats were split rather than collapsed and these artificially increased the total assembly size while providing little novel information about the physical map itself.

Given these results, the decision was made to reassemble the Chinese Spring genome using publicly available reads from the flow-sorted chromosome arms produced by the IWGSC. The protocol used for assembly was different from the one used by the IWGSC but similar to that used successfully (Berkman et al., 2011b, Berkman et al.,

2012a, Berkman et al., 2013a) in the assembly of group 7 chromosome arms. Preliminary tests, to select the most adequate pre-processing steps and kmer size for assembly, were performed on chromosome arm 1DS due to its smaller size. Removal of vector leftovers and low quality stretches resulted in more contiguous assemblies with metrics similar to those obtained with similar approaches. Kmer sizes from 61 to 101 were tested and N50 metrics were compared. For chromosome arm 1DS, the longest N50 metric was obtained with a kmer size of 71. This assembly configuration produced metrics similar to those obtained by Berkman et al (2013) in the assembly and analysis of homeologous group 7. In that study, the author used a kmer size of 68 and obtained average N50 values of 2Kb and contigs as long as 50Kb per chromosome arm. Furthermore, the assemblies contained all or nearly all genes expected to be present in these chromosomes (Berkman et al., 2013a). Based on these results, kmer size of 71 was selected for assembly of all other chromosome arms.

The N50 metrics of other chromosome arms were not as large as the one obtained for 1DS (Figure 2-6) and the average number of contigs per chromosome arm were also higher than that obtained for 1DS (Figure 2-7). High N50 values are not necessarily good predictors of gene content and smaller N50 can still contain a larger fraction of the genes present in the genome (Bradnam et al., 2013a). Furthermore, by forcing larger N50 upon an assembly the number of missassemblies tend to increase (Hunt et al., 2013, Salzberg and Yorke, 2005). These missassemblies are usually caused by repeats in the target genome and should be of particular concern in genomes as repetitive as wheat (Salzberg and Yorke, 2005).

Assessment of the assembly size showed that it was equivalent to two thirds of the expected genome size (10.7 Gb, 67.6%) comprised mostly by unique sequences. In contrast, the IWGSC reference has a total length equivalent to 60% of the expected genome size (9.5Gb) and with a high amount of long duplicated sequences. Similarly, the assembly of group 7 chromosome arms resulted in assemblies with an average length equivalent to 63% of the expected size (Berkman et al., 2013a, Sehgal et al., 2012). The missing 32% may have collapsed into repetitive contigs and remains hidden from the typical assembly metrics. Having 32% of the genome as collapsed contigs suggests that at least 64% of the genome is repetitive sequence of 500 bp or longer that could not be solved by the data used during the assembly. These results are similar to early DNA renaturation studies on wheat where it was estimated that around 70% of the genome was composed of repetitive sequence of rapid reassociation kinetics (Smith, 1976).

2.4.2 Assessment of the genome assembly

2.4.2.1 Horizontal and vertical coverage

Pre-processed reads from each chromosome arm were mapped back to its specific chromosome arm assembly to assess the horizontal and vertical coverages of the assembly. *De novo* assemblies of short reads into full genomes are prone to missassemblies generated either by the nature of the genome sequence itself (repeats) or by unfiltered contamination during DNA extraction or library preparation. An example of both was reported by Ruperao et al. (2014) who found missassemblies in both chickpea reference genomes (desi and Kabuli genotypes) and large regions of the desi genome that were not present in either whole genome shotgun read data nor in chromosome isolated read data, suggesting that these were not really part of the desi physical map (Ruperao et al., 2014). Their results suggest that the missing sequences could be an artefact produced by the assembly method used or by unfiltered contamination that remained after library preparation. External contamination is usually not an issue in chromosome sorted libraries (Šafář et al., 2010) and interchromosomal contamination has been reported to have no discernible effect on the construction of physical maps in rye (Šafář et al., 2010, Bartos et al., 2008).

The mapping efficiency observed for the pre-processed data suggests that 90% of the sequence contained in the reads is present in the assembly. Nevertheless, the local assembly size (10.7Gb) represents two thirds of the estimated genome size. Assuming a uniform read sampling of each chromosome arm and no bias in their isolation prior to library preparation, we can assume that at least part of the 10% of reads that did not map to the local assembly, represent the fraction of the wheat genome that was discarded due to their small contig size (<200 bp). Also, the ratio between expected and real coverage suggests that part of the assembly has a higher than expected coverage as would happen with contigs that represent collapsed repeats. The average ratio of 1.4 suggests that at least 40 percent of the genome sequence is found on collapsed contigs in the local assembly.

Additionally, whole genome shotgun raw 454 single end reads were mapped to the local assembly and the coverage was assessed. The 454 sequencing technology is known to contain a high number of insertions and deletions, particularly in long homopolymer stretches (Huse et al., 2007, Archer et al., 2012). By mapping raw reads directly to the reference genome, a greater portion of the reads will not map due to the high number of

sequencing errors that were neither trimmed nor corrected. As a consequence, the mapping efficiency drops sharply as is evident from the mapping efficiency observed for this data (50%) with an average vertical coverage of 2X. The same data produced a horizontal coverage of 70%, which means that 30% of the assembly was not contained in the 454 reads. The large unmapped fraction of the reference cannot be explained by random error in the sampling process which would be 13.5% based on the Lander-Waterman model (Lander and Waterman, 1988).

In Figure 2-2 a non-homogeneous vertical distribution can be observed along chromosome 1D. Two regions can be observed using both datasets (454 reads and Illumina reads) around position 375 Mbp. Interestingly this position coincides with the merging point of the two assemblies 1DL and 1DS. The marked difference between both regions could be attributed to a higher number of collapsed repeats in the 1DS assembly than in the 1DL assembly. This is supported by the level of compression of the 1DS assembly compared to the 1DL. Despite the fact that both assemblies show no significant number of genes missing (Figure 2-9, 1 gene missing in 1DL and 0 missing in 1DS), the total assembly size of chromosome 1DS represents only 57% of the expected size (Šafář et al., 2010), whereas the 1DL assembly represents 75% of the expected size for this chromosome arm. This suggests that chromosome arm 1DS assembly is more compressed than the 1DL without losing any of the genes expected to be in the assembly. This compression may be related to the total content of repetitive sequence. Another source of evidence comes from the comparison of the total fraction of duplicated sequences in the IWGSC assembly between the 1DS and 1DL arms (Figure 2-8) which shows that in 1DS 6.4% of the assembly was duplicated, whereas only 2.0% was duplicated in chromosome arm 1DL. Taken together, these results explain the unusual vertical distribution observed between the two chromosome arms of 1D.

2.4.2.2 Core eukaryotic genes

A reliable method to assess the usefulness and completeness of a *de novo* assembly is the assessment of the number of functional genes found in it (Bradnam et al., 2013a). As estimates of the total gene content in the wheat genome vary greatly (Brenchley et al., 2012b, Choulet et al., 2010), the presence of core eukaryotic genes (CEGs) can be used as an estimate of the completeness of the assembly due to its high correlation with the total fraction of genes present (Parra et al., 2009). Core eukaryotic genes are a collection of 548 genes that are conserved in the genomes of most known eukaryotic organisms.

The Assemblathon paper recommended its use to assess the completeness of *de novo* assemblies in the absence of external validation sources or as a complement to those external validation procedures (Bradnam et al., 2013a). The authors of CEGMA (Bradnam et al., 2013b) have made available a subset of 248 CEGs specific for plant genomes which was screened against the local assembly. The results showed that 245 (98.7%) CEGs were present while 33 (13%) of these genes were incomplete or truncated (Figure 2-3). The fact that 13% of the genes were found as partial alignments is probably a consequence of the fragmented nature of the assembly. This result confirms that the local wheat genome assembly contains most of the genes that are expected to be in the wheat genome. A complete annotation of the local assembly is discussed later in this chapter.

A different subset of conserved genes found in plants was generated by Simão et al (2015) and they propose that this subset is a better predictor of gene repertoire completeness than the core eukaryotic gene approach (Simão et al., 2015). The BUSCO plant database was screened against the peptide sequence of the final gene annotation and found 98% of the genes as either complete or partial matches. This result is in agreement with the previous CEGMA results which found 98.7% of all plant CEGs present in the local wheat assembly. A difference between these two results is the proportion of partial or incomplete gene matches they found, whereas CEGMA found 13% of incomplete genes, BUSCO classified 26% of the genes as partial matches. Further comparison of the gene annotation to other grass proteomes show that 35,339 putative genes could be merged into 14,556 genes based on their non-overlapping alignments to the same *T. urartu* proteins, thus supporting the results obtained from CEGMA.

It is possible that these split genes were the result of unsolved repeats present in intronic sequences which kept both ends of the gene in separate contigs as is evident from the 14% of genes that were split. This unsolvable intronic repeats could be common in gene clusters along the wheat genome. Ancient whole genome duplications and direct gene duplications are known to be common in the evolutionary history of angiosperms (Wang et al., 2012). In *Arabidopsis thaliana*, it has been shown that genes involved in regulations pathways are preferentially retained after duplication (Freeling, 2009). Furthermore direct gene duplications comprise more than 10% of all Arabidopsis and rice genes (Rizzon et al., 2006) which is consistent with the hypothesis that tademly arrayed genes are the main driver behind split genes in the wheat annotation.

2.4.2.3 Comparison with the IWGSC v2 published reference

2.4.2.3.1 Comparison of assembly metrics

Comparison between the public IWGSC v2 assembly and the local assembly showed that the new assembly contained a larger fraction of the wheat genome sequence with much lower duplication levels (Figure 2-5 and Figure 2-8). DeBruijn graph assemblers cannot solve repeats longer than the insert size of the paired-end reads used in the assembly (Zerbino and Birney, 2008), these repeats are merged into a single node in the graph and then are collapsed into single contigs in the final assembly. Some protocols add an extra step of using paired-end or mate-pair data to bridge long repeats and produce scaffolds, but this approach is limited by the insert size of the libraries produced and any repeat longer than the insert size of the libraries cannot be bridge using this approach. The libraries used by the IWGSC were paired-end reads with an average insert size between 300 and 500 bp and no mate-pair libraries were included.

This raw data makes it unlikely to produce contigs of 1Kb or longer with identical sequence in a single chromosome arm. But, as shown in Figure 2-8, some chromosome arms contained over 30% of their sequence as identical duplications of 1kb or longer. In contrast, the local assembly contained very little sequence duplication with a maximum of 0.02% of the total assembly size of chromosome arms 7BS and 5BL. Overall, 7% of the total IWGSC assembly were sequence duplications of 1Kb or longer, whereas only 0.004% of the local assembly were duplications. The *de novo* assembler used by the IWGSC was AbySS (Simpson et al., 2009) which can take advantage of multiple cores to speed up the assembly process and may be responsible for the occurrence of high levels of sequence duplications. By solving the nodes in parallel, it is possible that the same node is solved by separate processes which fail to communicate and cause the assembler to produce separate contigs with the same sequence.

In order to ensure that all coding sequences predicted to be in the IWGSC assembly were also present in the local assembly, we extracted the gene sequences from the IWGSC reference and aligned them to the local assembly. We found 97.7% (91,352) of all the genes and further characterized 2172 genes that were not found in the local assembly. We assessed their length, GC content and intron lengths and discovered that these genes were smaller, had a higher GC content and contained either no introns or very small ones compared to the rest of genes (Figure 2-10, Figure 2-11 and Figure 2-12). These

characteristics set them apart from all other genes and support the idea that these could be truncated copies of other functional genes and not real ones.

2.4.2.3.2 Comparison of gene content

The presence of core eukaryotic genes (CEGs) in a new genome assembly is an important clue to determine the level of completeness of such assembly in the absence of other external evidence like RNA-seq, ESTs or flcDNAs (Bradnam et al., 2013a). As expected, the local assembly contained a higher proportion of CEGs than the IWGSC v2 assembly, but also a higher number of partial matches. This suggests that a higher number of the genes present in the assembly are split across different contigs. This was later confirmed by aligning the protein sequences of the genes annotated in the local assembly to the proteins of the *T. urartu*, *B distachyon* and *Ae. tauschii* genomes.

The number of universal single copy orthologous genes found in the local wheat assembly confirmed that 98% of them are present and around 25% were partial or incomplete. Similarly, around 25% of the CEGs found in the local assembly were found to be partial alignments. Alignments to the proteomes of close relatives revealed that 35K genes of the 139K annotated (25%) could be merged into larger genes.

The IWGSC used these assemblies along with a comprehensive set of RNA-seq libraries to predict gene models using the MIPS methodology. This resulted in nearly 100 thousand high-confidence gene models which have been used to predict the accuracy of our own assemblies. The DNA sequence of the IWGSC genes was aligned to the local assembly and 97.7% of the genes were found in our assembly. The remaining 2172 genes that could not be found were further analysed and found to be shorter, with higher GC content and fewer and shorter introns genes (Figure 2-10, Figure 2-11 and Figure 2-12). These results suggest that these genes are not actually active genes, but pseudogenes without activity.

These results confirm that the local assembly is an improvement on the IWGSC v2 reference wheat genome because it captured a higher proportion of the wheat genome, has less duplicated regions and contains more plant core eukaryotic genes.

2.4.2.3.3 Comparison with the TGAC v1 assembly

More recently, a new version of the wheat genome assembly cv. Chinese Spring was announced and made publicly available in Ensembl plants (Kersey et al., 2016). This

assembly was produced from whole genome shotgun reads of 250 bp paired end Illumina reads and a combination of mate-pair libraries with different insert sizes for scaffolding. The libraries have not been made publicly available, but the final assembly and annotation are available for download from the Ensembl plants web site (http://plants.ensembl.org/Triticum_aestivum/Info/Index).

In the early 2000s, it was considered infeasible to reconstruct the wheat genome physical map from whole genome shotgun reads (Gill et al., 2004) due to its high content of repetitive sequence distributed along the entire genome. Those estimates were correct given the sequencing technology and assembly algorithms available at the time. However, recent advances in the chemistry of Illumina sequencing machines and in the *de novo* assembly algorithms (Vyahhi et al., 2012, Bož̌a et al., 2014) have made it possible to tackle large and complex genomes like wheat using whole genome shotgun reads. The assembly was performed using the w2rapcontigger assembler which is a modification of the Discovar assembly (Weisenfeld et al., 2014) and was specifically designed for 250 bp paired-end libraries and uses two stages of assembly first with a fixed kmer size and then with a range of kmer sizes starting with 200 bp to avoid collapsing nearly identical repeats that could be true heterozygous sequences in complex genomes (BIOINFOLOGICS, 2016). TGAC assembled over 13.4 Gbp of sequence (78.8%) of the wheat genome in scaffolds 500 bp or larger.

Similar assembly statistics were reported by the NRgene company in association with the IWGSC using the trademarked *de novo* assembler De novoMagic2.0 (NRGene, 2016). The development team reports that 82% (14 Gbp) of the wheat genome has been assembled on scaffolds 5Kb or larger and 97% of the scaffolds have been placed into pseudomolecules. The assembly is currently available for IWGSC members and for signers of the Toronto agreement. The details of the assembly and the assembler have not been revealed but general details were reported during the Plant and Animal Genome Conference in early 2016 (IWGSC, 2016). The assembler requires a very specific combination of paired-end and mate-pair libraries with insert sizes ranging from 300 bp to over 12 Kbp and a combined sequencing depth of over 200X.

Given the big differences in the assembly metrics, the comparisons between this assembly and the local assembly were based on gene content and sequence collinearity. All the contigs in the local assembly were aligned to the largest 100 TGAC scaffolds with Blast and significant alignments with more than 99% of horizontal coverage and >99% of

sequence identity were further analyzed. The total length of the contigs aligned was 10.8Mb which is roughly a sixth of the total length of the scaffolds analysed (61.3Mb). The scaffolds that had been assigned to a chromosome arm in the wheat genome were preferentially aligned to the local contigs that belonged to the same chromosome arm (Figure 2-13 and Figure 2-14). The TGAC scaffolds were classified based on the amount of CSS42 reads that were mapped to its sequence. Given that CSS42 reads were the raw data used for the local assembly and for the original IWGSC assembly, this coincidence was expected.

The alignment of contigs to the TGAC scaffolds revealed the presence of chimeric scaffolds joining loci from different chromosome arms (Figure 2-15). The alignments can point with high precision to the breaking point of these scaffolds and could prove helpful to improve the current TGAC assembly.

Alignment of the TGAC genes to the local assembly revealed that 97.4% of the genes were present. This is in agreement with the results obtained from CEGMA and BUSCO which report the presence of 98% of their respective databases present in the local assembly. Based on non-overlapping sequence alignments between the contigs and the TGAC genes, we found that most of the genes are found in a single contig per locus, but some genes are split between 2-4 contigs. Similar results have been observed by comparing the protein sequences of the local assembly with the proteome of close relatives including *Triticum urartu, Aegilops tauschii, Brachypodium dystachion, Hordeum vulgare* and *Oryza sativa*.

2.4.3 Gene annotation

The total number of gene coding loci in the TGAC genome was 100,568 close to the 99,000 loci found in the IWGSC assembly. In contrast, the local assembly contained 118,463 gene loci. Previous estimates of gene content in the wheat genome ranged from 77,000 (Berkman et al., 2012a) to 150,000 (Choulet et al., 2010). The annotation of 118 thousand genes in the local assembly is within the range proposed by these estimates. Gene annotation in the IWGSC v2 assembly reported a total of 99 thousand high-confidence genes whereas the TGAC assembly reported 100 thousand genes. All the genes reported in this chapter were supported by at least two external sources of evidence including RNA-seq alignments, flcDNA alignments, and similarity to wheat ESTs, grass proteins or cDNAs. The inclusion of external evidence ensures that the genes reported in this chapter are real, although some may still be fragments of larger genes that were split

in the assembly as was shown for 35 thousand genes after aligning them to the proteome of *Triticum urartu*.

The distribution of these genes across the three subgenomes shows a higher than expected number of genes in the D genome. Previous studies had shown that the B genome contained a significantly higher number of genes than either of the other two genomes (Qi et al., 2004, IWGSC, 2014). The recent annotation of the TGAC assembly confirms this trend as does the IWGSC annotation with the A genome containing the least number of genes, closely followed by the D genome and the B genome with the most number of genes. The fewer gene numbers in the A and D genomes has been partially explained by the two rounds of polyploidization they went through compared to a single round for the D genome (Berkman et al., 2013a). Every round of hybridization and genome duplication results in the non-random loss of genes in the genomes involved (Berkman et al., 2013a, Marcussen et al., 2014, Kenan-Eichler et al., 2011, Salmon et al., 2005).

3 Chapter 3 Assembly and annotation of the wheat pangenome

3.1 Introduction

Reconstruction of a single individual genome is an important first step towards understanding the structure and evolution of a species' genome. However, since the beginning of the genomics era, it was clear that one single individual's genome could not be considered representative of an entire population, let alone a species. Small scale sequence divergence is regularly found and annotated in every organism studied and even though resequencing approaches have been very successful in the discovery of single nucleotide polymorphisms and small indels, copy number variations and their extreme representative presence-absence variations have received little attention.

Among the many differences that can be found between two individuals of the same species, single nucleotide polymorphisms (SNPs) and copy-number variants (CNVs) are the most wide-spread and useful polymorphisms. In wheat, millions of SNPs (Lorenc et al., 2012, Trick et al., 2012, Forrest et al., 2014, Wang et al., 2014, Lai et al., 2012a, Lai et al., 2015b) and thousands of simple sequence repeats (SSRs, microsatellites) (Plaschke et al., 1995, Lelley et al., 2000, Ishii et al., 2001, Somers et al., 2004, Sourdille et al., 2004) have been identified and used to produce high throughput genotyping methods that helped understand wheat evolution and to analyse its genetic diversity, providing valuable resources for genetic breeding (Dvořák et al., 1993, Martin et al., 1995, Plaschke et al., 1995, Lelley et al., 2000). These markers have also been used in the construction of genetic maps that assist in the map-based isolation of genes of agronomic importance (Poland et al., 2012b, Gill et al., 1996, Röder et al., 1998, Stephenson et al., 1998, Somers et al., 2004).

Among genomic variations CNVs have been the least studied, mostly due to the lack of the technology necessary to identify them efficiently. Even after the development of second generation sequencing (SGS) technologies, most genomic studies focused on the discovery of SNPs. However, evidence keeps accumulating revealing the importance of CNVs in the phenotypic plasticity and adaptability of varieties to different environments. Presence-absence variation (PAVs) are an extreme form of CNVs where the sequence is completely missing in one individual and present in another. Genes are also subject to

such variations and can be found in one individual while being absent in another of the same species. These gene PAVs have recently been associated with heterosis in crop plants (Springer et al., 2009, Swanson-Wagner et al., 2010, Kaeppler, 2012).

The origins of these gene presence-absence variations are still unclear. Genome duplication via interspecific hybridization usually results in reproducible patterns of gene loss within the first generations after the appearance of the amphipolyploid, although some extent of differential gene loss can be identified. This rapid gene loss has been observed in studies of newly synthesized allopolyploids of different plant species including wheat (Smet et al., 2013, Schnable et al., 2011, Kashkush et al., 2002, Wendel and Doyle, 2005, Adams and Wendel, 2005). Following the stabilization of the genome, a process of diploidization takes place and affects genome evolution by allowing greater freedom of mutation in duplicated genes. This often results in preferential neo-functionalization or sub-functionalization of one of the copies of duplicated genes, which in turn increases the differentiation between homeologous chromosomes that is crucial for diploidization (Tate et al., 2009, Lukens et al., 2006b, Prince and Pickett, 2002, Irish and Litt, 2005). Thus evolutionary processes can explain the differential gene content between individuals of the same species and have been studied using comparative genomics approaches.

Comparative studies of gene content between isolates of the pathogenic *Streptococcus agalactiae* showed that around 20% of the genes were absent in at least one of the isolates, while the remaining 80% was present in all samples analysed (Tettelin et al., 2005, Medini et al., 2005). Mathematical analysis of the gene content increase as a function of the number of genomes included revealed that some species had an upper limit to the number of genes present in their genepool that was estimated by the asymptote of the regression curve, while others lacked such limit and could apparently contain an infinite number of novel genes in their genepool. These cases are referred to as closed and open genomes (Tettelin et al., 2008, Lapierre and Gogarten, 2009, Bentley, 2009) depending on the gene number limit estimated by regression ("*closed"* means there is a limit and "*open"* means there is no such limit). These studies defined the pangenome as the sum of all the genes that would be found if all individuals of a clade (more commonly a single species) were sequenced and annotated. The first studies were performed in bacteria because genome assembly an annotation was easier and cheaper and there was an enormous wealth of unexploited genomic data already available.

Pangenomic studies in higher organisms can trace their origins to the comparative genomic analysis between assembled genomes of different species. The use of molecular markers for the study of population genomic structure and genome evolution required large quantities of genotypic data from multiple individuals of different races. In humans, not long after the publication of the first genome, a plan was being devised for the sequencing of 1000 human genomes and the establishment of a database of human genomic variation (The Genomes Project, 2015, Iafrate et al., 2004, Feuk et al., 2006). Similar initiatives were proposed for other model organisms including Arabidopsis (Alonso-Blanco et al.), mouse (Keane et al., 2011) and the fruit fly (Wang et al., 2015). As a result of the many whole genome resequencing projects and large scale production of genomic data, evidence started to accumulate showing the extent of CNVs in every organism studied. However, it is only recently that large scale pangenomic studies of plants have been undertaken. The 1001 Arabidopsis genome project (Alonso-Blanco et al.), the rice pangenome assembly and gbrowse (Sun et al., 2017, Yao et al., 2015), the maize pan-transcriptome and pangenome analysis (Jin et al., 2016, Lu et al., 2015, Hirsch et al., 2014), the construction of soybean pangenome draft (Li et al., 2014c) and the recent construction and analysis of the *Brassica oleracea* pangenome (Brassica genome C) (Golikz, 2016, Golicz et al., 2016b) are all the first attempts at unveiling the hidden genetic diversity of crop plant gene pools.

These first studies have shown the most plant species contain a large core genome that comprises between 60-80% of all the genes in the pangenome. Maize, on the other hand shows a smaller core genome and seems to be more prone to accumulating structural variants that could account for up to 50% of sequence divergence in some loci (Hirsch et al., 2014, Swanson-Wagner et al., 2010, Lai et al., 2010, Eichten et al., 2011). Analysis of the variable genes revealed that they are enriched with genes involved in the defence response, response to environmental stress, and intracellular signalling pathways. These results highlight the importance of the variable genome in individual adaptation to local environment and pathogens. Mathematical modelling of pangenome expansion has shown that all these plant species have a closed pangenome, but agree that the inclusion of more distant varieties, landraces or synthetic varieties could harbour yet unexplored sequence variants that may further increase the gene content of the pangenome in each of the species.

The re-assembly of the Chinese Spring genome described in the previous chapter, though an important step in wheat genomic studies here is only considered as the first

step towards the construction of a complete wheat pangenome. Chinese Spring has been an important source of cytogenetic stocks mainly due to its readily crossability with rye and the production of aneuploid and chromosomal deletions clones that were essential in the early days of wheat cytogenetic studies (Sears and Miller, 1985, O'mara, 1953, O'Mara, 1951, Sears, 1969). Nevertheless, it has not been widely used in breeding programs, mostly because of its high susceptibility to pathogens and lack of traits of agronomical importance (Sears and Miller, 1985). Chinese Spring is the most studied wheat cultivar with plenty of cytogenetic resources that have been used from gene identification and isolation to chromosome-based genome assembly using Chinese Spring derived ditelosomic lines (IWGSC, 2014). However, its absence in the pedigree of most of the modern elite wheat cultivars and from parental lines in wheat breeding programs limit the use of its genome sequence in breeding programs. That is why, the extension of Chinese Spring reference genome with new genes that are absent in its sequence is essential to increase the usefulness of wheat genomic data.

In this chapter, the Chinese Spring reference genome is expanded to include additional sequences contained in all wheat cultivars for which there is enough sequence data available. To do this, sequences that were not found in the Chinse Spring reference genome, but were present in the 19 cultivars analysed, were assembled and annotated. This analysis permitted the assignment of annotated genes to either the core or variable genomes of wheat, the estimation of the approximate gene content in the wheat pangenome and the functional characterization of its variable genome. This can be considered the first draft of the wheat pangenome and is a valuable resource for the identification of genes of agronomic importance that cannot be found in a single reference genome.

3.2 Materials and Methods

3.2.1 Raw data

Whole genome shotgun reads from 16 wheat Australian cultivars were downloaded from the Bioplatforms diversity sequencing set (https://downloads.bioplatforms.com/wheat_cultivars/samples). Also, WGS reads from cultivars OpataM85 and synthetic W7984 and their double haploid offspring were downloaded from the sequence read archive (SRA) at NCBI. RNA-seq reads from 9 of the 16 wheat cultivar were downloaded from the NCBI SRA database (Appendix I).

3.2.2 Construction of the wheat pangenome

3.2.2.1 De novo assembly of unmapped reads

In order to extend the sequence of the wheat genome and include sequences from all available sequenced cultivars, we mapped the raw reads of the cultivars to the complete Chinese Spring reference genome and extracted those reads that could not be mapped to it.

The complete Chinese Spring reference genome was constructed by adding the previously assembled group 7 chromosomes (Berkman et al., 2013a), the chromosome 3B (Paux et al., 2006) and both the chloroplast (Middleton et al., 2014) and the mitochondrial (Cui et al., 2009) genomes of wheat (Genbank: KJ614396.1 and AP008982., respectively) to the genome assembly of groups 1 to 6 shown in the previous chapter.

Raw reads from the 16 Australia wheat cultivars were mapped to the reference wheat genome with Bowtie2 v2.2.9 (Langmead and Salzberg, 2012) allowing an insert size from 0 to 1000 bp. Libraries with mapping efficiencies below 60% were not included in the analysis. Paired-reads that could not be mapped to the reference genome were extracted with samtools v1.3 (Li et al., 2009a) using the samtools view command –f 0x4. Unmapped reads were evaluated using FastQC (Andrews) (k=7). Trimmomatic v0.32 (Bolger et al., 2014) was used to remove low quality stretches and adapter leftovers using a sliding window approach of 6 bp with less than 20 base quality scores. Trimmed reads smaller than 73 bp and unpaired reads were no longer considered for further analysis. Paired-end reads were assembled with IDBA-UD (Peng et al., 2012) using default parameters.

After removing contaminant sequences, a second reference genome was produced by adding the selected scaffolds to the previous reference genome. Subsequently, raw reads from the cultivars OpataM85, W7984 and 90 double-haploid offspring were mapped to the new reference genome as described above, and unmapped read pairs were extracted, trimmed and pooled for assembly with IDBA-UD.

3.2.2.2 Contamination removal

The *de novo* scaffolds were aligned to NCBI nucleotide database (NT database) with Blastn v 2.2.30+ (Camacho et al., 2009) using a minimum e-value of 1e-10 (-e-value 1e-10). Blast results were parsed into tabular format with the Bioperl module Bio::SearchIO and an in house script parseBlast.pl. Only the best hit for each scaffold was considered

and scaffolds with a significant top hit outside of the green plants clade (Viridiplantae, Taxon ID 33090) were excluded from the assembly and considered contamination. Classification of the top 100 genus and species hits in the Blast results was done using the NCBI taxonomy database (ftp://ftp.ncbi.nih.gov/pub/taxonomy/).

3.2.3 Comparison of mapping efficiency to the reference genome and to the pangenome

As an initial validation of the assembled sequences, the mapping efficiency of the raw reads to the reference genome and to the extended genome (pangenome) was compared. Reads from each cultivar were mapped to the reference genome and to the pangenome with Bowtie2 v 2.2.9 (Langmead and Salzberg, 2012) with standard parameters and mapping efficiencies, and the results were were plotted for each cultivar. Chinese Spring chromosome sorted reads were included as a control to confirm the absence of the additional scaffolds in the raw sequence of Chinese Spring.

3.2.4 Placement of unmapped scaffolds in the Chinese Spring reference genome

3.2.4.1 Placement based on read pair information

Reads from the 16 wheat cultivars were mapped to the pangenome and reads that mapped to the last 300 bp of either end of the unmapped scaffolds were used to anchor the scaffolds to Chinese Spring contigs. At least two anchor reads were required for placing the scaffolds and at least 80% of the reads should place the scaffolds to the same position in the reference genome.

3.2.5 Gene Annotation

RNA-seq reads from 9 of the 16 wheat cultivars were mapped to the pangenome assembly with HiSat2 v 2.0.4 (Kim et al., 2015) using default parameters. Accepted alignments were transformed into hints with the program bam2hints from the Augustus package. In parallel, the unmapped assembly was aligned to the green plants ESTs database from NCBI using tBlastx. Significant alignments (eval ≤ 1e-5) were also transformed into hint files using Bioperl (Stajich et al., 2002).

Augustus (Stanke and Morgenstern, 2005) was used for genome annotation of the unmapped assemblies using RNA-seq and EST alignments as hints for gene structure. Gene models were further selected based on their size (≥ 300bp) and sequence similarity to known EST or proteins in the green plants sequence dabatase from NCBI. Finally,

genes with high similarity to transposable elements or that overlapped masked repeats were removed from the final annotation.

3.2.6 Presence-absence variation of genes

Genes were identified as either present or absent in every cultivar analysed based on the alignment of their raw reads to the exonic features of individual genes. The protocol used was a modification of the SGSGeneLoss pipeline (Golicz et al., 2015b): raw reads were mapped to the reference pangenome with Bowtie2 v 2.2.9 using standard parameters. Read coverage per gene per cultivar was obtained using the command samtools depth from Samtools v1.3 package. Genes were identified as present if they comply with two conditions: 1) exon coverage was at least 5% of the entire exonic sequence of the gene and 2) per base read coverage in the exonic regions was at least 2. Genes that were annotated in the assemblies but found to be absent from all the cultivars involved were removed from further analysis.

3.2.6.1 Validation of gene presence-absence variation

RNA-seq reads from 11 wheat cultivars were mapped to the wheat pangenome with HiSat2 (Kim et al., 2015). The horizontal coverage of exonic sequence per gene was measured using the bed annotation file and the command samtools depth –b <bed_annotation>. The exonic horizontal coverage was measured as the total number of bases in exons covered by 2 or more reads as a fraction of the total length of all exons in the gene. The average exonic horizontal coverage of genes predicted to be present was measured and used as a threshold to assess the present or absent status of genes predicted to be absent according to WGS data. The number of genes predicted to be absent from WGS data, but present according to the RNA-seq data was measured.

3.2.7 Pangenome modelling

PanGP (Zhao et al., 2014) was used to count the total number of genes in the core and variable genome for all possible combinations of cultivars from 1 to 19 cultivars. The averages of these counts were used to model the expansion of the pangenome and the contraction of the core genome using non-linear models. For pangenome expansion, a power law model ($f_{(x)} = Ax^B+C$) was used, whereas an exponential model ($f(x) = Ae^{Bx}+C$) was used to fit the core genome contraction (Tettelin et al., 2005). Both models were fitted in R (Suzuki and Shimodaira, 2006) with the nls function.

Genes that were present in every cultivar were considered to be part of the wheat core genome, whereas the rest were considered the variable genome.

3.2.8 Functional enrichment of the wheat variable genome

Translated protein sequences of all annotated genes were aligned to the *Arabidopsis thaliana* proteome database (ftp://ftp.arabidopsis.org//Sequences/Blast_datasets/other_datasets/CURRENT/At_GB_refseq_prot.gz) using BlastP with standard parameters. Functional annotation was then performed using the command line version of Blast2GO v2.5 (Conesa and Götz, 2008) with standard parameters. The TopGO package (Adrian Alexa, 2006) from Bioconductor was used to determine functional enrichment of biological processes in the variable genome using a Fisher exact test ($p \leq 0.01$) and the full genome annotation as background.

3.3 Results

3.3.1 Assembly of unmapped reads

Whole genome shotgun sequencing reads of 19 wheat elite cultivars were used to identify and assemble regions that were present in any of these 19 cultivars but absent from the Chinese Spring reference genome constructed in the previous chapter. We used a two-step approach where the unmapped reads from 16 wheat elite cultivars were pooled and assembled to extend the reference sequence and, in a second step, the SynOpDH population (Sorrells et al., 2011) and their parents were mapped to the extended reference genome, unmapped reads were isolated, pooled and assembled.

In the first step, whole genome shotgun reads from 16 elite wheat cultivars were mapped to the reference genome of Chinese Spring (Figure 3-3). On average, the mapping efficiency of the libraries was 83% with the exception of 5 libraries from the cultivar Baxter whose mapping efficiency was only ~30%. Aggressive quality clipping of these libraries to remove low quality sequence stretches, adapter sequences and overrepresented kmers did not result in higher mapping efficiency. Finally, 1,000 random reads from each library were compared to the non-redundant nucleotide database from the National Centre for Biotechnology Information (NCBI). The results showed that these libraries contained mostly bacterial and fungal sequences and were therefore excluded from further analysis (Appendix II).

Unmapped reads from the 16 cultivars were extracted and assessed for sequence quality, adapter content and overrepresented kmers. Then, reads were quality trimmed, PCR adapters were removed and reads smaller than 73bp were discarded. In total, 870 million paired reads were used for assembly. The reads from the 16 cultivars were pooled to compensate for the low coverage per cultivar which was estimated to range from 8X to 20X and was insufficient for *de novo* assembly of individual samples (Table 3-1).

Table 3-1. Total number of bases and sequencing depth of the cultivars used in this Thesis

Year	Cultivar	Experiment	Read Length (bp)	Insert Size (bp)	Total bases (Gbp)	Sequencing depth
2016	ABC-1	WGS	100	300	166.045	09.77
2016	Alsen	WGS	100	280	189.837	11.17
2016	BX-1	WGS	100	320	196.977	11.59
2016	CH7	WGS	100	200	251.540	14.80
2016	Drysdale	WGS	100	300	173.100	10.18
2016	Excalibur	WGS	100	300	161.200	09.48
2016	Gladius	WGS	100	300	184.000	10.87
2016	H45	WGS	100	430	171.900	10.12
2016	Kukri	WGS	100	300	247.000	14.53
2016	OpataM85	WGS	216	680	195.600	11.51
2016	Pastor	WGS	100	300	214.120	12.60
2016	RAC	WGS	100	300	166.500	09.79
2016	Volcani	WGS	100	300	168.800	09.93
2016	W7984	WGS	150	400	340.700	20.04

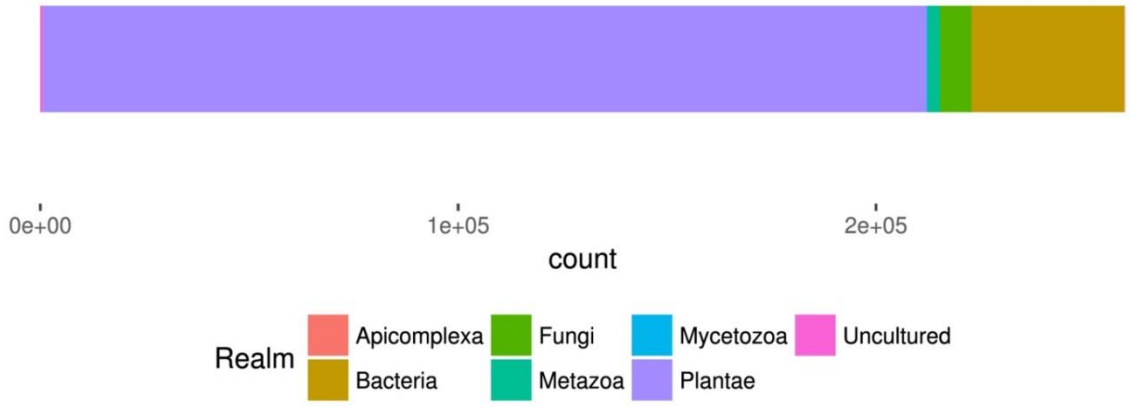
2016	Westonia	WGS	100	260	142.479	08.38
2016	Wyalkatchem	WGS	100	350	338.575	19.92
2016	Xi-1	WGS	100	140	243.672	14.33
2016	Yp-1	WGS	100	150	222.288	13.08

Following the same approach, raw reads from the cultivars OpataM85, W7984 and 90 F1-derived double haploid individuals from the SynOpDH population (Sorrells et al., 2011) were mapped to the extended reference genome and the unmapped paired reads were extracted, pre-processed, pooled and assembled. Four libraries from cultivar W7984 with mapping efficiency below 60% were not used in the assembly. The unmapped reads from the SynOpDH population was pooled with the unmapped reads of both parents to compensate for the low sequencing depth of each cultivar. This raised the average sequencing depth to ~60X. In total, 7 million paired-end reads were used for assembly. As in the previous step, the raw assembly was compared to the NT database from NCBI to identify and remove contaminant sequences.

3.3.2 Contamination identification and removal

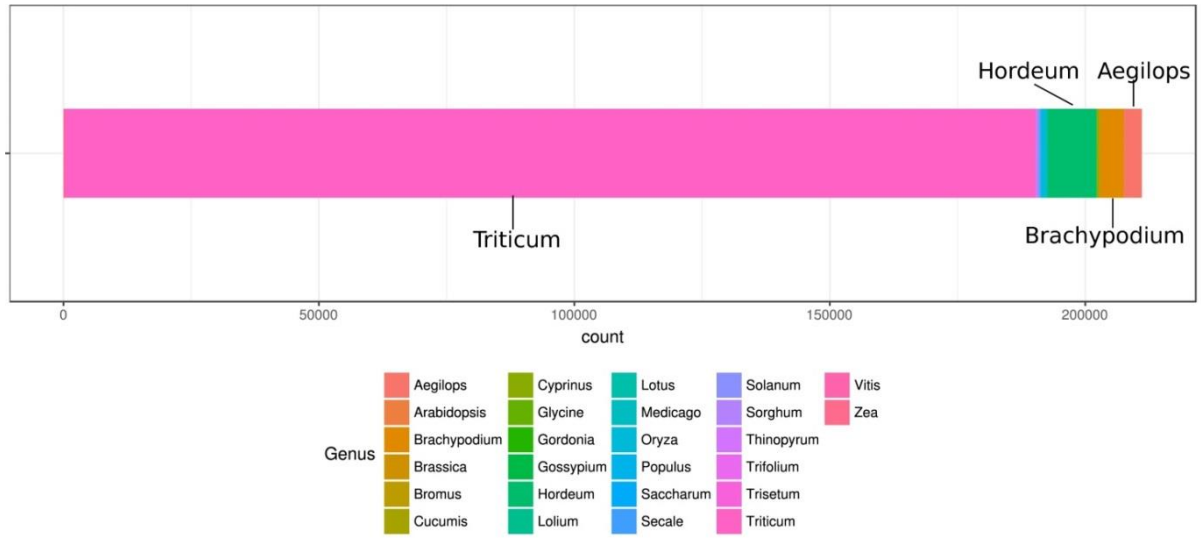
Combined, both assemblies produced a total of 659.7 Mb (659,703,067bp) of raw sequence in 328,783 scaffolds. In order to remove contaminant sequences, the raw assembly was aligned to the nucleotide database at NCBI (NT) and all the scaffolds whose top alignment was outside of the green plants clade (Viridiplantae - NCBI Taxon ID: 33090) were removed from the assembly. Scaffolds with no hits in the NT database were further compared to the genome survey sequence database (GSS) and genomic Reference sequence database (Genome). As shown in Figure 3-1, Plantae was the group with the largest number of best hits to the raw scaffolds representing 60% of all the best hits. Another 4% of the scaffolds did not have any hits to any database and were also kept for further analysis. The remaining 36% of the scaffolds with top hits outside of the green plants group were no longer considered in this study.

Figure 3-1 Source of the best Blast hits for the raw scaffolds. The graph includes the top 100 most frequent genus hits which appear in over 250,000 scaffolds. The most frequent Blast hit was with group Plantae, followed by Bacteria, Fungi and Metazoa.



The contribution of each genus annotation to the final scaffold selection is shown in Figure 3-2. The majority of selected scaffolds have a significant sequence identity to other sequences in the *Triticum* genus, followed by the *Hordeum, Brachypodium* and *Aegilops* genera. Within the *Triticum* genus, *Triticum urartu, Triticum durum, Triticum turgidum* and *Triticum aestivum* were the most frequent Blast hits, with the latter as the most common one. The average sequence similarity between the additional scaffolds and the Plantae hits was 88.4% with an average alignment length of 543bp, suggesting that these sequences were close enough to be recognised, but different enough to be considered novel. These results do not allow us to speculate as to the origin of these sequences.

Figure 3-2. Number of scaffolds with best hits to group Plantae. The family Poacea is the most frequently represented in the final selection with the Triticum, Hordeum, Brachypodium and Aegilops as the most frequent representatives.



After removal of contaminant sequences, the statistics of both assemblies are shown in Table 3-2. The assembly of unmapped reads from the 16 wheat cultivar contained 343 Mb of additional sequence (343,277,182 bp) in 210,792 scaffolds. The average scaffold length was 1,629 bp and the N50 was 1,830 bp The second assembly iteration using unmapped reads from the cultivars OpataM85, W7984 and their 90 double haploid offspring contained 6.7 Mb (6,722,169 bp) of additional sequence in 11,199 scaffolds with an average length of 600 bp and an N50 of 960 bp.

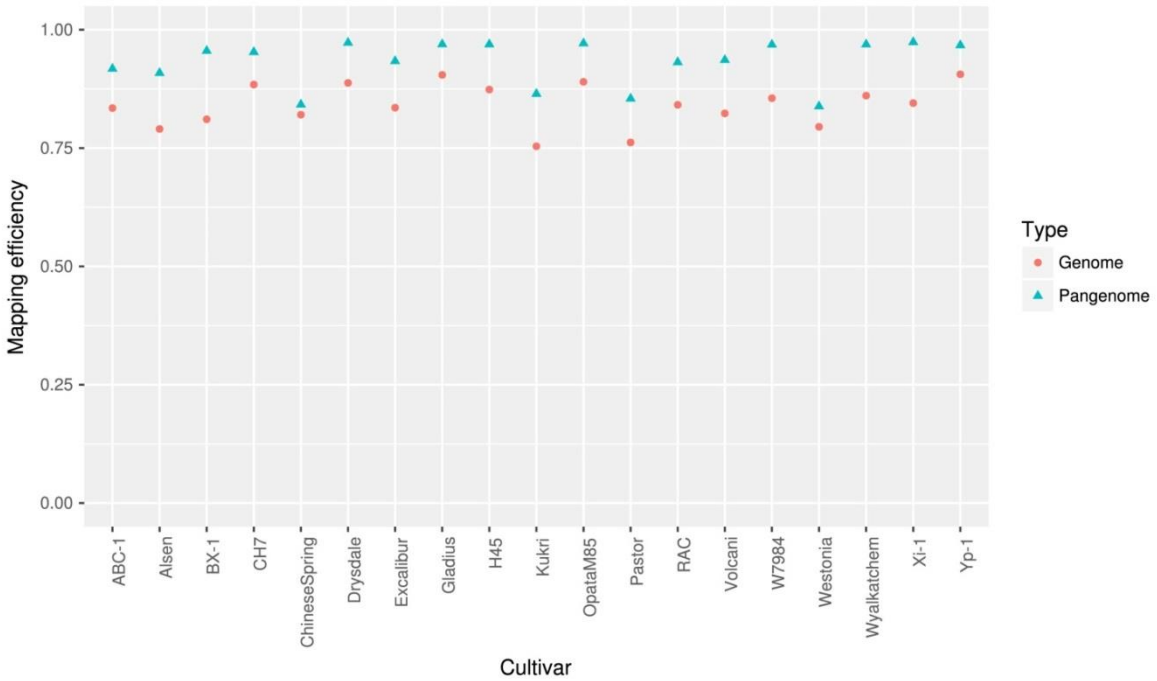
Table 3-2. Assembly metrics of the unmapped reads. The Bioplatforms subset includes 16 Australian wheat cultivars. The OpataM85-W7984 assembly includes the 90 double-haploid individuals of the SynOpDH family, and the parental cultivars OpataM85 and W7984.

Assembly	**Total bases**	**Avg size (bp)**	**N50 (bp)**	**Biggest Contig (bp)**
BioPlatforms	343,277,182	1,629	1,830	97,138
OpataM85 W7984	6,722,169	600	960	12,975

3.3.3 Validation of the assembly

Reads from the 19 cultivars (18 cultivars + Chinese Spring CSS libraries) were mapped to the wheat pangenome and the mapping efficiency was obtained for each cultivar. Comparison with the mapping efficiency obtained using the Chinese Spring reference genome shows that on average 9% more reads were mapped to the pangenome than to the Chinese Spring genome with the exception of Chinese Spring itself, which only increased its mapping efficiency by 2%. This increase in the mapping efficiency for all libraries suggests that the pangenome offers mapping space that is unavailable in the Chinese Spring reference. Furthermore, the small increase in mapping efficiency found for Chinese Spring in comparison with all other cultivars shows that this additional mapping space is barely used by Chinese Spring CSS reads, confirming that this most of this additional sequence was not present in the Chinese Spring genome.

Figure 3-3. Comparison of mapping efficiency for all cultivars between the pangenome and the Chinese Spring genome. Red circles are the mapping efficiency against the Chinese Spring assembly; blue triangles represent the mapping efficiency against the wheat pangenome assembly.



3.3.4 Placement of unmapped scaffolds into the reference Chinese Spring reference genome

Two approaches were used to place newly assembled scaffolds into the reference pangenome: paired-end information and a genetic map (see Chapter 4). The first approach allowed us to anchor 30.2% (67,024) of the newly assembled scaffolds to specific chromosome arms. In most cases, however, the scaffolds were anchored to contigs that had not been placed in the pseudomolecules. Table 3-3 shows the distribution of the unmapped scaffolds across the wheat pangenome. Fewer scaffolds were placed in the D genome compared to the A and B genomes whereas more scaffolds were placed in the homeologous group 7 than in the other homeologous groups mostly driven by an increase in chromosome 7B.

Table 3-3. Distribution of new scaffolds assigned to a chromosome by mate-pair information

	A	**B**	**D**	**Total**
1	1290	1667	1405	4362
2	1828	2157	4158	8143
3	7453	1953	2186	11592
4	2116	2902	1899	6917
5	3399	2651	1427	7477
6	2665	3842	1264	7771
7	5693	9535	5534	20762
Total	24444	24707	17873	**67024**

3.3.5 Gene annotation and clustering

Augustus was used for gene annotation using RNA-seq (Wang et al., 2014) and alignments to green plant ESTs as hints for gene prediction. Gene models were filtered based on their size (>100bp), support from either RNA-seq or ESTs, their overlap to

repeat-masked regions and their similarity to TE-related proteins. To avoid overestimation of the number of genes in the unmapped reads, only genes with sequence identity to other genes from green plant members were considered for further analysis. The final gene set contained 21,653 gene models with an average gene length of 950 bp. Along with the 118,463 genes present in the Chinese Spring reference genome, a total of 139,747 genes have been identified in the complete wheat pangenome.

The genes were clustered based on protein sequence similarity with the proteomes of *Arabidopsis thaliana*, *Triticum urartu*, *Hordeum vulgare*, *Brachypodium dystachyon*, *Sorghum bicolor*, *Setaria italica* and *Aegilops tauschii*. In total 45,779 clusters were found with an average of 7.2 genes per cluster and a median of 4 genes.

3.3.6 Gene presence-absence variation

On average, each cultivar contains 128,656 genes ranging from 118,288 genes in Chinese Spring to 132,445 in Xiaoyan-54 (Xi-1). In total 49,952 genes (35.7%) show presence-absence variation in the 19 cultivars and represent the variable genome of wheat. The remaining 89,795 genes (64.3%) were present in all cultivars and represent the wheat core genome. Baxter-1 (BX-1) contains the highest number of unique genes (712) whereas ABC-1 and Kukri both had 0 unique genes. The average number of genes missing per cultivar was 11,091 with Chinese Spring being the one with the most genes lost (21,459) and Xiaoyan-54 the one with fewest genes absent (7,477). Mapping of reads from Chinese Spring isolated chromosome arms to the pangenome assembly showed that none of the additional annotated genes were present in the Chinese Spring genome.

The presence of orthologous and paralogous genes in the wheat genome creates a redundancy of functions that can result in neo-functionalisation or specialisation of one or all of the genes clustered. Thus, core functions may be present even if one member of the family members is missing from the individual. To account for this, 31,433 wheat gene clusters were evaluated for presence absence variation. From these, 26,014 gene clusters were present in all cultivars and represent the core functions of the wheat genome whereas the remaining 5,419 clusters are present only in a subset of the cultivars analysed and their function is not considered essential.

3.3.6.1 Validation of presence-absence variation

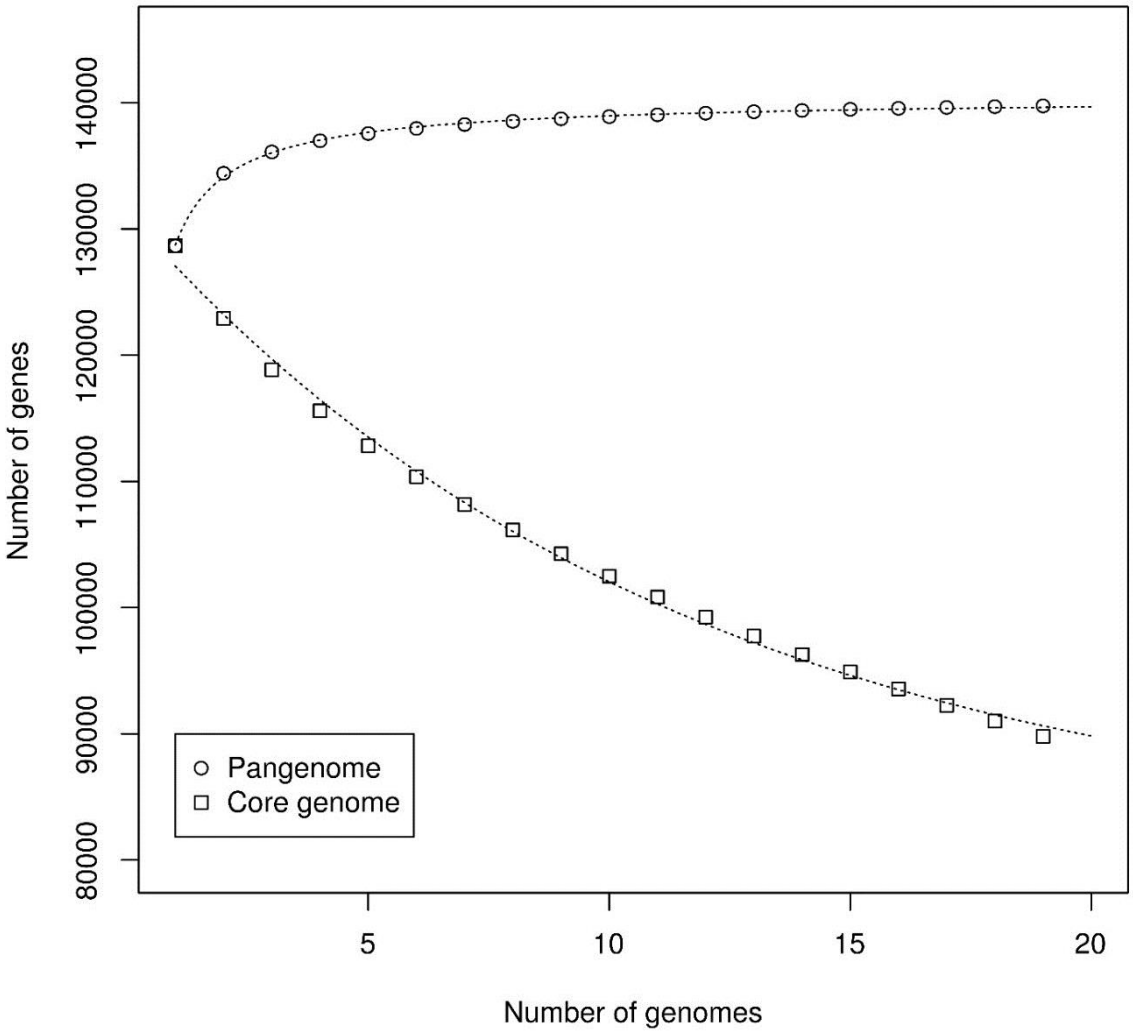
RNA-seq from 11 cultivars was used to confirm the missing status of the genes predicted to be absent by the PAV pipeline. The average mapping efficiency was 75% for

the eleven libraries used. For every gene in the pangenome, the fraction of the coding sequence (CDS) covered by RNA-seq data was measured. The subset of genes predicted to be present was used as a positive control to estimate the average CDS fraction covered by RNA-seq data. On average, 68% of the CDS was covered by RNA-seq reads with 95% of these genes having 60% or more of their CDS covered. Using this 60% as a threshold, the missing genes were assessed. In average, 80 genes predicted to be absent from a single cultivar were found in the RNA-seq data with an average horizontal coverage of 80% of the CDS. This represents an error rate of 0.7% in the PAV calling pipeline for this specific dataset.

3.3.7 Pangenome expansion modelling

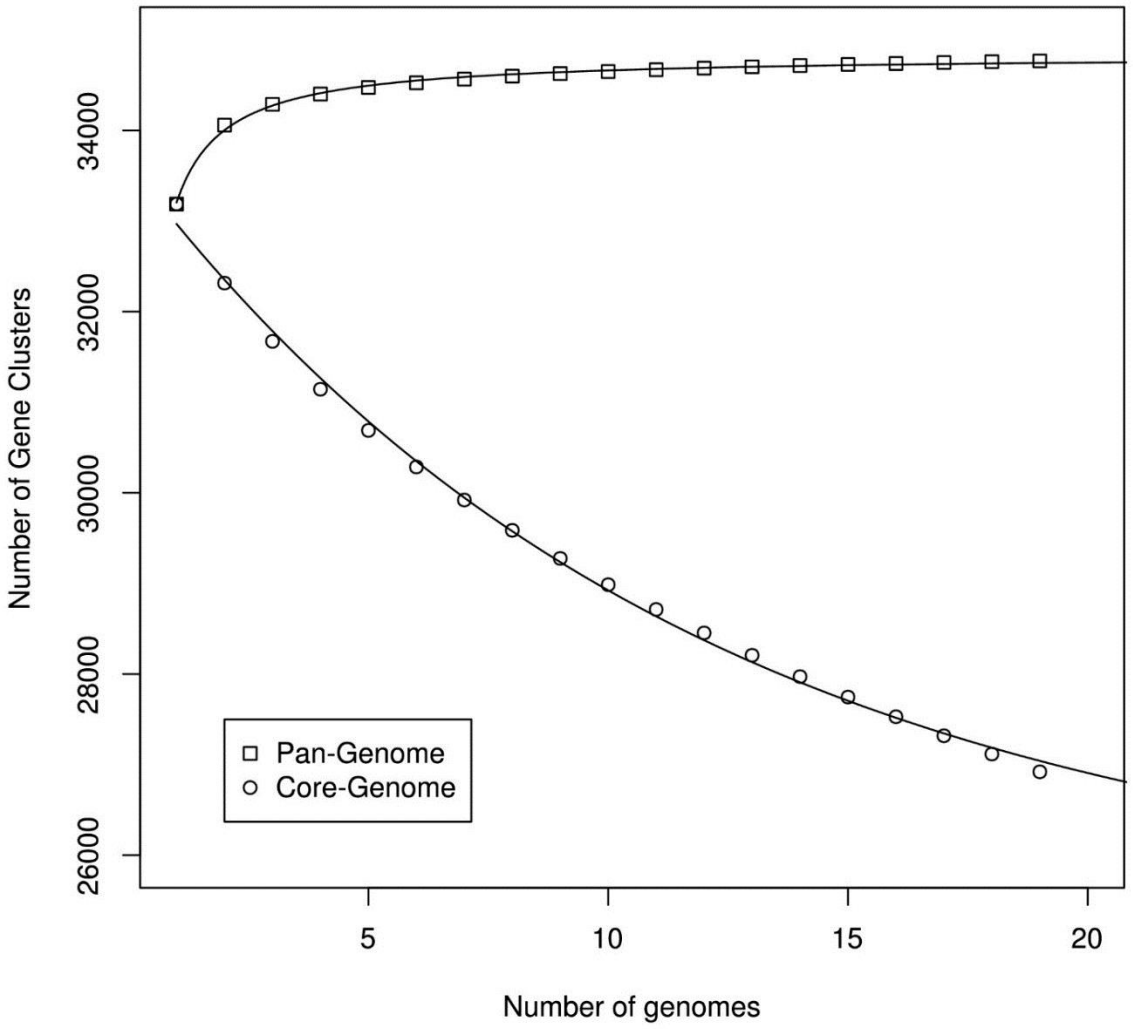
To estimate the gene content of the wheat pangenome, the total gene number was modelled as a function of the number of genomes included and adjusted to a power law. The total number of core genes was also modelled as a function of the number of genomes sequenced. The analysis shows that the pangenome of wheat is closed and contains approximately 140,500 +/- 102 gene models for this specific dataset (Figure 3-4). The reduction of the core genome was also used to estimate its gene content in 81,070 +/- 1,631 gene models.

Figure 3-4. Modelling of the wheat pangenome. Pangenome expansion in gene content was modelled as a function of the number of genomes included in the analysis. Mean gene counts for all combinations of "x" genomes are presented in the figure.



Similarly, the expansion of gene clusters was modelled as a function of the number of genomes sequenced. Based on the model, the approximate number of clusters contained in the wheat pangenome was 34,837 +/- 17 clusters and 25,402 +/- 238 were considered part of the core clusters (Figure 3-5).

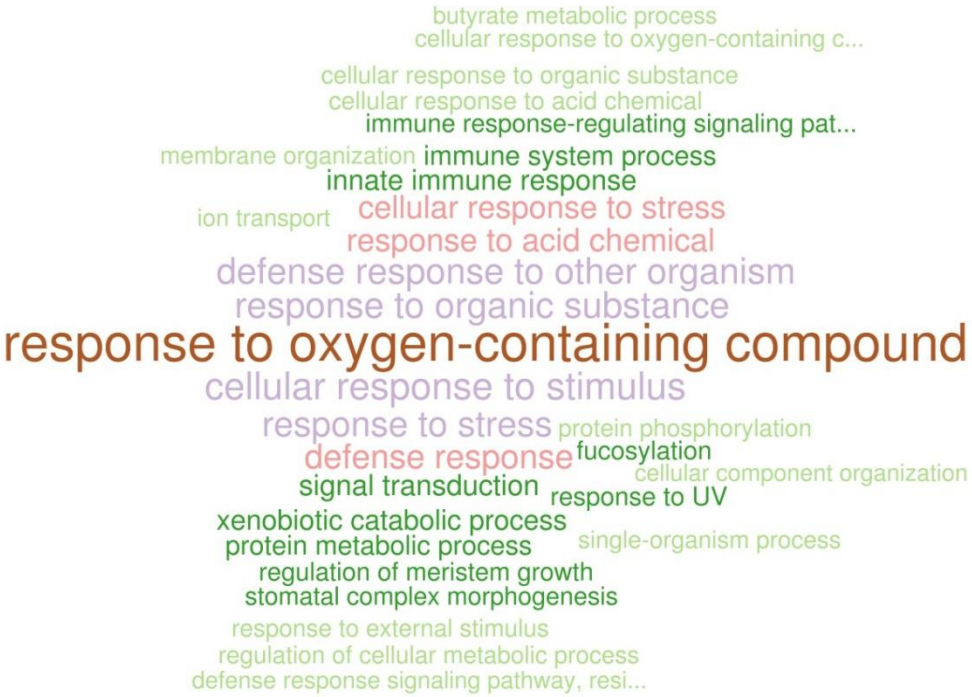
Figure 3-5. Modelling the expansion of gene clusters in the wheat pangenome.



3.3.8 Functional enrichment of the variable genome

Functional annotation of the wheat pangenome was performed with Blast2GO and GO terms were assigned to 64,998 genes in the wheat pangenome. Of these, 17,738 genes were variable genes. These variable genes with functional annotations were tested for enrichment in comparison with the entire dataset. Figure 3-6 shows the biological processes that were found enriched in the variable genome. Response to biotic and abiotic stress were the predominant biological processes that were statistically more abundant (Fisher exact test, $p \leq 0.01$) in the variable genome. Several enriched terms represent processes that play a role in different pathways e.g. signal transduction is an essential part of any response to the environment (biotic or abiotic stimulation) or to hormonal induction.

Figure 3-6. Functional enrichment of the variable genome of wheat. The size of the font reflects the significance of the enrichment with larger font size being the most significant ($p<1e^{-30}$) and the smallest font size the least significant ($p < 1e^{-5}$)



3.4 Discussion

3.4.1 Assembly of the pangenome

In this chapter, eighteen elite wheat cultivars were used to extend the known sequence of the hexaploid wheat genome, assembled in the previous chapter. This extended reference genome contains additional genes that were not present in the Chinese Spring assembly and can thus be referred to as the first draft pangenome of wheat. The methodology used for the assembly and annotation of the pangenome was designed to make the best use of all the wheat sequencing reads publicly available and to reduce the effect of DNA contamination and low quality reads that may be enriched in the unmapped fraction of a sequencing library. In many aspects, the methodology used was limited by the quality and quantity of sequencing data publicly available and the results are a reflection of the diversity found in the germplasm included in this study.

After the initial reassembly of the wheat genome, we mapped all available reads from eighteen cultivars to the reference genome which included the sequences of the chloroplast and mitochondrial genomes. A study on soybean found that unmapped reads from 8 diverse soybean lines showed an enrichment of chloroplast, mitochondria and transposable-elements derived reads (Sonah et al., 2013). Including these reads in the assembly would have provided some insights into the extranuclear diversity in wheat. However, both organelles have shown little diversity between polyploid species and their diversity could be better examined through a mapping approach (Ogihara and Tsunewaki, 1988, Ishii et al., 2001) rather than a *de novo* assembly. The organelle reads were excluded from the unmapped reads dataset in order to decrease the computing time for assembly and reduce the complexity of the dataset for *de novo* assembly.

Mapping efficiency of raw reads to the wheat genome varied from 70% to 90%, with the exception of 5 libraries from the cultivar Baxter (BX-1) that had mapping efficiencies of ~30%. The mapping efficiencies of these 5 libraries could not be improved despite several rounds of stringent quality trimming and more relaxed mapping conditions. Further analysis of a subset of reads from these libraries showed a large number of bacterial and fungal sequences present. These libraries were subsequently removed from the analysis altogether to avoid overrepresentation of contaminating sequences in the assembly. For the rest of the libraries, the mapping efficiencies were similar to those found in other plant resequencing projects (Duitama et al., 2015, Bekele et al., 2013, Gordon et al., 2014). Mapping efficiency is related to the quality of the library used. Small template sizes, unintended introduction of foreign DNA during sample collection/processing, or low quantities of DNA extracted will all affect the mapping efficiency of the reads produced. High quality libraries usually have high mapping efficiencies when mapped to a close reference genome. Using more distant genomes can still produce valid alignments, for example mapping of shotgun reads from wheat chromosome arm 7DS to the *Brachypodium dystachyon* reference genome (International Brachypodium, 2010) showed high mapping efficiency to the coding sequences (CDS) of conserved genes in syntenic regions, despite being separated by 32 million years of evolution (Berkman et al., 2011b). Unmapped reads are usually enriched with low quality reads, reads with adapters and PCR primer leftovers, unintended foreign DNA contamination, transposable elements and reads from highly diverged loci of the genome that do not have a counterpart in the reference used. It is the latter, the ones that we aim to assemble to gain a better understanding of the diversity in the wheat germplasm that are of interest, and therefore,

the methodology was designed to reduce the impact of the other categories of reads present in the unmapped reads dataset. A comparison of non-human sequences in human sequencing data revealed that foreign DNA contamination could be found in every DNA sequencing centre and all libraries had different levels of contamination (Tae et al., 2014). Furthermore, common contaminant sequences appeared in samples from different sequencing centres such as human adenovirus. A similar pattern in the libraries could explain the presence of large scaffolds of bacterial and fungal origin in the raw pangenome assembly (Figure 3-1). Pathogenic or commensal DNA could have been extracted alongside the plant DNA, introduced in the libraries and enriched in the unmapped reads datasets after mapping to the wheat genome.

3.4.1.1 Pre-processing of the raw data

To deal with low quality sequences in the dataset, two methods are commonly used in *de novo* assembly projects: (1) quality control removing low quality stretches and end clipping using known PCR primers and adapter sequences or (2) the correction of erroneous reads by kmer analysis. The latter is based on the analysis of kmer frequency in the raw reads to modify the sequence of very low frequency kmers (1X) which usually appear only in erroneous and low quality reads (Marcais and Kingsford, 2011, Marçais et al., 2015). These low frequency kmers are modified in as few bases as necessary until they are identical to a higher frequency kmer. This approach was used successfully in the assembly of the 22 Gb loblolly pine genome where the authors constructed a database of 24mers to identify erroneous kmers (multiplicity ≤ 15), repetitive kmers (multiplicity >120X) and single copy kmers (multiplicity ~60X) and used this database to correct the erroneous kmers prior to the assembly (Neale et al., 2014, Zimin et al., 2014). However, there are two assumptions of this approach that are not held by pooled read datasets. First, uniform read coverage across the sample's genome. This assumes that every locus in the genome has equal chance of being sequenced and therefore equal chance of producing similar amounts of reads. As discussed before, in our pooled data, loci shared by many individuals have a much higher read count than the loci present in only a few individuals effectively skewing the read distribution. In second place, the approach assumes all low frequency kmers are erroneous, which is mostly true in libraries produced from a single sample, but false in libraries from extremely diverse samples, like environmental samples that may contain entire communities. In the latter, legitimate kmers may have an extremely low frequency due to the low abundance of a variant in the community relative to other much more common variants and this difference is exacerbated by PCR amplification prior

to sequencing. For these reasons, no error correction was performed in the dataset prior to *de novo* assembly of the pooled reads and the reads were pre-processed to eliminate highly erroneous reads and trim adapter leftovers that would have prevented correct identification of overlapping nodes in the assembly graph (Bolger et al., 2014)

3.4.1.2 Assembly methodology

If deeper sequencing of the cultivars had been available, the unmapped reads from every individual would have been assembled separately and the reference genome would have been expanded based on the additional contigs obtained from every individual after a thorough comparison with the reference genome (Golicz et al., 2016b). In the soybean pangenome, the authors sequenced and assembled 7 wild *Glycine soja* accessions individually using an average of 111.9X sequencing depth per accession (Li et al., 2014c). For the *Brassica oleracea* pangenome, the author performed *de novo* assembly of every sample with sequencing depth ≥ 30X and pooled the reads of the remaining samples prior to *de novo* assembly (Golikz, 2016). In *Solanum tuberosum,* 55.7Mb of additional sequence were found after resequencing 12 potato landraces from 30X to 69X and *de novo* assembly of the unmapped reads (Hardigan et al., 2016). This additional sequence had not been captured in the original reference genome assembly and can be considered part of the potato pangenome (PGSC, 2011). Also, in *Arabidopsis thaliana*, 80 diverse inbred samples were sequenced to a depth between 10X and 20X and unmapped reads were assembled *de novo* revealing 43,003 contigs that could be anchored to the reference genome (Cao et al., 2011). Unfortunately, the individual coverage of the samples used in this thesis was not high enough to allow an efficient *de novo* assembly of the unmapped reads per individual. That is why the reads from several individuals were pooled together prior to assembly. This approach was successfully used by Yao et al. (2015). In their study, they used very low sequencing depth (1-3X) from 1483 cultivated rice accessions and assembled the reads from two pools: the *indica* and the *japonica* groups. Yao recognise that the main challenge in assembly of pooled reads was the uneven read distribution across the target genome (Yao et al., 2015). This means that regions shared by multiple individuals in the pool will have a higher sequencing depth than those shared by fewer individuals. Assuming an average sequencing depth of 10X per sample, a loci shared by 2 samples would have a sequencing depth of 20X, those shared by 3 samples, would have 30X and those shared by all samples would have 160X. Uneven read distribution may not be an issue when pooling reads from few closely related individuals where the sequencing depth does not vary in more than 10X to 40X as was done for

Brassica oleracea (Golikz, 2016), but it is a concern when the average coverage per site varies dramatically.

Our pooled data has an uneven sequencing distribution that ranges from 10X to 160X which is similar to that found in metagenomic samples where every locus has a significantly different read abundance (Peng et al., 2012a, Davenport and Tümmler, 2013). Peng et al. (2010) showed that using a single kmer size for *de novo* assemblies could lead to either over branching in higher coverage regions or gaps (missing kmers) in low coverage regions which results in suboptimal assemblies. This is especially true when the dataset has a non-uniform read distribution as is the case for metagenomics samples, single cell sequencing experiments, samples obtained by multiple displaced amplification (MDA) or pooled reads from several different individuals. The assembler IDBA-UD faces this problem by using different kmer sizes in iterative steps and combining the results into a single graph to produce a consensus assembly (Peng et al., 2012). IDBA-UD is one of the most used metagenome *de novo* assemblers and the proposed used of multiple kmer sizes has been implemented in many state-of-the-art *de novo* assemblers including SPAdes (Bankevich et al., 2012), metaSPAdes (Nurk, 2016), HGA (Al-okaily, 2016), HyDA (Movahedi et al., 2016) and a new Superstrings graph algorithm (Cazaux et al., 2016). The metagenomics assembly approach has been successfully used by Yao et al (2015) to study the dispensable genome of rice from thousands of rice accessions that had been sequenced to an average 3X sequencing depth (Yao et al., 2015).

3.4.1.3 Contamination identification and removal

The raw assembly contained 659.7 Mb in 329,000 scaffolds. Contaminant sequences were detected by alignment to the NCBI non-redundant nucleotide database (NT) and only the best alignments for each scaffold were kept. A total of 309 Mb in 105 thousand scaffolds were removed from the assembly due to its similarity to sequences outside of the green plants clade (Viridiplantae, TAXON ID: 33090). The majority of scaffolds aligned to *Triticum aestivum* and *Hordeum vulgare,* and most of the contaminant sequences were part of the bacteria, fungi and metazoan groups (Figure 3-1). This result contrasts with that of the *B. oleracea* pangenome, where the amount of sequences discarded as contamination was very low for all samples except broccoli which contained 21% of the contigs as contamination from *Herbaspirillum seropedicae* (Golicz, 2016). In this wheat assembly nearly half of the assembly (41% of the assembly, 36% of the contigs) were removed. These contrasting results may be explained by the different methodologies and

libraries used for the assembly. The low sample coverage prevents unique sequences from being assembled and this reduces the total assembly size and increases the relative abundance of contamination in the assembly. The level of contamination is highly dependent on the sample although it is possible that the combination of unmapped reads from multiple libraries can increase the coverage of common pathogenic and commensal DNA sequences that were extracted alongside the sample DNA.

3.4.2 Assessment of the pangenome

After contamination removal, 350 Mb of additional sequence contained in 221,991 scaffolds were obtained from the unmapped reads of the 18 wheat cultivars. The sequence similarity between the scaffolds and green plants database (Blast, e-threshold ≤ 1e-10) provides evidence that the assembly represents plant sequences (Figure 3-1). Furthermore, analysis of the genera found in the Blast hits, shows that the great majority of Blast hits were to *Triticum*, *Hordeum*, *Brachypodium* and *Aegilops* all close relatives of *Triticum aestivum* (Figure 3-2). Mapping of pre-processed reads from the wheat chromosome arms to the wheat pangenome shows that these sequences were not present in the read dataset and represent a real expansion of Chinese Spring reference genome. Additionally, comparison of the efficiencies between mapping the reads to the Chinese Spring reference and to the pangenome reference shows an increase of mapping efficiency when using the pangenome (Figure 3-3), which confirms that the pangenome contains sequences that are not in the Chinese Spring reference, but are in the other 19 cultivars. The small increase seen when mapping reads from Chinese Spring to its own reference genome and to the pangenome confirms that most of the additional sequences present in the pangenome do not appear to have a counterpart in the Chinese Spring genome.

These results provide further evidence that the use of a single reference genome is insufficient to capture the full genomic diversity present in a species. Similar conclusions have been drawn from other pangenomic studies. Golicz et al. (2016) assembled an additional 99 Mb of the *B. oleracea* pangenome that were not found in the TO1000 reference genome by iterative mapping and assembly of unmapped reads (Golicz et al., 2016b). In rice, Yao et al. (2015) assembled 88.8 Mb and 57.0 Mb of additional sequence that was not present in the Nipponbare reference genome from the Group *indica* and Group *japonica* respectively (Yao et al., 2015). The sequencing of 80 *Arabidopsis thaliana* accessions also produced additional sequence that was not found in the Col-0 reference

genome. Due to the low sequencing depth per sample for each of the 80 accessions, the authors acknowledge the additional sequences are an underestimation of the total sequence diversity present in the samples (Cao et al., 2011). More recently, thousands of deleted genes were identified in a panel of 12 monoploid accessions of *Solanum tuberosum* Group *Phureja*. The authors found evidence that 21.9% of the genes annotated in the reference genome (DM) are dispensable and that only 30,401 genes were shared by all sequenced accessions of potato. This is still a conservative approximation of the real size of the dispensable genome of potato, because the monoploid panel used in this study has been freed from deleterious or dysfunctional haplotypes that could still be present in diploid or polyploid varieties, but would not survive anther culture and *in vitro* propagation (Hardigan et al., 2016). Similarly, the additional sequence assembled in this wheat study is likely an underestimation of the total sequence available in the wheat gene pool. The raw data available determined the approach taken to construct the pangenome. This means that the total assembly contains mostly sequences present in more than one cultivar and unique loci that do not have enough coverage are unlikely to be present. Even though IDBA-UD uses an iterative assembly approach with different kmer sizes to optimise the assembly of low coverage and high coverage regions, simulated data of bacterial metagenomics samples showed that it was only able to assemble 81% of the sequence with coverage of 10X (Peng et al., 2010) which is the expected coverage for loci unique to one cultivar. The expected assembly efficiency may drop even further if we take into consideration that our real case scenario uses plant genomes, which are more complex than bacterial genomes, and that the actual data contained more contamination than would appear in a simulated sample, which in turn affects the actual coverage of legitimate cultivar-specific sequences.

A recent study by Liu et al (2016) assembled unmapped reads from chromosome 3B of cultivar CRNIL1A and determined that up to 8.3 Mbp of sequences present in the CRNIL1A 3B chromosome were absent in its Chinese Spring counterpart and estimated that 159.3 Mbp present in the CRNIL1A could be absent in the Chinese Spring reference genome. Additionally, the authors assembled RNA-seq data from 28 wheat cultivars and aligned the assembled transcripts to the IWGSC reference genome, the TGAC v1 genome, the W7984 assembly and the assembly of 16 wheat cultivars described here. The authors found that 10% of the transcripts not found in any Chinese Spring reference assembly were found in the assembly of the 16 Australian wheat cultivars we provided (Liu et al., 2016). The assembly provided by us did not include the assembly of unmapped

reads from the W7984 and OpataM85 genomes. Nevertheless, only 10% of the non-CS transcripts could be found in our assembly, while an additional 22.1% of the transcripts were found on the W7984 assembly. Taken together, the full assembly of 18 wheat cultivars presented here contains 32.1 % of non-CS transcripts. Further analysis of the non-CS transcripts showed that only 45.9% were detected in wheat or close relatives and 68.8% had significant Blast hits against the NCBI NR protein database. Taking the 45.9% of these transcripts as the most reliable subset assembled given their presence in wheat and close wheat relatives, our assembly accounts for 69.9% of these high confidence transcript set.

3.4.3 Placement of scaffolds to the Chinese Spring reference genome

Newly assembled scaffolds were anchored to the reference genome based on paired-end information, and 31% of all the scaffolds could be anchored to a chromosome arm in the Chinese Spring reference genome. This is similar to the 28% of scaffolds anchored to the TO1000 reference genome in the *B. oleracea* pangenome following a similar approach (Golikz, 2016). The placement of these sequences into the reference genome was usually supported by two or more cultivars which strengthens the accuracy of the placement and confirms the origin of these sequences as legitimate wheat derived sequences that were present in the additional cultivars but absent in the Chinese Spring genome. The distribution of the scaffolds across the reference genome show a preference for the A and B subgenomes (36% and 37%, respectively, Table 3-3) over the D genome (27%). This result is likely related to the significantly lower genetic diversity found in the D genome. This characteristic of the D genome has been extensively documented and related to the continuous gene flow between tetraploid and hexaploid varieties compared to the limited gene flow between the diploid and hexaploid varieties (Lelley et al., 2000, Siedler et al., 1994, Bryan et al., 1997, Talbert et al., 1995, Paux et al., 2006, Akhunov et al., 2010, Brenchley et al., 2012, Lai et al., 2012b, Berkman et al., 2013a, IWGSC, 2014). Across homeologous chromosomes, group 7 contained a significantly larger fraction of new scaffolds, whereas group 1 contained fewer scaffolds than the rest of the homeologous groups. The placement of 30% of the additional scaffolds to chromosomes in the group 7 is puzzling. The fact that large continuous assemblies like the 3B chromosome, the 1DS chromosome arm or the 2BS and 5B chromosomes do not contain a similarly high number of assigned scaffolds (Table 3-3), discards the possibility of assembly quality as a factor influencing the placement of scaffolds. Additionally, the fact

that this distribution pattern remains constant for each individual sub-genome suggest that this distribution of additional scaffolds is not random.

3.4.4 Annotation of the wheat pangenome

Annotation of the additional sequences was done following the same approach as the annotation of the wheat genome in the previous chapter. The annotation was supported by external evidence from aligned plant proteins, plant EST databases and RNA-seq data from the 11 cultivars for which it was available (Wang et al., 2014). This annotation strategy had been successfully applied in Chapter 2 to identify 98% of core eukaryotic genes and 97% of universal single copy orthologs. Putative genes without support from external sources were not considered for further analysis to prevent an overestimation of gene content in the variable genome. The lack of high quality alignments to known plant proteins (>90% similarity over >90% of the protein) prevented the identification of genes that had been split across different contigs. In total 21,459 gene models were annotated in the additional sequences assembled. This represents an increase of 15% in the number of genes and is comparable to results in other plant species (Golikz, 2016, Yao et al., 2015, Sun et al., 2017). Recent studies in plants have shown that the number of genes in a pangenome can be up to 30% higher compared to a single reference genome (Sun et al., 2017). The total gene content in a pangenome assembly depends on several factors including the number of genomes included on the analysis, the diversity of the samples analysed and the version of gene annotation used for the analysis. The first two factors have a direct correlation with the total number of genes, the more samples studied or the more diverse the samples analysed, the greater the number of genes detected. The correlation between the version of gene annotation used and the total number of genes is not so direct because annotation versions are continuously updated. For example, including unsupported gene models may artificially increase the total gene content, as would the addition of fragmented gene models. In the *Brassica oleracea* pangenome, 5,197 novel genes were found among the 9 cultivars analysed which represents an increase of 9% in the number of genes compared to the TO1000 reference genome (Golicz et al., 2016b, Golikz, 2016), whereas in rice, the total number of genes after sequencing 3010 rice varieties increased by 30% (Sun et al., 2017, Li et al., 2014b). An increase of 15% in the number of genes of the wheat pangenome is larger than the 9% observed in the *B. oleracea* pangenome and looks modest compared to the 30% observed in the *O. sativa* pangenome and can be explained by the intermediate number of samples included here. The total number of genes found here are probably an

underestimation of the gene pool in the 18 samples because genes that appear in only one or two samples are unlikely to be assembled by IDBA-UD due to their low coverage (Peng et al., 2012). Furthermore, the inclusion of more distant wheat varieties such as wheat landraces and European or American elite cultivars may increase the number of novel genes which are not found in the Australian germplasm and that contain adaptations to other environments, day-light cycles and resistance to different pathogens absent from the Australian mainland.

3.4.5 The core and variable genomes of wheat

To determine the extent and content of the wheat core and variable genomes we first determined the presence or absence status of individual genes in each of the 19 cultivars based on the horizontal coverage of their coding sequences. The method has been successfully used in different projects including the identification of 57 genes that were missing in at least one of the isolates of the canola pathogen *Leptosphaeria maculans* (Golicz et al., 2015b). It was also successfully used to detect the loss of the ethylene biosynthetic pathway in the seagrass *Zostera muelleri* (Golicz et al., 2015b) and in the identification and analysis of the core and variable genomes of *Brassica oleracea,* showing low rates of false-positive prediction (0.05) for genes present in contigs 1 Kb or larger, which correspond to 97% of the total number of genes annotated in the *Brassica oleracea* pangenome (Golikz, 2016, Golicz et al., 2016b). Based on the Lander-Waterman analysis (Lander and Waterman, 1988), the probability of missing a base in the 17 Gbp wheat genome with a sequencing depth of 10X is 5×10^{-5} which translates into 772 Kbp. Analysis using RNA-seq data of 11 of the 19 cultivars showed an average false-negative rate of 0.7% which is slightly higher than the rate observed in *B. oleracea* or *L. maculans*. The higher error rate may be due to the more fragmented nature of the wheat pangenome. In the *B. oleracea* pangenome, including genes present in contigs smaller than 1 Kbp increased the error rate to values similar to those observed here (Agnieszka Golicz, personal communication). The fact that more than half of the wheat pangenome assembly is contained in scaffolds smaller than 1 Kbp made it impossible to use this filter to reduce the error rate. However, although this error may confound the boundaries between core and variable genomes by making some core genes appear to be variable, it does not fundamentally change the estimation of gene content in the pangenome, nor does it significantly change the estimated sizes of the core and variable genomes based on mathematical regression (Figure 3-4 and Figure 3-5).

Our results show that 64.3% of the genes in the wheat pangenome belong to the core genome and the remaining genes are part of the variable genome. These results are comparable to those found in the rice pangenome; where over 30% of the annotated genes in more than 3,000 rice accessions belonged to the variable genome (Sun et al., 2017, Yao et al., 2015, Li et al., 2014b). However, these numbers differ from the results in *B. oleracea* and *G. max* in which nearly 80% of the annotated genes were assigned to their core genome (Li et al., 2014c, Golicz et al., 2016b). In maize, the size of the core and variable genomes has not been determined accurately. A first study using 6 inbred lines from different heterotic groups from China uncovered hundreds of genes with presence-absence variation, but the large majority of the nearly 40,000 genes in the maize genome was considered core (Lai et al., 2010, Springer et al., 2009). However, a follow up study of more than 500 maize accessions, showed that nearly 50% of all the representative transcript assemblies (RTAs) were absent from the B73 reference genome (Hirsch et al., 2014). The relative size of the core genome partially depends on the number of accessions included in the construction of the pangenome. Fewer genomes usually results in a pangenome with a higher content of core genes and as more genomes are added, more core genes are found to be affected by PAV. Similarly, the use of closely related varieties also results in a larger core genome due to the larger number of similarities between the samples. The results in the wheat pangenome could be caused by an interplay between the higher number of accessions compared to the 9 used for *B. oleracea* or the 6 accessions in the soybean, and a more complex genome structure which includes high content of orthologous genes and transposable elements. It is possible that some genes that legitimately belong in the core genome have been classified as variable due to a false-negative call. Yet, even correcting for the error rate in the PAV calling pipeline, the size of the core genome would not increase dramatically and this would not affect the total gene content estimated for the pangenome.

3.4.6 The wheat pangenome is closed

After modelling the pangenome expansion as a function of the number of genomes added using the average gene content of all possible genome combinations for each value of x (Figure 3-4) we found that the wheat pangenome is closed and estimate its upper limit corresponds to 140,500 genes. This is in agreement with previous findings in other plant pangenome projects and bacterial pangenome projects. Closed pangenomes are likely a reflection of the molecular mechanics involved in the origin of variable genes. In bacteria, lateral gene transfer is a major mechanism in the acquisition of new genetic material,

species with high rates of horizontal gene transfer tend to have larger pangenomes, and the fraction of their core genome tends to be smaller (Rouli et al., 2015, Tettelin et al., 2008, Lapierre and Gogarten, 2009, Medini et al., 2005). Meanwhile, the origin of variable genes in plants is largely affected by gene loss, genome duplication followed by diversification and activity of transposable elements (Jin et al., 2016, Hirsch et al., 2014). Given these differences, it is expected that plant pangenomes are usually closed as has been seen in *B. oleracea* (Golicz et al., 2016b) and *G. max* (Li et al., 2014c).

Modelling the pangenome expansion also allowed us to estimate the total gene content in the wheat genepool. The current estimation is based on closely related cultivars selected mostly from the Australian germplasm and does not necessarily reflect the gene content if other wheat germplasm were to be evaluated. Including more diverged wheat cultivars such as landraces or newly synthesised hexaploid wheat designed to harbour new genes from wild *Ae. tauschii* accessions. Including such divergent varieties would impact the pangenome size estimations by increasing the total gene content and probably reducing the core genome size.

3.4.7 Local adaptation to pathogens and environment shape the wheat variable genome

We have found that the wheat variable genome is enriched with genes involved in the response to biotic and abiotic stress and intracellular signalling pathways. The GO terms enriched are highly concordant with what has been found in the all other major plant pangenome studies done so far (Sun et al., 2017, Liu et al., 2016, Jin et al., 2016, Hardigan et al., 2016, Golikz, 2016, Yao et al., 2015, Li et al., 2014b, Gordon et al., 2014). In bacteria, the variable genomes confer selective advantages such as niche adaptation, antibiotic resistance, the ability to colonize new hosts and other pathogenic and virulence properties (Vernikos et al., 2015, Tettelin et al., 2008). Crop plant genomes have been strongly selected for agricultural production and therefore, their variable genes also contain characteristics that are advantageous for production, pathogen tolerance/resistance and adaptations to local weather patterns that differentiate them from cultivars in other geographical regions. Therefore, gene PAV contribute to the repertoire of genetic variants available to wheat for adaptation to local environments. These genetic differences have been found to be starker between distant landraces which are more specialised to narrow environmental conditions (Iwaki et al., 2001, Villa et al., 2005, Zeven, 1998). Overall, the wheat variable genome adheres to this pattern of local adaptation to

pathogens and environmental conditions through human selection as is evident from its enrichment with defence response genes, environmental stress response and the molecular signalling pathways necessary to react to both.

4 Chapter 4: SNP diversity analysis of the wheat pangenome

4.1 Introduction

The construction of the wheat pangenome in the previous chapter offers the opportunity to explore the genetic diversity of wheat outside of the boundaries of a single cultivar reference genome. Such an approach would give us insights into the evolutionary forces that shape the characteristics of both core and variable genes. Marker discovery in the wheat pangenome and particularly SNP discovery would also be useful in the development of highly integrated high-density genetic maps that permit the anchoring of newly assembled sequences into a framework genetic map. Finally, this could also be used for the assessment of relatedness between the cultivars.

The construction of the wheat infinium array (Bachlava et al., 2012) could be considered a first attempt to explore the diversity of the wheat pangenome. Its development included the *de novo* assembly of transcripts from many different elite cultivars and wild relatives and the identification of homologous gene clusters (HGCs) that were used for SNP discovery. However, the high level of paralogous and orthologous sequences present in wheat and the difficulty in telling them apart from true homologous clusters led to an overall increase in the false-posititve rate in SNP discovery.

In rice, the pangenome sequence was used to identify nearly half a million SNPs and linkage disequilibrium was used to assign position to the SNP-containing contigs in the Nipponbare reference genome (Yao et al., 2015). In soybean, the assembly of wild soybean accessions led to the discovery of more than 4 million SNPs that were used to characterize the variable and core genes. They found that variable genes had a higher SNP density and a higher rate of non-synonymous and deleterious SNPs than the core genes (Li et al., 2014c).

In wheat, several groups have produced genome wide markers and used them to characterize large populations of hexaploid, tetraploid and diploid wheat, get a better understanding of the phylogenetic relationships between different groups of wheat and find interesting candidates for gene introgresion into elite cultivars (Dvorak et al., 1988, Dvorák et al., 1993, Friebe and Gill, 1994, Plaschke et al., 1995, Heun et al., 1997, Dvorak et al., 1998).

RFLP were among the first genetic techniques used to characterize wheat populations. Kam'Morgan et al. (1988) used probes of α-amylase to construct a partial genetic linkage map for hexaploid wheat (Kam-Morgan, 1988). Similarly, Harcourt et al. (1991) found a highly polymorphic RFLP probe, PSR454, isolated from a random genomic library which showed up to 73% polymorphism in a single locus (Harcourt and Gale, 1991). In the late 1990's and early 2000's microsatellites and AFLPs were the preferred tools to study the phylogenetic and evolutionary relationships between wild and domesticated wheat as well as between elite cultivars developed during the green revolution and landraces (Plaschke et al., 1995, Ishii et al., 2001, Akbari et al., 2006, Akhunov et al., 2010).

The advent of next generation sequencing and the genomics era saw the exponential increase in the number of resequencing projects. Genome wide SNPs can now be identified and used for the characterization of hundreds of individuals through genotyping by sequencing. For example, the wheat SNP infinium array was generated from the consensus sequences of RNA-seq libraries from several cultivars (Wang et al., 2014). The use of RAD-seq genotyping by sequencing allowed identification of thousands of SNPs without the need of a reference genome and the construction of high density genetic maps for wheat and barley (Poland et al., 2012b).

More recently, the completion of draft genomes of the wheat chromosomes 7A, 7B and 7D (Berkman et al., 2011b, Berkman et al., 2013a), allowed the identification of over 4 million intervarietal SNPs in group 7 (Lai et al., 2015b), the largest SNP dataset identified prior to this project. This milestone is set to be surpassed with the recent release of several wheat genome assemblies (IWGSC, 2016a, Clavijo et al., 2016, Chapman et al., 2015, Mayer et al., 2014) which will allow more efficient and cost-effective approaches to SNP discovery and genotyping of diverse germplasm. Nevertheless, it is the pangenome that will provide the largest sequence available for the SNP discovery of distant relatives whose genes are not present in any single-cultivar genome assembly.

In this chapter, the SNP diversity of the 19 wheat cultivars is explored using the pangenome assembly as a reference. Then the distribution patterns of the SNPs in both the core and variable genes are characterized across the three subgenomes of wheat. The SNPs identified in this chapter are finally used to construct a high density genetic map that is useful for placing additional scaffolds in the context of the wheat pangenome and

assessing the relatedness of the cultivars through principal component and phylogenetic analysis.

4.2 Materials and methods

4.2.1 SNP discovery

Raw reads from the 19 hexaploid wheat cultivars (Appendix I) were mapped to the reference pangenome assembled in the previous chapter with Bowtie2 v2.2.9 (Langmead and Salzberg, 2012) and standard parameters. Reads mapping with MAPQ < 20 were removed from the alignments with the module "view" of Samtools v 1.2.0 (Li et al., 2009a). The SGSautoSNP pipeline (Lorenc et al., 2012) was used to identify homozygous intervarietal SNPs requiring at least 2 reads per variant for SNP calling and one read per cultivar for genotyping. A minor modification was also used: bases with base quality (BAQ) below 20 were ignored for conflict resolution of intravarietal SNPs.

4.2.2 SNP validation

4.2.2.1 Comparison with the 90K Infinium array

Flanking sequences of the 90K SNPs in the wheat Infinium array (Wang et al., 2014) were obtained and aligned to the wheat pangenome with the megaBlast module from the NCBI-Blast+ v2.2.30 package (Camacho et al., 2009). Only alignments greater or equal to 99% sequence identity and with unique alignment positions in the pangenome were used for comparison. All other partial alignments and multiple alignments were not considered any further. Comparison was performed using the 90K infinium array as the gold standard. Using all common SNPs as the universe, the false-positive rate of SNP discovery was determined as the fraction of SNPs considered monomorphic in the infinium array.

4.2.2.2 SNP distribution and Transition/Transversion (Ts/Tv) ratio

SNP density was measured based on the total assembly size of each chromosome and compared between all chromosome groups. Transition-transversion ratios were calculated with SNP effector v 4.2 (Baets et al., 2012) in a chromosome-wise fashion.

4.2.2.3 Effects of SNPs on the genes of the pangenome

Wheat chromosome assemblies and annotations produced in chapters 2 and 3 were configured in the SNPeffector v4.2 (Baets et al., 2012) database as described in the reference manual (http://snpeff.sourceforge.net/SnpEff_manual.html). The effects of the

SNPs were measured only in the coding sequences (CDS) of the genes. The total number of synonymous, non-synonymous and deleterious SNPs per gene per chromosome were obtained from the final report.

4.2.3 Construction of a genetic map using pangenome-wide SNPs

Whole genome shotgun data from the SynOpDH population was downloaded from the NCBI SRA database (Chapman et al., 2015) (Appendix 1). Reads from the 90 double-haploid individuals of the SynOpDH population were mapped to the wheat pangenome using Bowtie2 v2.2.9(Langmead and Salzberg, 2012). PCR clones and reads with low mapping quality (MAPQ ≤ 20) were removed from the alignments with Samtools v1.3 (Li et al., 2009a). Based on the SNPs found between the parents of the segregating population (OpataM85 and W7984), genotyping by sequencing (GBS) was used to determine the haplotypes of each offspring in the family. After GBS, the missing alleles were imputed based on the genotype of their flanking alleles, if and only if the same genotype was present on both sides of the missing alleles. Finally, SNP markers with similar segregation patterns in the population and located in the same scaffolds were merged into metaSNPs with a maximum of 1 recombination between the merged SNPs. MetaSNPs with 0 missing alleles were used to produce a framework genetic map of the wheat pangenome. MSTMap (Wu et al., 2008) was used on metaSNPs with no missing data. The map was constructed with the parameters shown in Table 4-1. Linkage groups with less than 50 members or less than 10 bins were not considered for further analysis. A reduced metaSNP dataset was selected by pooling one representative metaSNP from each bin in the framework genetic map.

Table 4-1. Parameters used in the construction of the genetic map with MSTMap

population_type	DH
population_name	SynOpDH
distance_function	kosambi
cut_off_p_value	0.000001
no_map_dist	10
no_map_size	30
missing_threshold	0.6
estimation_before_clustering	Yes
detect_bad_data	Yes
objective_function	ML.COUNT
number_of_loci	109137
number_of_individual	90

4.2.3.1 Validation of the genetic map

For every metaSNP in the framework genetic map, its chromosome of origin was used to measure the chromosomal enrichment of every linkage group and to assess the distribution of chromosome-specific metaSNPs across the linkage groups. The data was normalized based on either the total number of metaSNPs per linkage group for chromosomal enrichment or on the total number of metaSNP per chromosome assembly for the distribution of the metaSNPs. MetaSNPs from unplaced scaffolds were divided in two groups based on their assembly of origin: “BP” for those metaSNPs found in scaffolds of the 16 wheat varieties sequenced by Bioplatforms Australia; and “CH” for those found in the scaffolds of the SynOpDH parents OpataM85 and W7984.

4.2.4 Anchoring of unmapped scaffolds to the wheat physical map

The reduced metaSNP dataset was used to construct broader genetic maps with lower quality SNPs by iterative map construction with MSTMap. These genetic maps were used to place and orient contigs and scaffolds based on the location of their SNPs. The algorithm for placement consisted of three main steps:

1) Assignment of genetic position of the contigs/scaffolds: for every contig more than 80% of the SNPs needed to be placed in the same linkage group. A weight was assigned to every SNP in the contig/scaffold based on the number of missing values in their genotype, the more missing values the lower the weight. A weighted average was used as a positional value of the contig/scaffold in the genetic map.
2) Assessment of physical to genetic coherence: in sequences with more than one SNP, all SNPs with the highest quality and different genetic positions were used to determine a positional range of the contig in the linkage group. If the range of a contig overlapped the range of 2 or more contig bins in the same direction, the sequence with the largest range was removed. For contigs with 3 or more high quality SNPs with different positions in the genetic map their order in the physical map was required to be similar to their order in the genetic map, otherwise these sequences were removed from analysis.
3) Placement and orientation: for all remaining contigs that had unique and non-conflicting positions in the genetic map and had a coherent order of markers in the physical and genetic maps, orientation was assigned when possible (two or more SNPs required) and the pseudomolecules were constructed following the order of the contigs based on their weighted genetic position calculated in step 1. Contigs placed in the same position formed a contig bin. Where possible, contig-bins were further ordered based on synteny to Brachypodium distachyon.

A custom program ContigMapper, was designed to implement these criteria using MSTMap results directly along with a table containing basic information on the SNPs. The program was written in GO and is available as a binary or as source-code in this address: https://github.com/jdmontenegroc/contigMapper.

4.2.5 Principal component analysis of pangenome-wide SNP markers

The libraries SNPrelate and gdsfmt (Zheng et al., 2012) from Bioconductor (R Core Team, 2014, Giovanni Parmigiani, 2003) were used to find the top 4 principal components

that explain the largest amount of variation in the pangenome SNP dataset. The SNP dataset was pruned to retain SNPs in relative linkage-equilibrium using the "snpgdsLDpruning" module with an LD-threshold of 0.2. Eigenvectors were calculated with the pruned SNP dataset and the top 4 principal components were used to generate a plot showing the distribution and dissimilarity measure (distance) of the 19 wheat cultivars analysed.

4.2.5.1 Relatedness of the 19 wheat cultivars

A dissimilarity matrix was calculated from the entire wheat pangenome SNP dataset using the snpgdsHCluster function of the SNPRelate package in Bioconductor. The module heatmap.2 from the package gplots (Gregory R. Warnes, 2015) was then used to cluster hierarchically the 19 cultivars based on their dissimilarity measure. Additionally, the programs vcf-tools (Danecek et al., 2011) and tabix (Li, 2011) were used to generate cultivar-specific sequences of the pangenome based on the SNP profile of each cultivar. The auxiliary program 'gffread' from the package Cufflinks (Trapnell et al., 2012) was used to generate cultivar-specific coding sequences (CDS). All the cultivar specific CDS were concatenated into one fasta sequence per cultivar. Variable genes missing from a cultivar were replaced by dashes in the alignment ("-") to reflect a gap. These concatenated sequences were used as alignments for maximum likelihood estimation of a genetic tree with RAxML (Stamatakis, 2006) using 1000 iterations for bootstrap calculation. Also, binary matrices of gene presence-absence variations per cultivar were used for maximum-likelihood estimation of phylogenetic trees with RAxML.

4.3 Results

4.3.1 SNPs in the wheat pangenome

The diversity analysis of the wheat pangenome using 19 cultivars revealed the presence of over 36.4 million (36,413,491) intervarietal polymorphic SNPs. Of these, 2.91 million (2,911,482) were found in scaffolds absent from the Chinese Spring reference genome and represent 8.5% of the total SNPs. These SNPs were evenly distributed across the genome but were more widely spread in all chromosomes from the D genome which exhibited a lower SNP density consistently across all homeologous groups (

Figure 4-1). The SNP density in the unmapped sequences that were assembled in the previous chapter was higher than that found in most of the other chromosome arms, except for chromosome 2B. In all homeologous groups, the B genome exhibited a higher SNP density than the other 2 genomes, except for homeologous group 6 where chromosome 6A showed a higher SNP density.

The number of transitions (Ts) and transversions (Tv) were also measured and their ratio (Ts/Tv) was calculated for all chromosomes. The pangenome Ts/Tv ratio was 2.37 strongly driven by the Ts/Tv ratio found in the A and B genomes (Figure 4-2). Overall, the A genome showed a slightly higher Ts/Tv ratio than the B genome and both the A and B genomes had a significantly higher Ts/Tv ratio than the D genome in all the homeologous groups. The Ts/Tv ratio of the SNPs found in the unmapped assemblies was similar to that found for the entire wheat pangenome (2.36) and was closer to that found in the B genome than in either the A or D genomes.

Figure 4-1. SNP density across the wheat pangenome. For each homeologous group the SNP density per genome is shown with the following colour code: orange: A genome; yellow: B genome, green: D genome and brown: unplaced. For every homeologous group, SNP density in the D genome is always lower.

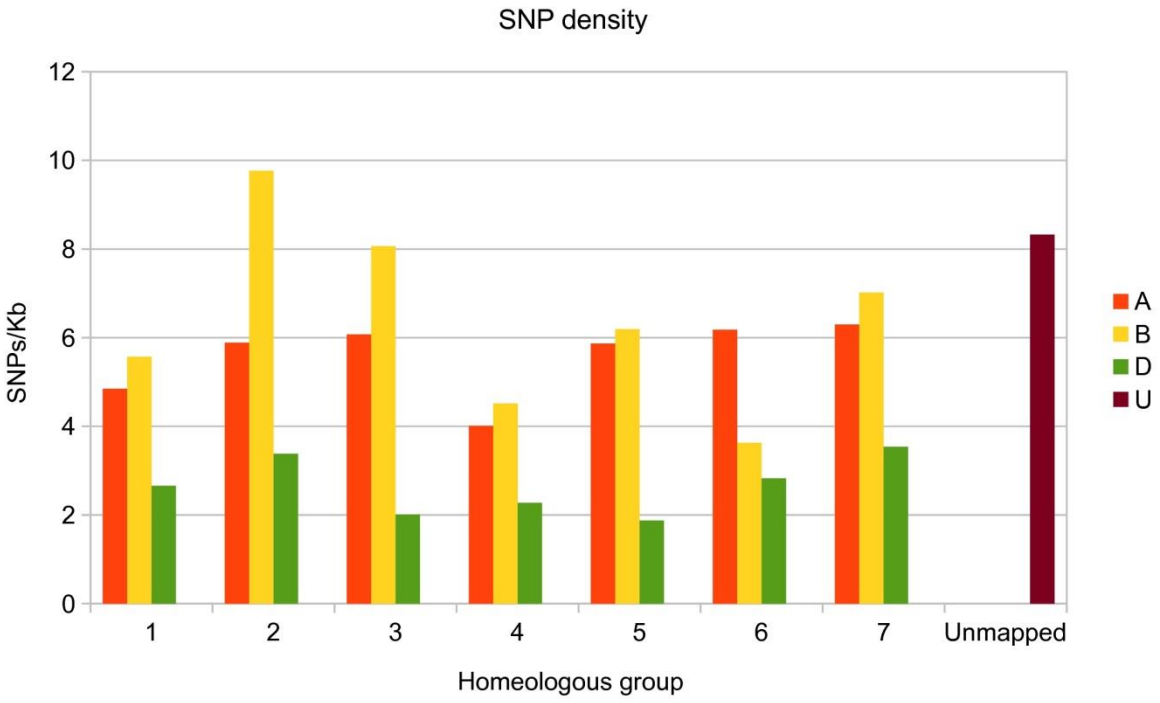
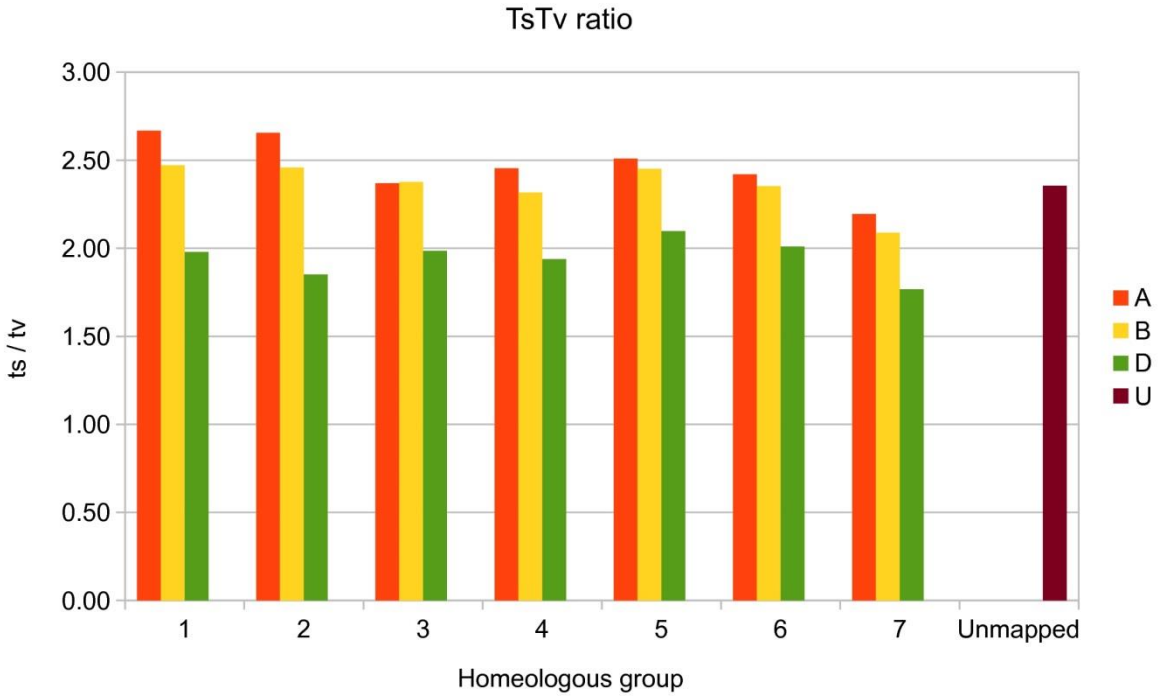


Figure 4-2. TsTv ratio of the wheat pangenome. The Ts/Tv ratio for the entire pangenome was 2.37 with the A genome having an overall higher Ts/Tv ratio, followed by the B genome and both significantly higher than the D genome.



4.3.1.1 Validation with the 90K Infinium array

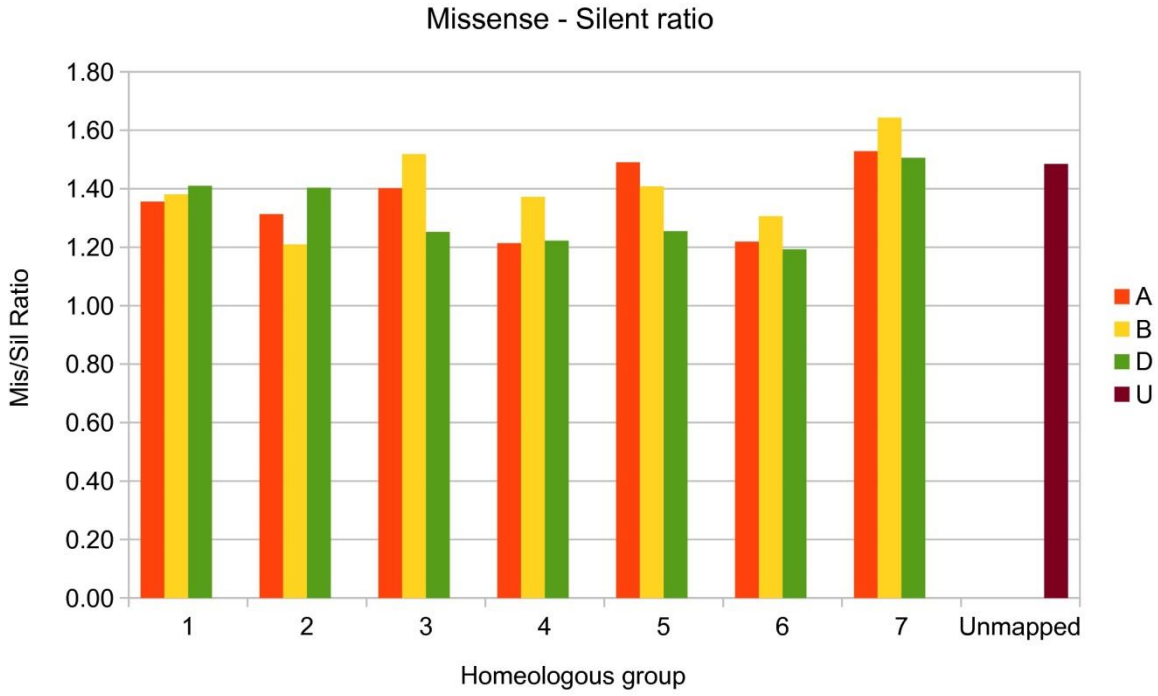
The SGSautoSNP (Lorenc et al., 2012) calls were compared to the SNPs from a recently published Infinium array (Wang et al., 2014). A total of 13,541 Infinium SNPs were identified as having matches at the same position as the SGSautoSNP calls. Out of these 59.8% were identified as polymorphic single locus, 36.4% as polymorphic multilocus, while 3.7% were monomorphic. Taken together 96.2% of the SNPs were validated as polymorphic by comparison with the Infinium array. Also, for every SNP loci analysed, the list of alleles were similar between the SGSautoSNP calls and the Infinium array.

4.3.2 Effects of the SNPs on the core and variable genomes

The majority of SNPs were found in intergenic regions and only 392,557 (1%) SNPs were located in coding regions. Of these, 225,310 (57.4%) are predicted to be non-synonymous mutations that could result in a potentially different functional protein. The ratio of non-synonymous to synonymous SNPs was 1.39 mostly driven by the results in group 7 chromosomes and the unmapped scaffolds assembled in Chapter 3. However, no

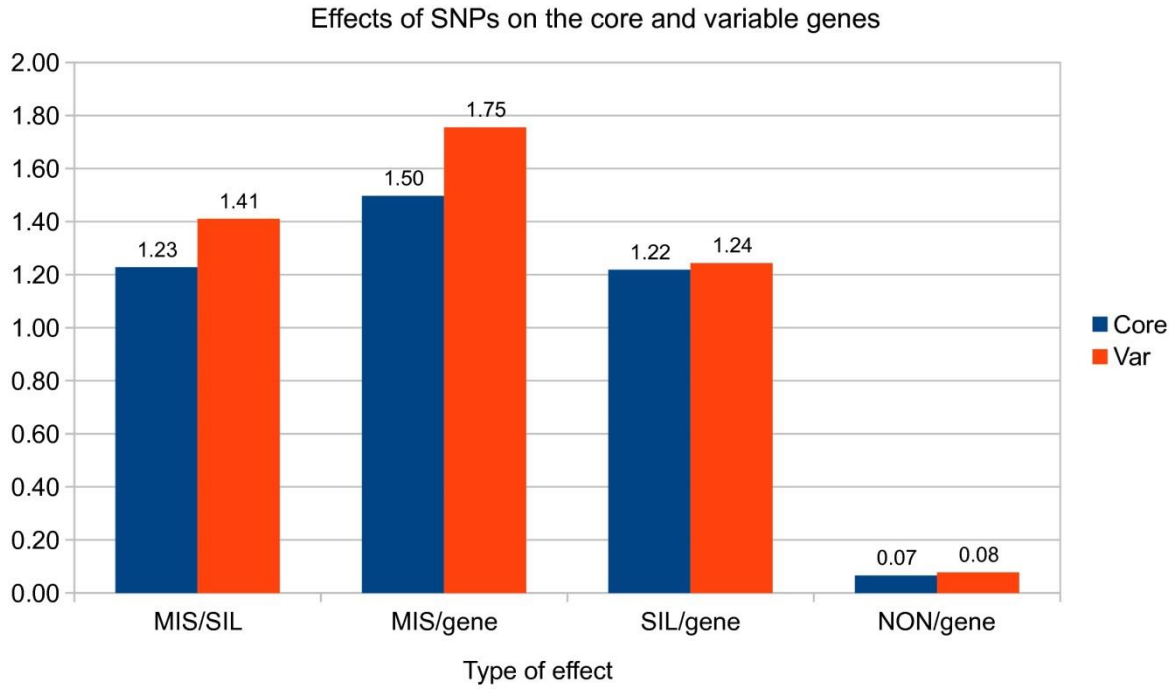
significant difference was found between the Non/Syn ratio of the different genomes or homeologous groups.

Figure 4-3. Non-synonymous to synonymous ratio in the wheat pangenome. The horizontal axis shows the 7 homeologous groups and the unmapped scaffolds assembled from unmapped reads of Chapter 2. The Missense-silent ratio for the entire pangenome was 1.39 with a maximum on chromosome 7B (1.64) and a minimum on 6D (1.19)



An analysis of the effects of SNPs on core and variable genomes showed that the variable genome has a higher SNP density (3.08 SNPs/gene) than the core genome (2.78 SNPs/gene). A decomposition by the effect of the SNPs on both datasets revealed a much higher non-synonymous SNP density in the variable genome (1.75) compared to the core genome (1.50). Whereas the SNP density of silent mutations and non-sense mutations also increased slightly, the largest increase was found in the SNP density of non-synonymous SNPs. This also affected the non-synonymous to synonymous SNP ratio with a higher ratio in the variable genome compared to the core genome (Figure 4-4).

Figure 4-4. Effects of SNPs on the core and variable genomes. Overall, the variable genome had a higher SNP density and a greater Non/Syn ratio. The rate of frequency of non-synonymous SNPs was higher in the variable genome whereas the synonymous and non-sense SNPs were roughly similar in both groups.



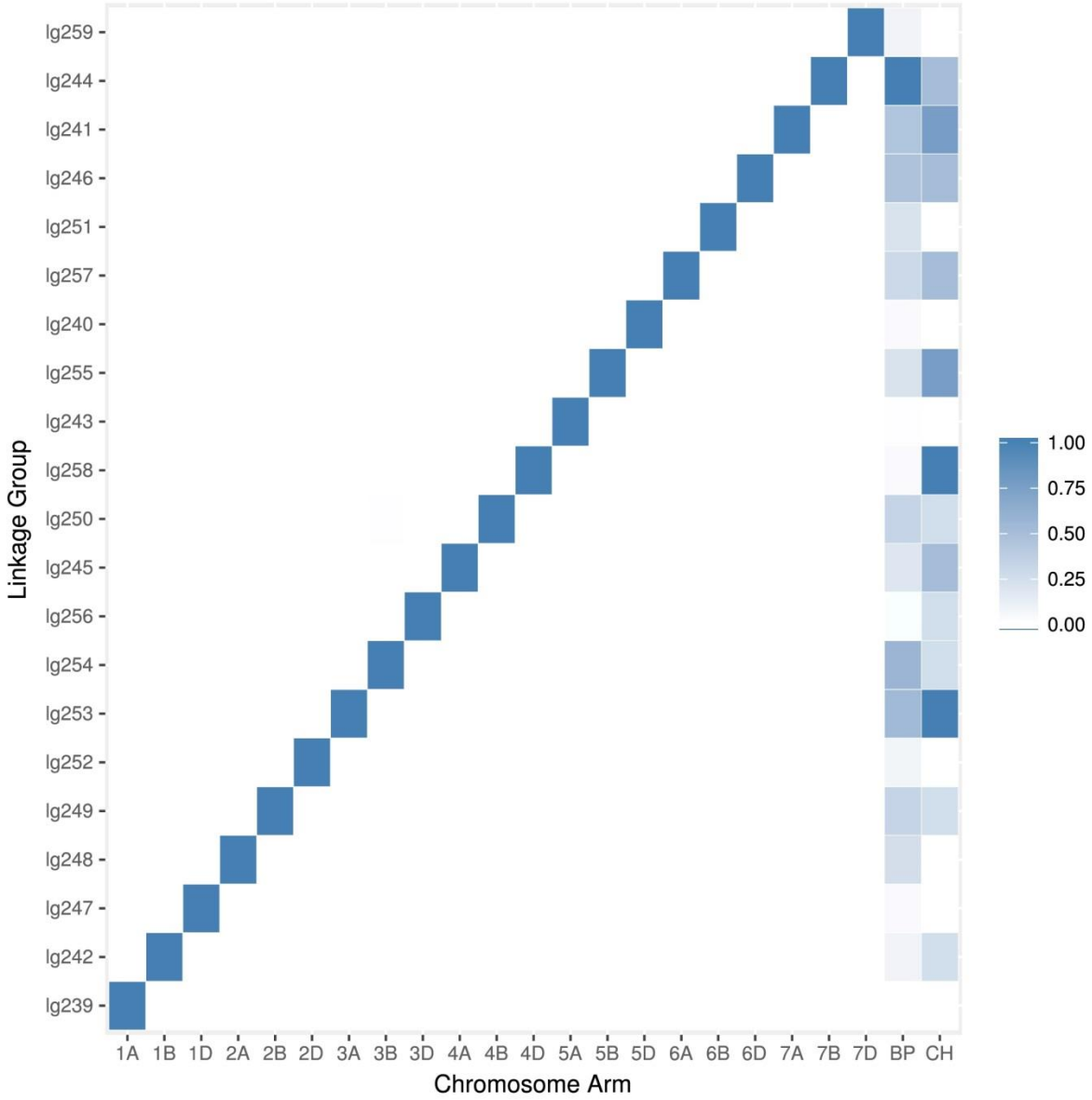
4.3.3 Construction of a genetic map using pangenome-wide SNPs

The SynOpDH population consists of 90 F1-derived double haploid individuals from the cross between OpataM85 and W7984. This mapping population was genotyped and used to build a genetic map of the wheat pangenome. After identification of SNPs between the 19 wheat cultivars, 11,001,655 polymorphic SNPs between the parental cultivars were used to genotype the SynOpDH population. The segregation data was imputed and merged into 2,237,807 metaSNPs. All metaSNPs with no missing data (109,137) were used to construct a framework genetic map of the wheat pangenome. Ninety-nine percent of the metaSNPs (108,808) were placed in one of 21 large linkage groups. The genetic map had a total length of 8,437 cM and contained 4,632 recombination bins. A total of 4,562 metaSNPs from unmapped contigs (27 from SynOpDH and 4535 from Bioplatforms) were also included in the genetic map and placed in the 21 linkage groups. This represents 98.1% of the unplaced metaSNPs used in the construction of the framework genetic map.

To estimate the accuracy of the genetic map two analysis were performed: first, the distribution of chromosome-specific SNPs across the linkage groups was assessed to determine if SNPs from a single chromosome showed a significant presence in more than 1 linkage group. For each of the 21 chromosome specific SNPs, over 99% of the sets were clustered in a single linkage group. The SNPs from unmapped scaffolds (BP and CH) were distributed evenly across all linkage groups with no particular preference for one or another (

Figure **4-5**). This result demonstrates that SNPs from any single chromosome belong together in the same linkage group and are not separated into different groups.

Figure 4-5. Heatmap of the distribution of chromosome specific SNPs across the 21 largest linkage groups. The SNPs from every chromosome arm were clustered in single linkage groups with 98% purity. SNPs from the the unmapped scaffolds (BP and CH) were evenly distributed across the 21 linkage groups.

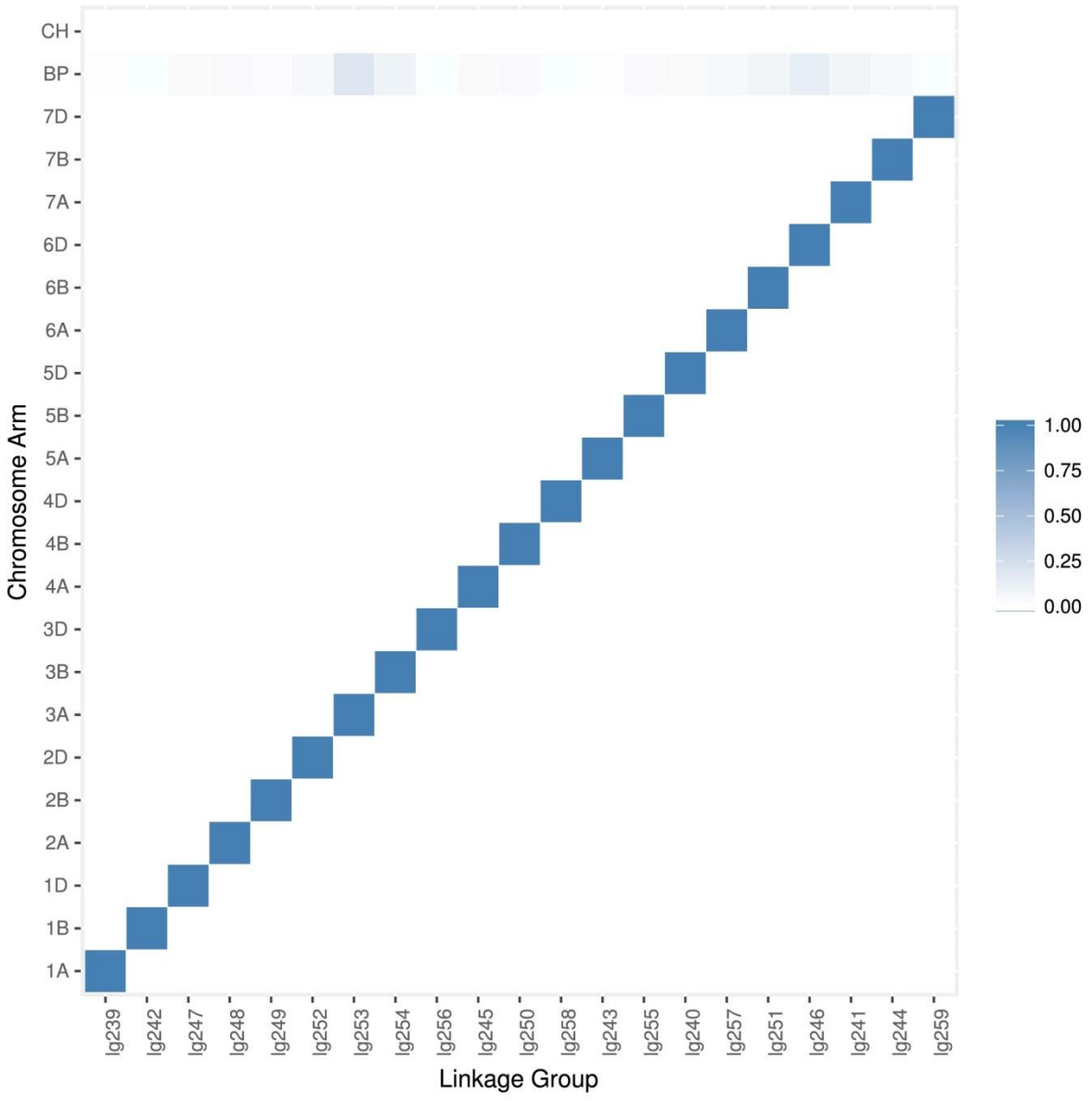


Similarly, the chromosome-specific enrichment of every linkage group was assessed to find out if a linkage group contained a significant amount of SNPs from different chromosomes. In every linkage group, the percentage of SNPs found to be from a single chromosome ranged from 88.6% in lg246 to 98.8% in lg239 with an average of 95%. Excluding the SNPs from unmapped scaffolds from the analysis, resulted in an increased enrichment of all linkage groups to an average of 99.9% that ranged from 99.4% in lg 242

to 100% for lg258, lg253, lg259 and lg 254. These results show that the linkage groups contain SNPs exclusively from one single chromosome arm and not a single linkage group contains a significant amount of markers from a different chromosome, although some of them do contain a significant amount of unmapped SNPs (Figure 4-6).

Taken together, these results show the high accuracy of SNP assignment to the genetic map where every linkage group contains nearly all the SNPs from a single chromosome and do not contain SNPs from any other chromosome as shown in Figures 5 and 6. The high accuracy obtained in the genetic map allows the unequivocal assignment of a chromosome to a linkage group. Furthermore, this genetic map can be used to determine the position of unmapped SNPs in relation to previously placed SNP markers.

Figure 4-6. Enrichment of linkage groups with chromosome-specific SNPs. Every linkage group contains markers exclusively from a single chromosome-specific SNP dataset. The fraction of SNPs from unmapped scaffolds in the pangenome assembly (BP and CH) is small compared to the number of SNPs found in chromosome-specific assemblies. No linkage group shows presence of a significant amount of markers from different chromosomes.



4.3.4 Anchoring of unmapped scaffolds to the genetic map

The framework genetic map constructed in the previous step was used to generate a minimal high quality SNP dataset that included one representative of every recombination bin to a total of 4632 metaSNPs. The minimal subset was used to place the remaining 106,076 metaSNPs from the unmapped scaffolds in the framework genetic map. In total, 78,517 metaSNPs (77.4%) could be placed in one of the 21 linkage groups. The other 22.6% were not placed in the genetic map either due to a high number of missing values (more than 40% of missing data) or due to placement in smaller 2 or 3 member linkage groups that were discarded from the analysis. This genetic map facilitated the placement of 50,035 novel scaffolds that had not been previously placed in the wheat pangenome. Along with the scaffolds placed by read pair information in the previous chapter, a total of 117,059 scaffolds have been placed in one of the 21 pseudomolecules. These scaffolds represent 52.7% of the additional scaffolds assembled in Chapter 3 that were absent in the Chinese Spring reference genome.

4.3.5 PCA analysis of the SNP dataset

The 36.4 million SNP dataset was analysed chromosome by chromosome to generate a subset of SNPs that was representative of the entire dataset based on LD values of markers within a window of 50 Kbp. A total of 190,456 SNPs were selected as being in relative linkage equilibrium (LD ≤ 0.2) and a Principal Component Analysis (PCA) of this data set was performed.

Only the top four principal components (PCs) were analysed because they explained a variation larger than 5% and together explained 36.8% of the total SNP variation found in the subset (12.5%, 8.5%, 8.1%, and 7.7% for each PC respectively). As shown in Figure 4-7, most cultivars analysed were clustered together in close proximity regardless of the combination of principal components used to plot them. However, some cultivars did behave differently depending on the PCs used and were not clustered with the rest of cultivars. For example, when PC1 is used W7984 is the most divergent sample, whereas for PC2, it is Xi1 (Figure 4-7 and Figure 4-8). Similarly, when PC3 is used, cultivar Alsen appears to be the most divergent, whereas for PC4, it is Volcani the clearest outlier (Figure 4-7 and Figure 4-9). Figure 4-8 and Figure 4-9 have been enlarged to better appreciate this behaviour. These results provide evidence that at least 4 cultivars W7984, Xi'1, Alsen and Volcani had a different behaviour than the rest of the samples depending on the principal component selected for the analysis.

Figure 4-7. Principal component analysis of the subset of SNPs in the wheat pangenome. The top 4 principal components (eigenvectors) explain 36.8 % of the total SNP diversity found int the SNP dataset. Most of the samples clustered together in all the plots regardless of the combination of PCs used. However, the varieties that bahaved like outliers did depend on the combination of PCs plotted.

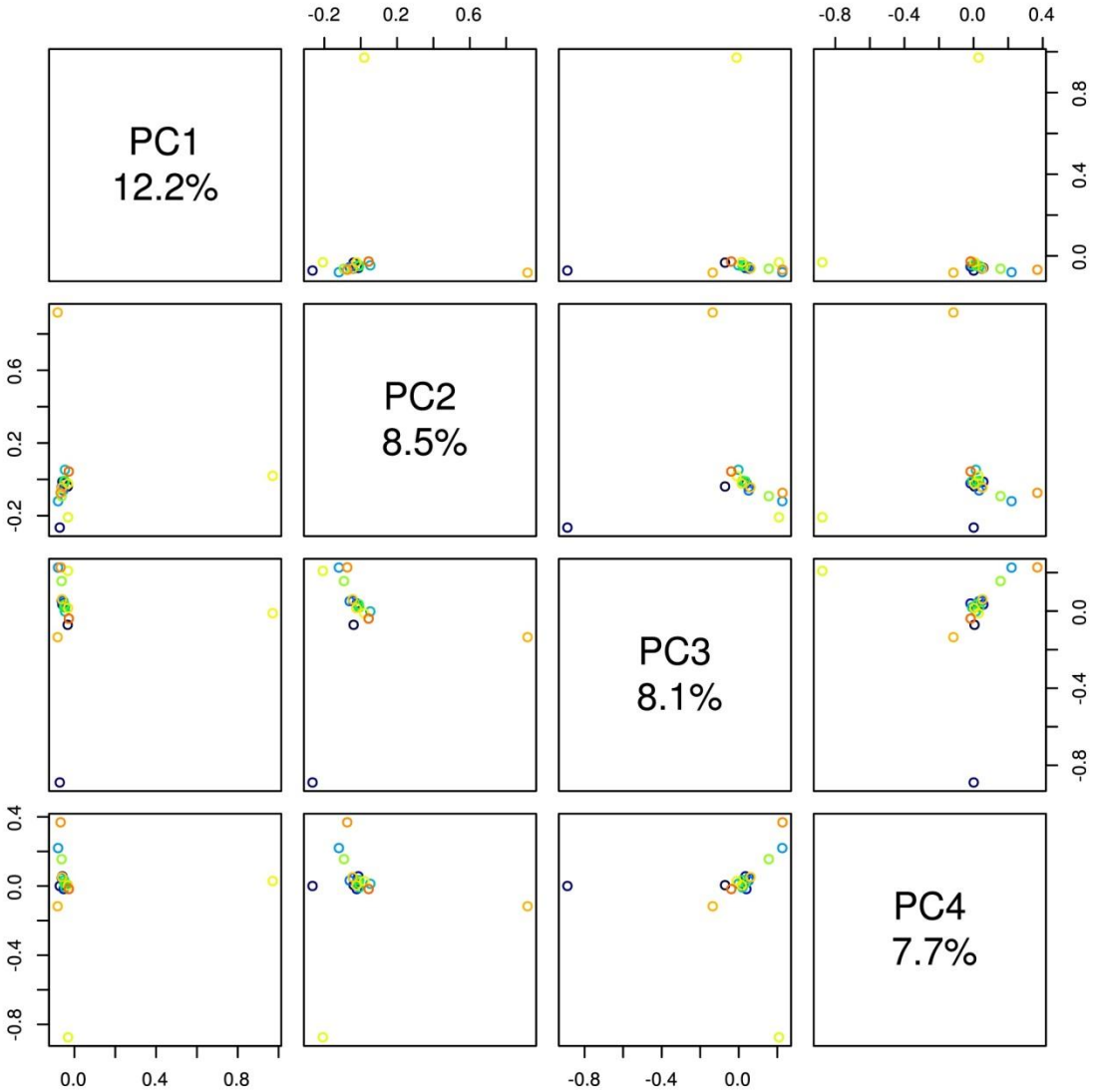


Figure 4-8. Plot of the top 2 principal components (eigenvectors) representing 21% of the total variance in the SNP dataset. All samples appear clustered except for W7984 and Xi-1 which appear to form separate cluster.

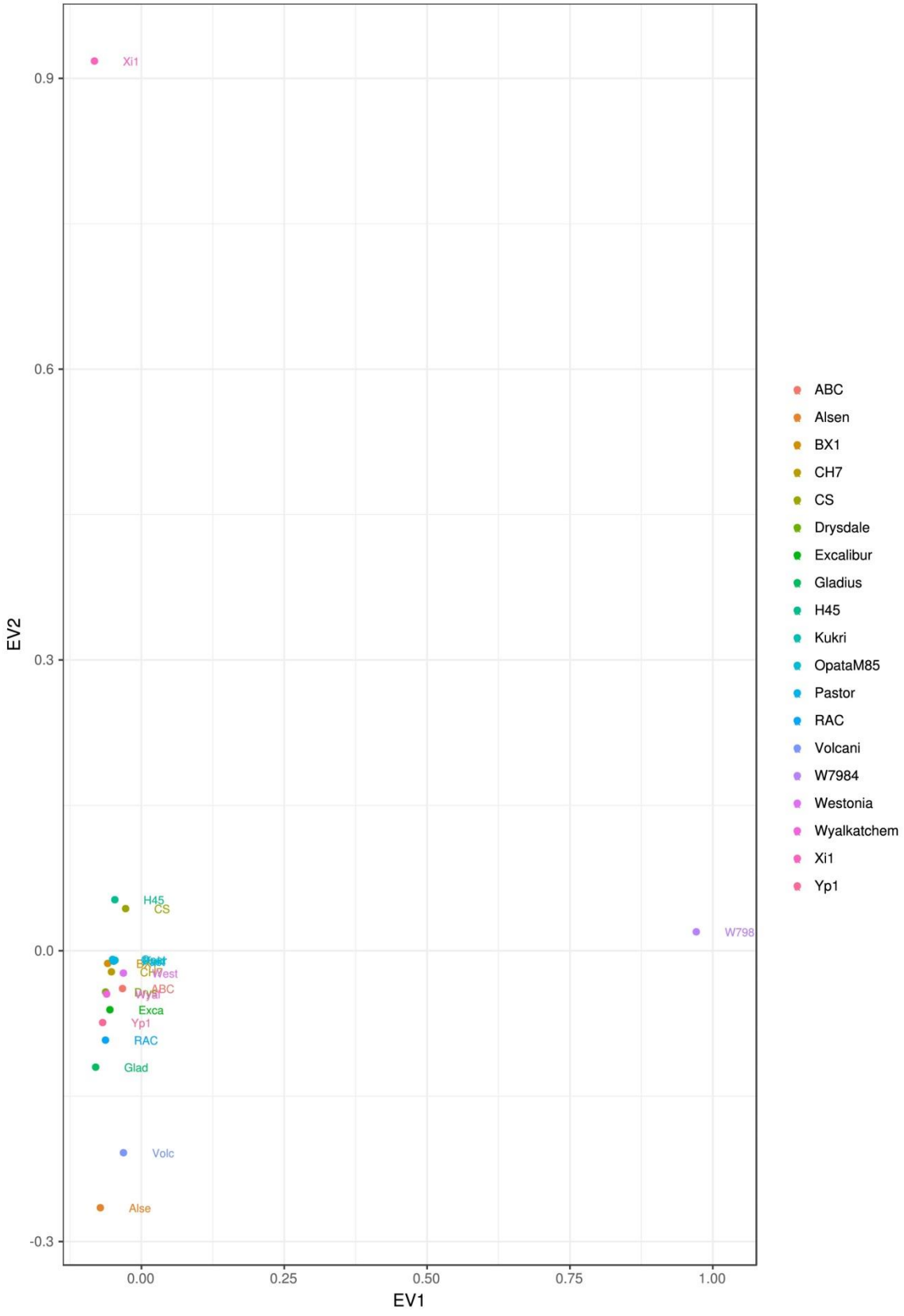
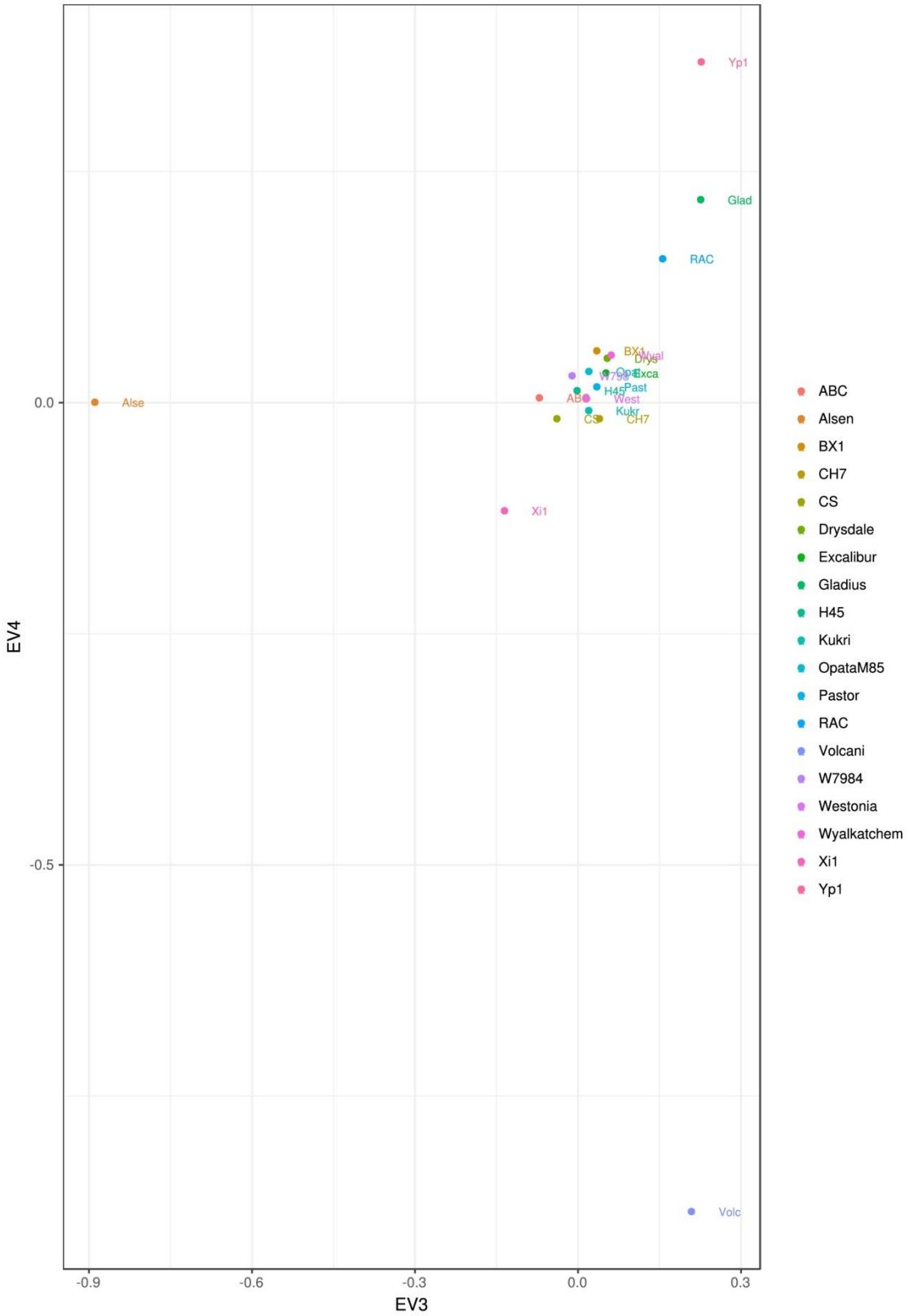


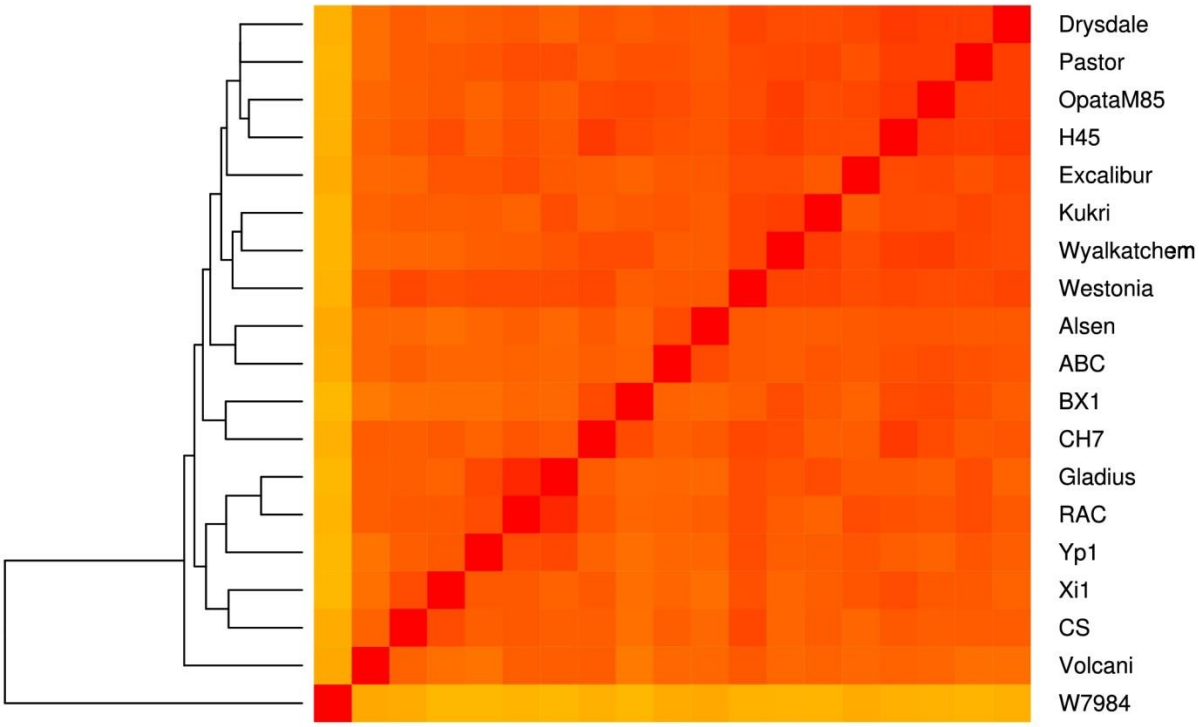
Figure 4-9. Plot of principal components 3 and 4 which explain 16.6% of the total diversity in the SNP dataset. Most of the samples appear clustered, but Volcani and Alsen, which appear to each form their own cluster away from the main cluster. This shows a different distribution and relationship between the samples.



4.3.6 Relatedness of wheat cultivars

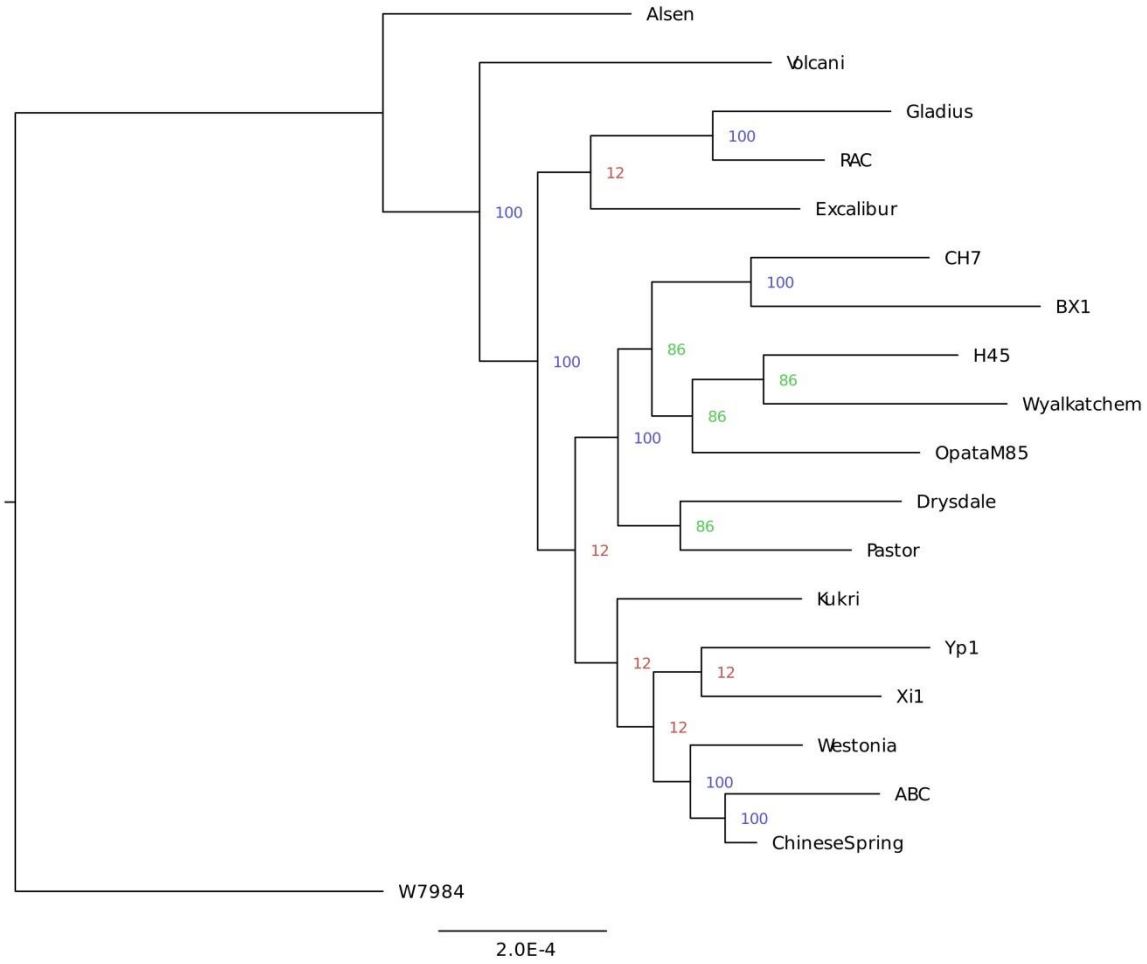
A dissimilarity matrix was produced from the pangenome SNP data and revealed pairwise distances that ranged from 0.12 between RAC and Gladius to 0.54 between BX-1 and W7984. Figure 4-10 shows a distance sorted heatmap and dendrogram inferred from the entire wheat pangenome SNP database. W7984 appears as the most divergent cultivar compared to all others with an average distance of 0.5 to all others samples compared to an average 0.27 between all other samples. Volcani is the second most divergent cultivar and is placed at the root of the internal clade.

Figure 4-10. Dissimilarity matrix and dendrogram of the 19 cultivars. The color scale goes from red = 0 distance (identical genotypes) to white = 1 distance (completely different genotypes). On the left side of the dissimilarity matrix there is a dendrogram produced by hierarchical clustering and neighbour joining. W7984 is the sample with the highest average distance from all other samples in this set (0.5).



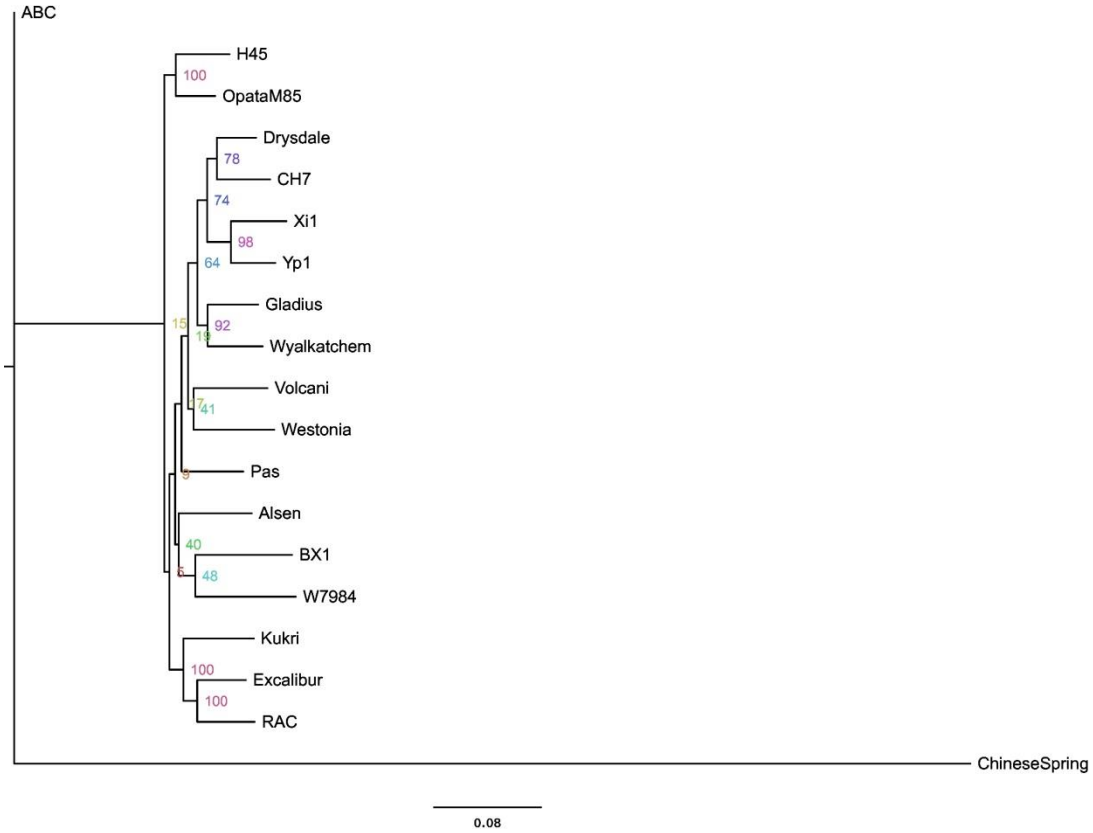
Based on the SNP data, 19 cultivar-specific genomes were generated and their specific coding sequences and protein sequences were extracted. The CDS of all genes were concatenated and the concatenated sequences were used for the construction of the phylogenetic tree shown in Figure 4-11. Cultivars W7984, Alsen and Volcani were the most divergent samples in the dataset in decreasing order of divergence. The remaining samples form a single monophyletic group. The internal monophyletic group was further subdivided in three groups: The Chinese Spring group which contains Kukri, Yp-1, Xi-1, Westonia, ABC and Chinese Spring; the Wyalkatchem group which contains CH7, BX1, H45, Wyalkatchem, OpataM85, Drysdale and Pastor; and the Gladius group which contains Gladius, RAC and Excalibur. Some nodes had little bootstrap support, particularly the node that splits the Wyalkatchem and Chinese Spring groups (bootstrap = 12).

Figure 4-11. Phylogenetic tree of the cultivar-specific CDS of the all genes. W7984 was found to be an outlier in this panel of wheat cultivars. All the rest form a single monophyletic clade which has Alsen at its root. Two additional monophyletic clades can be observed, the Chinese Spring clade and the OpataM85 clade.



Finally, a binary matrix of presence (1) - absence (0) variation of all genes in the wheat pangenome, was used to reconstruct an unrooted phylogenetic tree of the 19 cultivars. As shown in Figure 4-12, Chinese Spring appears at the root of the tree with a large distance to all others cultivars. This distance is supported by the large number of uniquely present and absent genes compared to all other 18 cultivars. The bootstrap values range from 5 to 100. All cultivars but Chinese Spring and ABC form a single monophyletic group, although the resolution of the internal nodes has little support from the bootstrap values.

Figure 4-12. Phylogenetic tree based on gene presence-absence variation. Chinese Spring appears as the most distant cultivar followed by ABC. The remaining cultivars form a single monophyletic group, although many of the internal nodes have little support from bootstrap values which range from 5 to 100.



4.4 Discussion

In this Chapter we have used all available whole genome shotgun data publicly available to identify over 36.4 million intervarietal SNPs across the entire wheat pangenome. This includes over 2 million SNPs identified in newly assembled contigs that were not present in the Chinese Spring genome. The polymorphic nature of the SNPs found was confirmed by comparison with a recently published 90K Infinium SNP array for wheat (Wang et al., 2014). The characteristics of the SNP distribution in the pangenome were explored. This is the largest SNP database available for wheat and it includes nearly 400K SNPs placed in coding sequences that could prove useful in marker assisted selection and genome wide association studies. These SNPs have been used to build a high density wheat genetic map that provided evidence for the placement and orientation of additional scaffolds of the pangenome into a reference sequence. Finally, the

relatedness and population structure of the 19 wheat cultivars was investigated based on the recently identified SNPs and gene PAVs found in the previous chapter.

4.4.1 Construction of a high quality pangenome-wide SNP database

Over 36.4 million polymorphic intervarietal SNPs were identified in the wheat pangenome sequence using the SGSautoSNP approach (Lorenc et al., 2012). This protocol was developed by Lorenc et al (2012) and was specifically designed for the identification of homozygous intervarietal polymorphic SNPs in complex genomes and with incomplete reference sequences with a SNP validation rate greater than 95%. Lorenc realized that the main challenge in SNP calling of complex and incomplete reference genomes is the presence of nearly identical sequences that have not or could not be properly represented in the genome sequence. The reads generated from these regions cannot be easily differentiated and could produce false heterozygous SNP calls (Hayward, 2012). This is particularly true for allopolyploid species in which homeologous or paralogous loci may have diverged little enough for the reads to be cross mapped. This pipeline was used by Lorenc et al (2012) in the discovery of over 800 thousand intervarietal SNPs in the homeologous group 7 of wheat (Lorenc et al., 2012) from whole genome shotgun data of four wheat cultivars. Lai et al (2015) further expanded the number of cultivars used for the discovery of SNPs in group 7 and identified over 4 million SNPs (Lai et al., 2015b). In this chapter, the original protocol was slightly modified to allow the use of a much larger reference. The original method used the Soap2 Aligner (Li et al., 2009c) which offered the option of keeping only uniquely mapping reads. Unfortunately, Soap2 was unable to handle reference sequences as large as the wheat pangenome. Therefore, the aligner Bowtie2 (Langmead and Salzberg, 2012) was used instead and additional filters were applied to ensure that only high quality alignments would be considered for SNP calling. Unpaired reads and read pairs with a mapping quality below 20 were removed from the alignments. Also, a base quality filter was included in the SNP calling step, where nucleotides below a quality value of 20 were not considered in the alignment. Previously, nucleotide quality had been ignored for genotyping due to the frequency of high quality erroneous nucleotide calls in the reads (Lorenc et al., 2012). The filter applied here aims only to remove low quality nucleotides that are expected to be erroneous and prevent the pipeline from ignoring otherwise homozygous SNPs. To ensure that SNPs identified with this modified approach were as good as those obtained following the original protocol, we validated them by comparing them to the 90K SNP infinium array (Wang et al., 2014). The validation found that 96.2% of the 13,541 common SNP loci were

polymorphic. This is slightly higher than the validation of the original protocol where Lai et al (2015) found 94% of the SNPs as polymorphic in the 90K Infinium array (Lai et al., 2015b). The fact that only 13,541 SNP loci from the Infinium array could be found in the dataset is a direct consequence of the SNP discovery methodology used by SGSautoSNP. Each probe in the Infinium array had an average of 6.3 hybridization sites and a median of 3 sites. The SGSautoSNP protocol explicitly discards reads with multiple equally scoring mapping positions which tend to occur in both homeologous or paralogous loci and nearly identical repeats and which might appear as false heterozygous SNPs due to the reads cross mapping between the loci or being forced into a collapsed repeat in the assembly. For this reason, many of the loci in the Infinium array may not have been picked up by SGSautoSNP.

The use of the wheat pangenome as the basis for SNP discovery has the added benefit of including sequences that would have been missed if a single cultivar reference had been used. The pangenome, not only increases the fraction of reads properly mapped to the reference as shown in Chapter 3, but can also be reliably used for genotyping more diverse wheat cultivars, landraces or wild tetraploid and diploid relatives to assess their utility in wheat breeding (Reif et al., 2005). Furthermore, the additional sequences present in the pangenome can improve the accuracy of the read mapping, by correctly mapping reads that otherwise would have been forced into a different locus causing a decreased sensitivity due to an increase of alignment conflicts that prevent the pipeline from calling SNPs.

The SNP density found in the wheat pangenome was similar to that found in the study of over 800 thousand and 4 million SNPs found in the homeologous group 7 of wheat where the number and density of SNPs in the A and B genomes were higher than in the D genome. (Lorenc et al., 2012, Lai et al., 2015b). Previous studies had shown a similar decreased diversity in the D genome compared to the A and B genomes. A study of 359 intronic SNPs in two diverse wheat panels found a ratio of polymorphic information content (PIC) between the A/B and the D genome of 1.7 and 1.9, highlighting the lower diversity found in the D genome (Chao et al., 2009). Similarly, 1,102 EST unigenes from 32 lines that included wild emmer, domesticated emmer, cultivated durum and aestivum germplasm were compared and 5,471 SNPs at 1,791 loci were found. Distribution of intravarietal SNPs in the 32 lines also found that the D genome of *T. aestivum* exhibited fewer haplotypes compared to the A and B genomes. Nevertheless, the number of haplotypes in the D genome of synthetic hexaploid wheats did not show a similar drop in

genetic diversity (Akhunov et al., 2010). Another study by Allen et al (2015) showed that the number of SNP markers in a framework genetic map of hexaploid wheat were in asymmetrical proportions of 30:50:17 for the A, B and D genomes respectively (Allen et al., 2011) which is also supported by our observation of a greater SNP density in the B genome compared to the A and D genomes (

Figure 4-1).

Comparison of RFLP marker diversity between the wild populations of *Ae. tauschii* and the D genome of hexaploid wheat shows that the former contains greater diversity (Caldwell et al., 2004) suggesting a limited gene flow between them. These results together with the description of pentaploid hybrids between hexaploid wheat and tetraploid wheat (Dvorak et al., 2006) and the low fertility rate observed between hexaploid wheat and diploid *Ae. tauschii* (Dvorak et al., 1998) led to the hypothesis that the low genetic diversity found in the D genome of hexaploid wheat is mostly driven by the little gene flow between the wild *Ae. tauschii* populations and the hexaploid wheat populations.

A more recent hypothesis tries to explain the differences in the diversity between the three subgenomes based on subgenome dominance after the polyploidization events (Pont et al., 2013). This hypothesis is based on the observation that there is a significant difference in the diversity content not only between the D and A/B genomes, but also between the A and B genomes, with B being the most diverse, followed by A and D. Pont et al (2013) found a similar pattern of genetic plasticity (B > A > D) after analysing paleo- and neo-polyploidization events using conserved orthologous sequences (COS) (Pont et al., 2013). In their hypothesis, after the first hybridization to form tetraploid wheat, the A genome took the dominant role over the B genome allowing greater freedom of mutation of the latter. Similarly, after the second polyploidization, the D genome took the dominant role allowing the A and B genome greater plasticity. This hypothesis is supported by the observation of biased genetic erosion during domestication (Cavanagh et al., 2013) an observation that has been supported by the study of metabolic networks in the wheat group 7 reference sequences (Berkman et al., 2013a). The differential SNP density found between the A and B genome could be explained by the genome dominance hypothesis, where the dominance of genome A in the original tetraploid restricted the plasticity of the A

genome and such effect is still visible in the hexaploids derived from those tetraploids. Now, the dominant role is played by the D genome and thus has a restricted plasticity compared to the other two.

The transition-transversion ratio found here (2.37) is higher than that reported by Lai et al (2015) (Lai, 2015) or by Winfield et al (2012) who found a Ts/Tv of 1.81 (Winfield et al., 2012), but it was similar to that found by Wang et al (2014) who found a Ts/Tv of 2.5 (72% transitions and 28% transversions) (Wang et al., 2014). Another interesting observation was the differential transition – transversion ratio found between the subgenomes which is preserved in the seven homeologous groups. The ratio is significantly lower in the D genome than in the A and B genomes and is concordant to previous findings by Lai et al (2015) which also observed a similar pattern in the analysis of homeologous group 7 (Lai et al., 2015b, Lai, 2015).

It has been suggested that the transition – transversion ratio can be considered an evolutionary footprint of methylation (Buckler and Holtsford, 1996), since transitions occur more often in highly methylated regions through the spontaneous deamination of 5-methylcytosine (Coulondre et al., 1978). Differential methylation between the subgenomes has been proposed to be a possible cause of the differential transition - transversion ratio (Lai, 2015). In that model, the A and B genomes which underwent two rounds of polyploidization, compared to one round for the D genome, would have accumulated more methylated cytosines, which are prone to mutation via deamination. Methylation remodelling has been reported to occur in synthetic allopolyploids shortly after hybridization in Brassicas (Lukens et al., 2006a, Xu et al., 2009), in *Spartina spp* (Salmon et al., 2005) and in wheat (Shaked et al., 2001). However, these changes include demethylation and *de novo* methylation and do not necessarily result in an overall increase in the methylation levels of the homeologous genomes (Shaked et al., 2001). Furthermore, a recent study of genome-wide methylation patterns in hexaploid wheat found similar levels of cytosine methylation in the three genomes with 69% of all methylated loci being equally methylated in all three genomes, 15% equally methylated in any two subgenomes and 16% were methylated in a single subgenome (Gardiner et al., 2015). Taken together, these studies do not support the idea that increased methylation occurs after polyploidization and thus the differential transition – transversion ratios between the subgenomes cannot be directly attributed to the extra polyploidization round of the A and B genomes.

The non-synonymous to synonymous ratio found in the wheat pangenome was close 1.4 (ranging from 1.2 to 1.6). Similar ratios have been reported previously in soybean where the authors found a whole genome ratio of 1.36 (Lam et al., 2010) and rice, where 55% of the SNPs in coding regions were non-synonymous (Arai-Kichise et al., 2011). Sequencing of 21 genes in 26 diverse wheat landraces and released cultivars showed that 50% of the genes did not contain any polymorphisms, but those that did, contained twice as many non-synonymous SNPs than synonymous SNPs (Ravel et al., 2006). The development of the 90K Infinium array for wheat revealed a much lower ratio where synonymous SNPs were far more abundant (Wang et al., 2014). However, validation of these SNPs has been relatively low (73%) compared to the SGSautoSNP pipeline which has a validation of 94% (Lai et al., 2015b, Lai, 2015, Lorenc et al., 2012).

The dominance of non-synonymous SNPs in some domesticated species has been proposed to be linked to relaxed selective constrains in the field compared to the wild (Lu et al., 2006). Human selection would interfere with natural selection in two ways: first by reducing the effectiveness of recombination in a population through inbreeding, and second by selecting for few traits of agronomical importance, while not selecting against deleterious mutations that may be part of the same selected genome. This effect has been studied in soybeans and helps explain the unexpected high rate of non-synonymous to synonymous SNPs in domesticated soybeans. The evolutionary history of hexaploid wheat fits this model well, as it appeared only as a domesticated hybrid and mostly reproduces via selfing, both of which reduce recombination effectiveness and increase the chances of maintaining deleterious mutations via relaxed selection constrains. Furthermore, its polyploid nature further decreases selection constrains on homeologous genes in such a way that complete deletion of homeologous loci has been reported in the early stages of newly synthesized wheats (Kashkush et al., 2002, Li et al., 2015a).

The use of the pangenome sequence for SNP discovery also allowed the characterization of the core and variable genes. It was observed that the variable gene set contains a higher SNP density mainly driven by a higher ratio of non-synonymous SNPs (Figure 4-4). Similar results have been observed in the *Brassica oleracea* pangenome (Golikz, 2016, Golicz et al., 2016b), the soybean pangenome (Li et al., 2014c) and rice (Yao et al., 2015, Li et al., 2014a, The 3000 Genome Project, 2014). It has been proposed that weaker purifying selection or greater positive selection may be the main determinants of the patterns and distribution of SNPs in the variable genome. In hexaploid wheat most of the sequence diversity appeared through gene flow from close relatives (Reif et al.,

2005, Haudry et al., 2007, Akhunov et al., 2010) so it is more likely that a weaker purifying selection affected the majority of genes and that greater positive selection only affected genes directly linked with specific traits. Variable genes, which appear to be enriched with functions associated with responses to biotic and abiotic stress (Figure 3.5) are usually more diverse than other functional categories (Tatarinova et al., 2016) and this diversity may have been kept in the wheat germplasm because it conferred important traits like disease resistance and climate adaptability.

4.4.2 A high quality framework genetic map of the pangenome allows the anchoring of novel sequences to a pseudomolecule

In this chapter, the utility of pangenome-wide SNPs is shown by the construction of a high density genetic linkage map and its use in the placement and orientation of newly assembled contigs into pseudomolecules. The framework genetic map constructed with nearly 110 thousand high quality polymorphic SNPs produced 21 large linkage groups as expected in hexaploid wheat (n = 21). An assessment of the origin of the SNPs in every linkage group (LG) showed that each LG was enriched with SNPs from a single chromosome assembly (**Figure 4-6**). Additionally, an assessment of the distribution of the SNPs across the 21 LGs showed that most of the SNPs from a chromosome assembly were concentrated in single linkage groups except for SNPs in newly assembled contigs (

Figure **4-5**). The Chinese Spring reference genome was constructed in a chromosome-wise fashion (see Chapter 2) using chromosome sorted DNA for the library preparation and sequencing (Mayer et al., 2014). The fact that the SNPs in these chromosome assemblies were clustered together and assigned to the same linkage groups confirms the accuracy of the genetic map constructed in this chapter and its possible use in map-based gene landing, QTL analysis and chromosome anchoring. Furthermore, the SNPs identified in newly assembled contigs that were absent from the Chinese Spring genome (Chapter 3) were evenly distributed across all linkage groups, suggesting that their contigs of origin were equally evenly distributed across the genome.

The SynOpDH population had been previously used in the assembly of the synthetic wheat W7984 genome (Chapman et al., 2015). In their study, Chapman et al (2015) used the “Bubblecluster” algorithm (Strnadova V, 2014) to quickly assign linkage groups to millions of SNPs and then used the POPSEQ approach (Mascher et al., 2013) to assign positions to those contigs in the pseudomolecules. The framework genetic map produced for that study contained 112 thousand SNPs and had a total length of 2,826cM in 1,335

recombination bins. That is smaller than the genetic map constructed in this chapter and also smaller than many other genetic maps produced for several wheat mapping populations (Song et al., 2005, Quarrie et al., 2005, Semagn et al., 2006, Li et al., 2007, Xue et al., 2008, Poland et al., 2012b, Torada et al., 2006, Wu et al., 2015, Li et al., 2015c, Li et al., 2015b, Wang et al., 2014). Overall, these maps had a length ranging from 3,000cM to 4,500cM in 1,500 to 2,500 recombination bins. The genetic map generated in this chapter is 8,437cM long in 4,532 recombination bins, nearly twice as large as those previously constructed.

The length of a genetic map is affected by different factors including the size of the mapping population (Ferreira et al., 2006), the density of the markers or their distribution along the genome. The different map sizes found between the study by Chapman and the map presented in this chapter can be explained by the combined effects of a smaller population size which reduces the chance of discovery of more recombination events between the markers and distortion of the marker distribution. Chapman et al (2015) used only 78 of the 90 individuals of the population due to evidence of large-scale deletions in 12 samples, whereas the current map was constructed with information from all 90 individuals. The inclusion of individuals with these deletions affected neither the accuracy nor the resolution of the genetic map. Even though the number of SNPs used in the construction of both maps was similar, Chapman et al (2015) selected the SNPs based on one condition: these should be present in scaffolds with 3 or more co-segregating SNPs. This selection may have affected the distribution of the SNPs used for the construction genetic map by using SNPs that were more closely clustered in the genome and were thus less prone to recombination. In contrast, the SNPs selected for this chapter were chosen on the condition of 0 missing data points. This condition was achieved first by the imputation of missing alleles based on the surrounding known alleles in a single contig and second, by merging highly similar consecutive SNPs in the same contig into consensus metaSNPs. Both steps, greatly decrease the amount of missing data in the SNP dataset.

The use of SNPs more evenly spread across the genome increases the chances of detecting recombination events that would not be detected with a more skewed SNP sampling. It is also possible that the larger map length is due to oversampling of recombination hotspots which lead to an overestimation of genetic distances between the SNPs. However, the fact that both maps have a similar density of 1 recombination bin every 2 cM suggests that this is not the case.

4.4.2.1 Use of the genetic map for anchoring new scaffolds

The genetic map was used for the anchoring of additional sequences assembled in the previous chapter to specific positions in the chromosomes. This was achieved through a protocol proposed here that attempts to balance a high accuracy of placement while including as many contigs as possible. The guiding principle is that all SNPs in a contig provide clues about the orientation and placement of the contig, but markers with fewer missing data points are more accurate and thus should have a greater effect in the calculation of the final position and orientation of the contig. That is why the first step is the calculation of a weighted position in the genetic map. The second step attempts to remove contigs with unusually long genetic distances between their SNPs. If a single contig occupies more than one recombination bins and spanned the space of more than one contigs in the genetic map, then it was likely that the contig was either incorrectly assembled or incorrectly genotyped. Either way it could not be reliably anchored in a specific position of the genetic map. The final step uses the coherence between the genetic and the physical position of SNPs within a contig to orient the contigs in the genetic map. This step relies in relative position of SNP in the contig and in the genetic map. All the SNPs in a single contig should either belong to the same recombination bin or should have a similar order in both the genetic map and the contig. If the order was different, the contigs were not oriented within the pseudomolecule.

This approach allowed us to anchor an additional 50,000 scaffolds which represents approximately 25% of the newly assembled scaffolds of the pangenome. Because the genetic map is based on the SynOpDH population which was produced from OpataM85 and W7984, it is unlikely that it could be used to anchor scaffolds that were not present in either of those cultivars. Removing the 11,199 scaffolds unique for OpataM85 and W7984, the remaining 40K scaffolds anchored to the reference pangenome were also present at least in one cultivar other than the parental varieties and represent a 17% of all the newly assembled scaffolds of the pangenome. Given the accuracy of the genetic map used for placement of these scaffolds, it is unlikely that there are many misplacements in the current version of the pangenome assembly.

A similar approach, POPSEQ, has been proposed and successfully used in complex genomes including barley (Mascher et al., 2013) and wheat (Edae et al., 2015, Chapman et al., 2015). It also relies on the construction of a framework genetic map for placement of the contigs, but it does not attempt to order them within the pseudomolecules. The

POPSEQ approach requires the use of high quality SNPs with very few missing data points per SNP and several cosegregating SNPs per contig for placement. In an assembly as fragmented as the one presented in Chapters 2 and 3, such an approach would ignore the great majority of the contigs with SNP information, since the rate of missing data is high and due to their small size, they have few SNPs. The high rate of missing data in the population is due to the low coverage of the samples which, after read mapping averaged 0.5X. Moreover, the imputation step and the consensus calculation requires at least 3 and 2 SNPs per contig to decrease the amount of missing data and that condition was not met by the majority of contigs which contain one single SNP.

4.4.3 W7984 is the most diverged cultivar in the dataset

The pangenome SNP data was used to estimate the structure and relatedness of the samples. Principal Component Analysis (PCA) showed that there was not much diversity between the samples as is expected in wheat (Huang et al., 2002b, Cavanagh et al., 2013). However, some cultivars exhibited quite different behaviours from the others (W7984, Xi-1, Alsen and Volcani) depending on the principal component investigated. While this may suggest the existence of stratification in the sample, the small sample size prevents us from drawing a definitive conclusion (Figure 4-7 and Figure 4-9).

In order to better understand the relatedness of the samples, cultivar-specific transcriptomes were produced for all genes in the pangenome and a phylogenetic tree was constructed (Figure 4-11). The maximum likelihood tree showed that W7984 was the most divergent of the cultivars which is consistent with the results from PC1 of the PCA which showed that W7984 behaved differently (Figure 4-8). Also, sample Volcani and Alsen were placed as outgroups in the phylogenetic tree which is consistent with the results found in the PCA analysis. The phylogenetic tree was also concordant with a dendrogram constructed from a dissimilarity matrix using all the SNPs identified in this chapter (Figure 4-10) which showed that W7984 had the highest number of unique alleles followed by Volcani which was the second most divergent genotype in the dendrogram. Overall, the three analysis performed (PCA, dissimilarity matrix and phylogenetic analysis) coincided in placing W7984, Alsen and Volcani as the most diverged samples in the dataset with W7984 being the most divergent of the three in every case. Cultivar Xi-1 was also found to be divergent in the PCA (PC2). However, no other analysis supports its placement as a divergent genotype. It is possible that the analysis performed in this chapter were unable to assess the characteristics that make Xi-1 stand out in the PCA and

that, despite its lack of support it can still be considered as divergent from mainstream elite cultivars.

Cultivar W7984 is a synthetic allohexaploid that has been widely used in the construction of genetic maps due to its high genetic distance from most common elite breeding material (Anderson et al., 1993). It also lacks the 1RS.1BL rye-wheat translocation, which is common in wheat elite cultivars but prevents recombination between the rye-derived segment (1RS) and the homologous chromosome arms of wheat (1AS, 1BS and 1DS) causing a depletion of segregating markers in that locus (Sorrells et al., 2011). Synthetic wheat cultivars are obtained by the hybridization of tetraploid “durum” wheat and different accessions of *Ae. tauschii* followed by induction of genome duplication of the amphiploids and are usually used to increase the diversity of the D genome in wheat breeding programs (Warburton et al., 2006). It has been shown that synthetic wheat cultivars are genetically distant from most elite wheat cultivars and contain a higher nucleotide diversity (Akhunov et al., 2010) which is why these are being increasingly used in breeding programs for introgression of genes from wild germplasm into elite cultivars (Mujeeb-Kazi et al., 2008, Warburton et al., 2006, del Blanco et al., 2001).

In contrast to the PCA, dissimilarity matrix and phylogenetic analysis of the 19 cultivars, a phylogenetic reconstruction based on gene presence-absence variation in the pangenome showed that Chinese Spring was the most distant cultivar in this set. This result is primarily the consequence of the large number of pangenome genes absent in the Chinese Spring germplasm. These results can be understood as a consequence of the origin of Chinese Spring which, despite being widely used as background in the construction of wheat cytogenetic stocks, has not been readily used in breeding programs due to its sensitivity to many pathogens and lack of resistance to many abiotic stresses (Sears and Miller, 1985). Also, Chinese Spring belongs to the group of wheat landraces which have been shown to be genetically different from and more diverse than modern cultivars (Cavanagh et al., 2013).

However, the phylogenetic tree constructed based on genes treated as binary state characteristics (presence/absence), do not reflect the true genetic variation within each gene and thus cannot be considered a representative reconstruction of the relations between the cultivars. Most of these genes occur in a multistate fashion in the wheat germplasm that is not represented in the PAV table with multiple haplotypes present in the accessions studied as shown in the cultivar-specific transcriptomes reconstructed for

them. Thus, a combination of the results from both approaches would better describe the relationship of the samples included in this study, where both W7984 and Chinese Spring appear to be the most divergent and the remaining Australian cultivars are closer to each other. Including a wild sample as an outgroup on both analysis would help understand the direction of the evolution in this dataset.

5 Chapter 5: Summary and outlook

This thesis describes the construction, annotation and analysis of the wheat pangenome based on a novel Chinese Spring reference assembly and the sequence data for 18 diverse wheat cultivars. The utility of the pangenome sequence was demonstrated by the identification of core and variables genes, their characterization with more than 36.4 million intervarietal SNPs spread across the pangenome reference, the construction of a high density genetic map and the assessment of the genetic relatedness of the 19 varieties included in this study. In many ways, the methods used in this thesis were conditioned by the amount and quality of the data publicly available and could be improved with emerging sequencing technologies and novel analysis algorithms.

5.1 Limitations of the current wheat pangenome

The genome assembly produced in Chapter 2 that served as the starting point for the construction of the pangenome contains a large fraction of the unique sequences present in the wheat genome and a large portion of the universal single copy orthologs and core eukaryotic genes. Nevertheless, due to its high level of fragmentation, accurate prediction of genes is challenging. This was confirmed by the detection of split gene models based on non-overlapping sequence identity to known proteins in the *T. urartu* genome (Chapter 2). Such fragmentation may also affect the mapping efficiency of reads, although such effect was not evident when mapping reads from the 19 wheat cultivars to the pangenome (Figure 3-3).

Another limitation of the current pangenome is the lack of information on the accurate placement of sequences into the pseudomolecules. Overall, 50% of the additional sequences were placed, but the remaining sequences are yet to be placed in their right position. Also, the current position of the scaffolds is only an approximation given that both methods used for placement (mate-pair information and genetic mapping) cannot point to the exact position where the new sequences should be inserted. Similarly, many genes are yet to be placed into the pseudomolecules, due to the lack of information linking their contig of origin to other markers already placed in the pseudomolecules.

Finally, the catalogue of genes found in the pangenome may be missing some rare or cultivar-specific genes that failed to be assembled due to the low sequencing depth of the samples. A recent study of variable gene content in the wheat transcriptome showed

that 32.1% of the genes absent in the Chinese Spring were present in the current pangenome (Liu et al., 2016). From the remaining 68% of the transcripts, half were homologous to known grass genes that had not been captured by the pangenome and the rest did not show homology to any known gene.

5.2 Impact of new genome assemblies

This study was started shortly after the publication and release of the first draft assembly of the wheat genome (IWGSC, 2014). The quality of the IWGSC reference genome prompted us to reassemble the wheat genome using an approach that had been previously successful in the assembly of group 7 chromosomes (Berkman et al., 2013a). In late 2015, the IWGSC announced the completion of a new draft assembly using whole genome shotgun reads by the TGAC team and this became available in mid 2016 (Clavijo et al., 2017). This assembly included a new genome annotation based on PACBIO long reads to produce full-length cDNA sequences that could be mapped to the reference genome. However, the raw data generated for that assembly has not yet been released. At the same time, another genome assembly was announced by the IWGSC based on the De novoMagic assembler from NRgene (NRgene, 2017). Unfortunately, this assembly was not publicly available at the time of this study and the details of the assembly protocol are still unknown.

Both of the new assemblies show improved metrics, with N50 exceeding the 100Kb mark and they also exhibit high collinearity which was confirmed by alignment to the previously assembled chromosome 3B (Paux et al., 2008). These improved metrics along with better sequencing technology for transcriptome data, have greatly improved the quality and accuracy of gene models. Their use could help improve the accuracy of gene PAV calls, increase the number of scaffolds anchored to the pseudomolecules and reduce the amount unmapped reads that need to be assembled to expand the pangenome sequence.

5.3 Impact of third generation sequencing technologies

The use of third generation sequencing (TGS) technologies could also have a great impact on the accuracy and contiguity of new genome assemblies and could help improve existing assemblies. The development of more accurate TGS platforms such as the PacBio SMRT system (McCarthy, 2010) and Oxford nanopore technologies (Eisenstein, 2012, Mikheyev and Tin, 2014) and recent advances in *de novo* assembly algorithms that

take advantage of these long reads (Vaser et al., 2017, Goodwin et al., 2015, Phillippy, 2017) have already resulted in the production of nearly complete genome sequences with phased haplotypes (Zimin et al., 2017a, Mochida et al., 2017) and improvements on previously released reference genomes (Zimin et al., 2017b). These new technologies open the possibility of directly assembling and placing highly diverged sequences directly onto a reference genome producing a more complete pangenome. Furthermore, these new TGS technologies are already being used to sequence and assemble full length transcripts from cDNA libraries and to explore the hidden isoform diversity that cannot be directly accessed through the traditional RNA-seq approaches (Cartolano et al., 2016, Abdel-Ghany et al., 2016). These developments will increase the accuracy and detail of current genome annotations which will have a major impact in the identification of gene presence-absence variations and understanding the effects of SNPs on isoform dominance in the genome.

Longer reads have been recently used in the identification of DNA modifications, particularly methylation (Simpson et al., 2017, Rand et al., 2017). Current strategies for detecting DNA methylation rely on the bisulphite treatment of DNA molecules to convert unmethylated cytosine into uracil and then into thymine through PCR (Frommer et al., 1992, Shapiro et al., 1970, Hayatsu et al., 1970). The coupling of bisulphate treatment and NGS led to the development of genome-wide methylation surveillance methods like whole genome bisulphite sequencing (WGBS) (Cokus et al., 2008) and reduced representation bisulphite sequencing (RRBS) (Meissner et al., 2005). However, these techniques are unable to detect any modification other than 5-methylcytosine like 5-formilcytosine, 5-hydroxymethylcytosine, 4-methylcytosine or 6-methyladenosine which have been shown to contribute to gene expression control in bacteria and eukaryotes (Meyer and Jaffrey, 2016, Koziol et al., 2015, Greer et al., 2015, Fu et al., 2015, Tahiliani et al., 2009, Ratel et al., 2006). In addition to that limitation, the generation of short reads after bisulphite conversion causes a higher level of ambiguity in the mapping stage which decreases the mapping efficiency and is unable to provide accurate information for some regions in the genome. Both difficulties can be overcome by the use of TGS technologies, which can reduce the mapping ambiguity because the reads are not being converted and are able to detect different types of modifications to individual DNA bases. These improvements will allow us to assess the effect of DNA modifications on the distribution and expression of variable genes in the pangenome.

5.4 Impact of improved digital representation

Currently, pangenomes are represented as a linear collection of core and variable genes, distributed along the hypothetical representation of a chromosome (pseudomolecule). However, such linear representation cannot accurately display the complex structural variations that occur between individuals of the same species (Marcus et al., 2014). Furthermore, the linear representation may negatively affect the mapping efficiency and thus hide genetic diversity present in the unmapped reads. Marcus et al (2014), proposed the use of *deBruijn* graphs for the representation of the closed pangenome of *Bacillus anthracis* (Marcus et al., 2014) with the tool splitMEM. This representation has been improved (Baier et al., 2016) and is growing more accepted in the scientific community (Sheikhizadeh et al., 2016). Furthermore, the *deBruijn* graph data structure for the storage and representation of the pangenome has been extended to allow variant analysis and pattern searching (Beller and Ohlebusch, 2016) which are the first steps towards accurate pangenome read mapping. Also, the rice pangenome project have released an online browser to mine genomic information for over 3000 rice varieties and include information such as presence-absence polymorphism, SNPs, geographical distribution, haplotype, functional annotation and more (http://cgm.sjtu.edu.cn/3kricedb/index.php).

In the near future, due to technical and methodological advances, genomic research will move away from the single genome reference paradigm into the pangenome paradigm. Such a move is currently ongoing with multiple pangenome projects being designed and developed for different species including many crop plants. We hope that the present study will contribute to the wider landscape of pangenomic studies and help clear the way for more ambitious studies on genomics of crop plants.

6 References

ABBERTON, M., BATLEY, J., BENTLEY, A., BRYANT, J., CAI, H., COCKRAM, J., COSTA DE OLIVEIRA, A., CSEKE, L. J., DEMPEWOLF, H., DE PACE, C., EDWARDS, D., GEPTS, P., GREENLAND, A., HALL, A. E., HENRY, R., HORI, K., HOWE, G. T., HUGHES, S., HUMPHREYS, M., LIGHTFOOT, D., MARSHALL, A., MAYES, S., NGUYEN, H. T., OGBONNAYA, F. C., ORTIZ, R., PATERSON, A. H., TUBEROSA, R., VALLIYODAN, B., VARSHNEY, R. K. & YANO, M. 2015. Global agricultural intensification during climate change: a role for genomics. *Plant Biotechnology Journal,* 14**,** 1095-1098.

ABDEL-GHANY, S. E., HAMILTON, M., JACOBI, J. L., NGAM, P., DEVITT, N., SCHILKEY, F., BEN-HUR, A. & REDDY, A. S. N. 2016. A survey of the sorghum transcriptome using single-molecule long reads. 7**,** 11706.

ADAMS, K. L. & WENDEL, J. F. 2005. Polyploidy and genome evolution in plants. *Curr Opin Plant Biol,* 8.

ADRIAN ALEXA, J. R. 2006. topGO: Enrichment Analysis for Gene Ontology.

AIR, G., SANGER, F., BARRELL, B., BROWN, N., COULSON, A., FIDDES, J., HUTCHISON, C., SLOCOMBE, P. & SMITH, M. NUCLEOTIDE-SEQUENCE OF DNA OF BACTERIOPHAGE-PHIX174. PROCEEDINGS OF THE AUSTRALIAN BIOCHEMICAL SOCIETY, 1977. PROC AUST BIOCHEMICAL SOC MONASH UNIV DEPT BIOCHEMISTRY, CLAYTON VICTORIA 3168, AUSTRALIA, 60-60.

AKBARI, M., WENZL, P., CAIG, V., CARLING, J., XIA, L. & YANG, S. 2006. Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. *Theor Appl Genet,* 113.

AKHUNOV, E. D., AKHUNOVA, A. R., ANDERSON, O. D., ANDERSON, J. A., BLAKE, N. & CLEGG, M. T. 2010. Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. *BMC Genomics,* 11.

AKHUNOV, E. D., SEHGAL, S., LIANG, H., WANG, S., AKHUNOVA, A. R. & KAUR, G. 2013. Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. *Plant Physiol,* 161.

AL-OKAILY, A. A. 2016. HGA: de novo genome assembly method for bacterial genomes using high coverage short sequencing reads. *BMC Genomics,* 17**,** 193.

ALLEN, A. M., BARKER, G. L. A., BERRY, S. T., COGHILL, J. A., GWILLIAM, R., KIRBY, S., ROBINSON, P., BRENCHLEY, R. C., D'AMORE, R., MCKENZIE, N., WAITE, D., HALL, A., BEVAN, M., HALL, N. & EDWARDS, K. J. 2011. Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (Triticum aestivum L.). *Plant Biotechnology Journal,* 9**,** 1086-1099.

ALLOUIS, S., MOORE, G., BELLEC, A., SHARP, R., RAMPANT, P. F., MORTIMER, K., PATEYRON, S., FOOTE, T. N., GRIFFITHS, S., CABOCHE, M. & CHALHOUB, B. 2003. Construction and characterisation of a hexaploid wheat (Triticum aestivum L.) BAC library from the reference germplasm 'Chinese Spring'. *Cereal Research Communications,* 31**,** 331-338.

ALONSO-BLANCO, C., ANDRADE, J., BECKER, C., BEMM, F., BERGELSON, J., BORGWARDT, K. M., CAO, J., CHAE, E., DEZWAAN, T. M., DING, W., ECKER, J. R., EXPOSITO-ALONSO, M., FARLOW, A., FITZ, J., GAN, X., GRIMM, D. G., HANCOCK, A. M., HENZ, S. R., HOLM, S., HORTON, M., JARSULIC, M., KERSTETTER, R. A., KORTE, A., KORTE, P., LANZ, C., LEE, C.-R., MENG, D., MICHAEL, T. P., MOTT, R., MULIYATI, N. W., NÄGELE, T., NAGLER, M., NIZHYNSKA, V., NORDBORG, M., NOVIKOVA, P. Y., PICÓ, F. X., PLATZER, A., RABANAL, F. A., RODRIGUEZ, A., ROWAN, B. A., SALOMÉ, P. A., SCHMID, K. J., SCHMITZ, R. J., SEREN, Ü., SPERONE, F. G., SUDKAMP, M., SVARDAL, H., TANZER, M. M., TODD, D., VOLCHENBOUM, S. L., WANG, C., WANG, G., WANG, X., WECKWERTH, W., WEIGEL, D. & ZHOU, X. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell,* 166**,** 481-491.

ANDERSON, J. A., CHURCHILL, G. A., AUTRIQUE, J. E., TANKSLEY, S. D. & SORRELLS, M. E. 1993. Optimizing parental selection for genetic linkage maps. *Genome,* 36**,** 181-186.

ANDREWS, S. *FastQC A Quality Control tool for High Throughput Sequence Data* [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [Accessed].

ARAI-KICHISE, Y., SHIWA, Y., NAGASAKI, H., EBANA, K., YOSHIKAWA, H., YANO, M. & WAKASA, K. 2011. Discovery of Genome-Wide DNA Polymorphisms in a Landrace Cultivar of Japonica Rice by Whole-Genome Sequencing. *Plant and Cell Physiology,* 52**,** 274-282.

ARCHER, J., BAILLIE, G., WATSON, S. J., KELLAM, P., RAMBAUT, A. & ROBERTSON, D. L. 2012. Analysis of high-depth sequence data for studying viral diversity: a

comparison of next generation sequencing platforms using Segminator II. *BMC Bioinformatics,* 13**,** 47.

ASHBY, M., EKHOLM, J. & HICKEY, L. 2017. Sequencing for Structural Variation and Isoform Discovery. *Genetic Engineering & Biotechnology News,* 37**,** 16-17.

ASSENG, S., EWERT, F., MARTRE, P., ROTTER, R. P., LOBELL, D. B., CAMMARANO, D., KIMBALL, B. A., OTTMAN, M. J., WALL, G. W., WHITE, J. W., REYNOLDS, M. P., ALDERMAN, P. D., PRASAD, P. V. V., AGGARWAL, P. K., ANOTHAI, J., BASSO, B., BIERNATH, C., CHALLINOR, A. J., DE SANCTIS, G., DOLTRA, J., FERERES, E., GARCIA-VILA, M., GAYLER, S., HOOGENBOOM, G., HUNT, L. A., IZAURRALDE, R. C., JABLOUN, M., JONES, C. D., KERSEBAUM, K. C., KOEHLER, A. K., MULLER, C., NARESH KUMAR, S., NENDEL, C., O/'LEARY, G., OLESEN, J. E., PALOSUO, T., PRIESACK, E., EYSHI REZAEI, E., RUANE, A. C., SEMENOV, M. A., SHCHERBAK, I., STOCKLE, C., STRATONOVITCH, P., STRECK, T., SUPIT, I., TAO, F., THORBURN, P. J., WAHA, K., WANG, E., WALLACH, D., WOLF, J., ZHAO, Z. & ZHU, Y. 2015. Rising temperatures reduce global wheat production. *Nature Clim. Change,* 5**,** 143-147.

BACHLAVA, E., TAYLOR, C. A., TANG, S., BOWERS, J. E., MANDEL, J. R., BURKE, J. M. & KNAPP, S. J. 2012. SNP discovery and development of a high-density genotyping array for sunflower. *PLoS One,* 7.

BADAEVA, E. D., DEDKOVA, O. S., GAY, G., PUKHALSKYI, V. A., ZELENIN, A. V., BERNARD, S. & BERNARD, M. 2007. Chromosomal rearrangements in wheat: their types and distribution. *Genome,* 50**,** 907-926.

BAETS, G., DURME, J., REUMERS, J., MAURER-STROH, S., VANHEE, P. & DOPAZO, J. 2012. SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res,* 40.

BAIER, U., BELLER, T. & OHLEBUSCH, E. 2016. Graphical pan-genome analysis with compressed suffix trees and the Burrows–Wheeler transform. *Bioinformatics,* 32**,** 497-504.

BANKEVICH, A., NURK, S., ANTIPOV, D., GUREVICH, A. A., DVORKIN, M., KULIKOV, A. S., LESIN, V. M., NIKOLENKO, S. I., PHAM, S., PRJIBELSKI, A. D., PYSHKIN, A. V., SIROTKIN, A. V., VYAHHI, N., TESLER, G., ALEKSEYEV, M. A. & PEVZNER, P. A. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology,* 19**,** 455-477.

BARBARA, T., PALMA-SILVA, C., PAGGI, G. M., BERED, F., FAY, M. F. & LEXER, C. 2007. Cross-species transfer of nuclear microsatellite markers: potential and limitations. *Molecular Ecology,* 16**,** 3759-3767.

BARTOS, J., PAUX, E., KOFLER, R., HAVRANKOVA, M., KOPECKY, D., SUCHANKOVA, P., SAFAR, J., SIMKOVA, H., TOWN, C. D., LELLEY, T., FEUILLET, C. & DOLEZEL, J. 2008. A first survey of the rye (Secale cereale) genome composition through BAC end sequencing of the short arm of chromosome 1R. *BMC Plant Biol,* 8**,** 95.

BATLEY, J. & EDWARDS, D. 2007. SNP Applications in Plants. *In:* ORAGUZIE, N. C., RIKKERINK, E. H. A., GARDINER, S. E. & DE SILVA, H. N. (eds.) *Association Mapping in Plants.* New York, NY: Springer New York.

BATLEY, J. & EDWARDS, D. 2016. The application of genomics and bioinformatics to accelerate crop improvement in a changing climate. *Current Opinion in Plant Biology,* 30**,** 78-81.

BEKELE, W. A., WIECKHORST, S., FRIEDT, W. & SNOWDON, R. J. 2013. High-throughput genomics in sorghum: from whole-genome resequencing to a SNP screening array. *Plant Biotechnology Journal,* 11**,** 1112-1125.

BELLER, T. & OHLEBUSCH, E. 2016. A representation of a compressed de Bruijn graph for pan-genome analysis that enables search. *Algorithms for Molecular Biology,* 11**,** 20.

BELÓ, A., BEATTY, M. K., HONDRED, D., FENGLER, K. A., LI, B. & RAFALSKI, A. 2009. Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theoretical and Applied Genetics,* 120**,** 355.

BENNETT, M. D. 1972. Nuclear DNA Content and Minimum Generation Time in Herbaceous Plants. *Proceedings of the Royal Society of London. Series B. Biological Sciences,* 181**,** 109-135.

BENTLEY, S. 2009. Sequencing the species pan-genome. *Nat Rev Micro,* 7**,** 258-259.

BERKMAN, P., SKARSHEWSKI, A., MANOLI, S., LORENC, M., STILLER, J., SMITS, L., LAI, K., CAMPBELL, E., KUBALÁKOVÁ, M., ŠIMKOVÁ, H., BATLEY, J., DOLEŽEL, J., HERNANDEZ, P. & EDWARDS, D. 2012a. Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theoretical and Applied Genetics,* 124**,** 423-432.

BERKMAN, P. J., MANOLI, S., MCKENZIE, M., KUBALÁKOVÁ, M., ŠIMKOVÁ, H., BATLEY, J., FLEURY, D., DOLEŽEL, J., EDWARDS, D., SKARSHEWSKI, A., LORENC, M. T., LAI, K., DURAN, C., LING, E. Y. S., STILLER, J., SMITS, L. &

IMELFORT, M. 2011a. Sequencing and assembly of low copy and genic regions of isolated Triticum aestivum chromosome arm 7DS. *Plant Biotechnology Journal,* 9**,** 768-775.

BERKMAN, P. J., SKARSHEWSKI, A., LORENC, M. T., LAI, K., DURAN, C., LING, E. Y., STILLER, J., SMITS, L., IMELFORT, M., MANOLI, S., MCKENZIE, M., KUBALAKOVA, M., SIMKOVA, H., BATLEY, J., FLEURY, D., DOLEZEL, J. & EDWARDS, D. 2011b. Sequencing and assembly of low copy and genic regions of isolated Triticum aestivum chromosome arm 7DS. *Plant Biotechnol J,* 9**,** 768-75.

BERKMAN, P. J., SKARSHEWSKI, A., MANOLI, S., LORENC, M. T., STILLER, J., LARS, SMITS, L., LAI, K., CAMPBELL, E., KUBALAKOVA, M., SIMKOVA, H., BATLEY, J., DOLEZEL, J., HERNANDEZ, P. & EDWARDS, D. 2012b. Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theoretical and Applied Genetics,* 124**,** 423-432.

BERKMAN, P. J., VISENDI, P., LEE, H. C., STILLER, J., MANOLI, S., LORENC, M. T., LAI, K., BATLEY, J., FLEURY, D., SIMKOVA, H., KUBALAKOVA, M., WEINING, S., DOLEZEL, J. & EDWARDS, D. 2013a. Dispersion and domestication shaped the genome of bread wheat. *Plant Biotechnol J,* 11**,** 564-71.

BERKMAN, P. J., VISENDI, P., LEE, H. C., STILLER, J., MANOLI, S., LORENC, M. T., LAI, K., BATLEY, J., FLEURY, D., SIMKOVA, H., KUBALAKOVA, M., WEINING, S., DOLEZEL, J. & EDWARDS, D. 2013b. Dispersion and domestication shaped the genome of bread wheat. *Plant Biotechnology Journal,* 11**,** 564-571.

BEUZEN, N. D., STEAR, M. J. & CHANG, K. C. 2000. Molecular markers and their use in animal breeding. *The Veterinary Journal,* 160**,** 42-52.

BIOINFOLOGICS. 2016. *The w2rap-contigger* [Online]. Available: http://bioinfologics.github.io/the-w2rap-contigger/ [Accessed 2016].

BOETZER, M., HENKEL, C. V., JANSEN, H. J., BUTLER, D. & PIROVANO, W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics,* 27.

BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-2120.

BOŽA, V., BREJOVÁ, B. & VINAŘ, T. 2014. GAML: Genome Assembly by Maximum Likelihood. *In:* BROWN, D. & MORGENSTERN, B. (eds.) *Algorithms in Bioinformatics: 14th International Workshop, WABI 2014, Wroclaw, Poland, September 8-10, 2014. Proceedings.* Berlin, Heidelberg: Springer Berlin Heidelberg.

BOŽA, V., BREJOVÁ, B. & VINAŘ, T. 2017. DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. *PLOS ONE,* 12**,** e0178751.

BRADNAM, K. R., FASS, J. N., ALEXANDROV, A., BARANAY, P., BECHNER, M., BIROL, I., BOISVERT, S., CHAPMAN, J. A., CHAPUIS, G., CHIKHI, R., CHITSAZ, H., CHOU, W.-C., CORBEIL, J., DEL FABBRO, C., DOCKING, T. R., DURBIN, R., EARL, D., EMRICH, S., FEDOTOV, P., FONSECA, N. A., GANAPATHY, G., GIBBS, R. A., GNERRE, S., GODZARIDIS, É., GOLDSTEIN, S., HAIMEL, M., HALL, G., HAUSSLER, D., HIATT, J. B., HO, I. Y., HOWARD, J., HUNT, M., JACKMAN, S. D., JAFFE, D. B., JARVIS, E. D., JIANG, H., KAZAKOV, S., KERSEY, P. J., KITZMAN, J. O., KNIGHT, J. R., KOREN, S., LAM, T.-W., LAVENIER, D., LAVIOLETTE, F., LI, Y., LI, Z., LIU, B., LIU, Y., LUO, R., MACCALLUM, I., MACMANES, M. D., MAILLET, N., MELNIKOV, S., NAQUIN, D., NING, Z., OTTO, T. D., PATEN, B., PAULO, O. S., PHILLIPPY, A. M., PINA-MARTINS, F., PLACE, M., PRZYBYLSKI, D., QIN, X., QU, C., RIBEIRO, F. J., RICHARDS, S., ROKHSAR, D. S., RUBY, J. G., SCALABRIN, S., SCHATZ, M. C., SCHWARTZ, D. C., SERGUSHICHEV, A., SHARPE, T., SHAW, T. I., SHENDURE, J., SHI, Y., SIMPSON, J. T., SONG, H., TSAREV, F., VEZZI, F., VICEDOMINI, R., VIEIRA, B. M., WANG, J., WORLEY, K. C., YIN, S., YIU, S.-M., YUAN, J., ZHANG, G., ZHANG, H., ZHOU, S. & KORF, I. F. 2013a. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience,* 2**,** 1-31.

BRADNAM, K. R., FASS, J. N. & KORF, I. F. 2013b. CEGMA gene predictions for Assemblathon 2 entries. *GigaScience Database.*

BRENCHLEY, R., SPANNAGL, M., PFEIFER, M., BARKER, G. L. A., D'AMORE, R., ALLEN, A. M., MCKENZIE, N., KRAMER, M., KERHORNOU, A., BOLSER, D., KAY, S., WAITE, D., TRICK, M., BANCROFT, I., GU, Y., HUO, N., LUO, M.-C., SEHGAL, S., GILL, B., KIANIAN, S., ANDERSON, O., KERSEY, P., DVORAK, J., MCCOMBIE, W. R., HALL, A., MAYER, K. F. X., EDWARDS, K. J., BEVAN, M. W. & HALL, N. 2012. Analysis of the breadwheat genome using whole-genome shotgun sequencing. *Nature,* 491**,** 705-710.

BRYAN, G. J., COLLINS, A. J., STEPHENSON, P., ORRY, A., SMITH, J. B. & GALE, M. D. 1997. Isolation and characterisation of microsatellites from hexaploid bread wheat. *Theoretical and Applied Genetics,* 94**,** 557-563.

BUCKLER, T. E. S. & HOLTSFORD, T. P. 1996. Zea ribosomal repeat evolution and substitution patterns. *Molecular Biology and Evolution,* 13**,** 623-632.

BUSH, S. J., CASTILLO-MORALES, A., TOVAR-CORONA, J. M., CHEN, L., KOVER, P. X. & URRUTIA, A. O. 2014. Presence–Absence Variation in A. thaliana Is Primarily Associated with Genomic Signatures Consistent with Relaxed Selective Constraints. *Molecular Biology and Evolution,* 31**,** 59-69.

CALDWELL, K. S., DVORAK, J., LAGUDAH, E. S., AKHUNOV, E., LUO, M. C. & WOLTERS, P. 2004. Sequence polymorphism in polyploid wheat and their d-genome diploid ancestor. *Genetics,* 167.

CAMACHO, C., COULOURIS, G., AVAGYAN, V., MA, N., PAPADOPOULOS, J. & BEALER, K. 2009. BLAST+: architecture and applications. *BMC Bioinformatics,* 10.

CAMPBELL, M. S., LAW, M., HOLT, C., STEIN, J. C., MOGHE, G. D., HUFNAGEL, D. E., LEI, J., ACHAWANANTAKUN, R., JIAO, D., LAWRENCE, C. J., WARE, D., SHIU, S. H., CHILDS, K. L., SUN, Y., JIANG, N. & YANDELL, M. 2014. MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol,* 164.

CANTAREL, B. L., KORF, I., ROBB, S. M., PARRA, G., ROSS, E., MOORE, B., HOLT, C., SANCHEZ ALVARADO, A. & YANDELL, M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res,* 18.

CAO, J., SCHNEEBERGER, K., OSSOWSKI, S., GUNTHER, T., BENDER, S., FITZ, J., KOENIG, D., LANZ, C., STEGLE, O., LIPPERT, C., WANG, X., OTT, F., MULLER, J., ALONSO-BLANCO, C., BORGWARDT, K., SCHMID, K. J. & WEIGEL, D. 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat Genet,* 43**,** 956-963.

CARTOLANO, M., HUETTEL, B., HARTWIG, B., REINHARDT, R. & SCHNEEBERGER, K. 2016. cDNA Library Enrichment of Full Length Transcripts for SMRT Long Read Sequencing. *PLOS ONE,* 11**,** e0157779.

CAVANAGH, C. R., CHAO, S., WANG, S., HUANG, B. E., STEPHEN, S., KIANI, S., FORREST, K., SAINTENAC, C., BROWN-GUEDIRA, G. L., AKHUNOVA, A., SEE, D., BAI, G., PUMPHREY, M., TOMAR, L., WONG, D., KONG, S., REYNOLDS, M., DA SILVA, M. L., BOCKELMAN, H., TALBERT, L., ANDERSON, J. A., DREISIGACKER, S., BAENZIGER, S., CARTER, A., KORZUN, V., MORRELL, P. L., DUBCOVSKY, J., MORELL, M. K., SORRELLS, M. E., HAYDEN, M. J. & AKHUNOV, E. 2013. Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars.

Proceedings of the National Academy of Sciences of the United States of America, 110**,** 8057-8062.

CAZAUX, B., SACOMOTO, G. & RIVALS, E. 2016. Superstring Graph: A New Approach for Genome Assembly. *In:* DONDI, R., FERTIN, G. & MAURI, G. (eds.) *Algorithmic Aspects in Information and Management: 11th International Conference, AAIM 2016, Bergamo, Italy, July 18-20, 2016, Proceedings.* Cham: Springer International Publishing.

CHANTRET, N., SALSE, J., SABOT, F., RAHMAN, S., BELLEC, A., LAUBIN, B., DUBOIS, I., DOSSAT, C., SOURDILLE, P., JOUDRIER, P., GAUTIER, M. F., CATTOLICO, L., BECKERT, M., AUBOURG, S., WEISSENBACH, J., CABOCHE, M., BERNARD, M., LEROY, P. & CHALHOUB, B. 2005. Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *Plant Cell,* 17**,** 1033-1045.

CHAO, S., ZHANG, W., AKHUNOV, E., SHERMAN, J., MA, Y., LUO, M.-C. & DUBCOVSKY, J. 2009. Analysis of gene-derived SNP marker polymorphism in US wheat (Triticum aestivum L.) cultivars. *Molecular Breeding,* 23**,** 23-33.

CHAPMAN, J. A., MASCHER, M., BULUÇ, A., BARRY, K., GEORGANAS, E., SESSION, A., STRNADOVA, V., JENKINS, J., SEHGAL, S., OLIKER, L., SCHMUTZ, J., YELICK, K. A., SCHOLZ, U., WAUGH, R., POLAND, J. A., MUEHLBAUER, G. J., STEIN, N. & ROKHSAR, D. S. 2015. A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biology,* 16**,** 1-17.

CHAUDHARI, N. M., GUPTA, V. K. & DUTTA, C. 2016. BPGA- an ultra-fast pan-genome analysis pipeline. 6**,** 24373.

CHEN, X., MIN, D., YASIR, T. A. & HU, Y.-G. 2012. Genetic Diversity, Population Structure and Linkage Disequilibrium in Elite Chinese Winter Wheat Investigated with SSR Markers. *PLOS ONE,* 7**,** e44510.

CHEN, Z. J. & NI, Z. 2006. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *Bioessays.,* 28.

CHOULET, F., WICKER, T., RUSTENHOLZ, C., PAUX, E., SALSE, J., LEROY, P., SCHLUB, S., LE PASLIER, M.-C., MAGDELENAT, G., GONTHIER, C., COULOUX, A., BUDAK, H., BREEN, J., PUMPHREY, M., LIU, S., KONG, X., JIA, J., GUT, M., BRUNEL, D., ANDERSON, J. A., GILL, B. S., APPELS, R., KELLER, B. & FEUILLET, C. 2010. Megabase Level Sequencing Reveals Contrasted

Organization and Evolution Patterns of the Wheat Gene and Transposable Element Spaces. *The Plant Cell,* 22**,** 1686-1701.

CINGOLANI, P., PLATTS, A., WANG, L. L., COON, M., TUNG, N., WANG, L., LAND, S. J., LU, X. & RUDEN, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. *Fly,* 6**,** 80-92.

CLARKSON, J. J., LIM, K. Y., KOVARIK, A., CHASE, M. W., KNAPP, S. & LEITCH, A. R. 2005. Long-term genome diploidization in allopolyploid Nicotiana section Repandae (Solanaceae). *New Phytologist,* 168**,** 241-252.

CLAVIJO, B. J., VENTURINI, L., SCHUDOMA, C., ACCINELLI, G. G., KAITHAKOTTIL, G., WRIGHT, J., BORRILL, P., KETTLEBOROUGH, G., HEAVENS, D., CHAPMAN, H., LIPSCOMBE, J., BARKER, T., LU, F.-H., MCKENZIE, N., RAATS, D., RAMIREZ-GONZALEZ, R. H., COINCE, A., PEEL, N., PERCIVAL-ALWYN, L., DUNCAN, O., TRÖSCH, J., YU, G., BOLSER, D. M., NAMAATI, G., KERHORNOU, A., SPANNAGL, M., GUNDLACH, H., HABERER, G., DAVEY, R. P., FOSKER, C., PALMA, F. D., PHILLIPS, A. L., MILLAR, A. H., KERSEY, P. J., UAUY, C., KRASILEVA, K. V., SWARBRECK, D., BEVAN, M. W. & CLARK, M. D. 2017. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Research,* 27**,** 885-896.

CLAVIJO, B. J., VENTURINI, L., SCHUDOMA, C., GARCIA ACCINELLI, G., KAITHAKOTTIL, G., WRIGHT, J., BORRILL, P., KETTLEBOROUGH, G., HEAVENS, D., CHAPMAN, H., LIPSCOMBE, J., BARKER, T., LU, F.-H., MCKENZIE, N., RAATS, D., RAMIREZ-GONZALEZ, R. H., COINCE, A., PEEL, N., PERCIVAL-ALWYN, L., DUNCAN, O., TRÖSCH, J., YU, G., BOLSER, D., NAAMATI, G., KERHORNOU, A., SPANNAGL, M., GUNDLACH, H., HABERER, G., DAVEY, R. P., FOSKER, C., DI PALMA, F., PHILLIPS, A., MILLAR, A. H., KERSEY, P. J., UAUY, C., KRASILEVA, K. V., SWARBRECK, D., BEVAN, M. W. & CLARK, M. D. 2016. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *bioRxiv*.

COGEPEDIA. 2017. *Sequenced Plant genomes* [Online]. online: onlie. Available: https://genomevolution.org/wiki/index.php/Sequenced_plant_genomes [Accessed 2017].

COKUS, S. J., FENG, S., ZHANG, X., CHEN, Z., MERRIMAN, B. & HAUDENSCHILD, C. D. 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature.,* 452.

COLLARD, B. C. Y. & MACKILL, D. J. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 363**,** 557-572.

COLLINS, R. E. & HIGGS, P. G. 2012. Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Mol Biol Evol,* 29**,** 3413-25.

CONANT, G. C., BIRCHLER, J. A. & PIRES, J. C. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr Opin Plant Biol.,* 19.

CONESA, A. & GÖTZ, S. 2008. Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. *International Journal of Plant Genomics,* 2008**,** 619832.

CONESA, A., GÖTZ, S., GARCÍA-GÓMEZ, J. M., TEROL, J., TALÓN, M. & ROBLES, M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics,* 21**,** 3674-3676.

CONSORTIUM, I. W. G. S. 2016a. IWGSC whole genome shotgun sequencing of Chinese Spring: Towards a Reference Sequence of Wheat. *Plan and Animal Genoem Conference XXIV.* San Diego.

CONSORTIUM, I. W. G. S. 2016b. Wheat Sequencing Consortium Releases Key Resource to the Scientific Community. wheatgenome.org.

CONSORTIUM, T. I. W. G. S. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science,* 345**,** 1251788.

COULONDRE, C., MILLER, J. H., FARABAUGH, P. J. & GILBERT, W. 1978. Molecular basis of base substitution hotspots in Escherichia coli. *Nature,* 274**,** 775-780.

CRONN, R., LISTON, A., PARKS, M., GERNANDT, D. S., SHEN, R. & MOCKLER, T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research,* 36**,** e122-e122.

CROS, D., DENIS, M., SANCHEZ, L., COCHARD, B., FLORI, A., DURAND-GASSELIN, T., NOUY, B., OMORE, A., POMIES, V., RIOU, V., SURYANA, E. & BOUVET, J.-M. 2015. Genomic selection prediction accuracy in a perennial crop: case study of oil palm (Elaeis guineensis Jacq.). *Theoretical and Applied Genetics,* 128**,** 397-410.

CROSSA, J., PEREZ, P., HICKEY, J., BURGUENO, J., ORNELLA, L., CERON-ROJAS, J., ZHANG, X., DREISIGACKER, S., BABU, R., LI, Y., BONNETT, D. &

MATHEWS, K. 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity,* 112**,** 48-60.

CUI, P., LIU, H., LIN, Q., DING, F., ZHUO, G., HU, S., LIU, D., YANG, W., ZHAN, K., ZHANG, A. & YU, J. 2009. A complete mitochondrial genome of wheat (Triticum aestivum cv. Chinese Yumai), and fast evolving mitochondrial genes in higher plants. *J Genet,* 88**,** 299-307.

DAMANIA, A. B. 1998. Diversity of major cultivated plants domesticated in the Near East.

DANECEK, P., AUTON, A., ABECASIS, G., ALBERS, C. A., BANKS, E., DEPRISTO, M. A., HANDSAKER, R. E., LUNTER, G., MARTH, G. T., SHERRY, S. T., MCVEAN, G., DURBIN, R. & GROUP, G. P. A. 2011. The variant call format and VCFtools. *Bioinformatics,* 27**,** 2156-2158.

DAVENPORT, C. F. & TÜMMLER, B. 2013. Advances in computational analysis of metagenome sequences. *Environmental Microbiology,* 15**,** 1-5.

DAVID, M., DURSI, L. J., YAO, D., BOUTROS, P. C. & SIMPSON, J. T. 2017. Nanocall: an open source basecaller for Oxford Nanopore sequencing data. *Bioinformatics,* 33**,** 49-55.

DE LA BASTIDE, M. & MCCOMBIE, W. R. 2007. Assembling genomic DNA sequences with PHRAP. *Curr Protoc Bioinformatics,* Chapter 11**,** Unit11.4.

DEL BLANCO, I. A., RAJARAM, S. & KRONSTAD, W. E. 2001. Agronomic Potential of Synthetic Hexaploid Wheat-Derived Populations I.A. del Blanco present address: Dep. of Plant Sciences, North Dakota State Univ., Fargo, ND 58105. Technical Paper no. 11547 of the Oregon State Univ. Agric. Expt. Stn. *Crop Science,* 41**,** 670-676.

DELCHER, A. L., PHILLIPPY, A., CARLTON, J. & SALZBERG, S. L. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res,* 30**,** 2478-83.

DÍAZ, A., ZIKHALI, M., TURNER, A. S., ISAAC, P. & LAURIE, D. A. 2012. Copy Number Variation Affecting the Photoperiod-B1 and Vernalization-A1 Genes Is Associated with Altered Flowering Time in Wheat (Triticum aestivum). *PLOS ONE,* 7**,** e33234.

DING, J., ARAKI, H., WANG, Q., ZHANG, P., YANG, S., CHEN, J.-Q. & TIAN, D. 2007. Highly asymmetric rice genomes. *BMC Genomics,* 8**,** 154.

DUBCOVSKY, J. & DVORAK, J. 2007b. Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* 316**,** 1862-1866.

DUITAMA, J., SILVA, A., SANABRIA, Y., CRUZ, D. F., QUINTERO, C., BALLEN, C., LORIEUX, M., SCHEFFLER, B., FARMER, A., TORRES, E., OARD, J. & TOHME,

J. 2015. Whole Genome Sequencing of Elite Rice Cultivars as a Comprehensive Information Resource for Marker Assisted Selection. *PLOS ONE,* 10**,** e0124617.

DVORAK, J., AKHUNOV, E. D., AKHUNOV, A. R., DEAL, K. R. & LUO, M.-C. 2006. Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat. *Mol Biol Evol,* 23.

DVORAK, J., LUO, M. C., YANG, Z. L. & ZHANG, H. B. 1998. The structure of the Aegilops tauschii genepool and the evolution of hexaploid wheat. *TAG Theor Appl Genet,* 97.

DVORAK, J., MCGUIRE, P. & CASSIDY, B. 1988. Apparent sources of the A genomes of wheats inferred from the polymorphism in abundance and restriction fragment length of repeated nucleotide sequences. *Genome,* 30.

DVOŘÁK, J., TERLIZZI, P. D., ZHANG, H.-B. & RESTA, P. 1993. The evolution of polyploid wheats: identification of the A genome donor species. *Genome,* 36**,** 21-31.

EDAE, E. A., BOWDEN, R. L. & POLAND, J. 2015. Application of Population Sequencing (POPSEQ) for Ordering and Imputing Genotyping-by-Sequencing Markers in Hexaploid Wheat. *G3: Genes|Genomes|Genetics*.

EDWARDS, D. & BATLEY, J. 2004. Plant bioinformatics: from genome to phenome. *Trends Biotechnol,* 22**,** 232-7.

EDWARDS, D., BATLEY, J. & SNOWDON, R. 2013. Accessing complex crop genomes with next-generation sequencing. *Theoretical and Applied Genetics,* 126**,** 1-11.

EDWARDS, D., WILCOX, S., BARRERO, R. A., FLEURY, D., CAVANAGH, C. R., FORREST, K. L., HAYDEN, M. J., MOOLHUIJZEN, P., KEEBLE-GAGNÈRE, G., BELLGARD, M. I., LORENC, M. T., SHANG, C. A., BAUMANN, U., TAYLOR, J. M., MORELL, M. K., LANGRIDGE, P., APPELS, R. & FITZGERALD, A. 2012. Bread matters: a national initiative to profile the genetic diversity of Australian wheat. *Plant Biotechnology Journal,* 10**,** 703-708.

EICHTEN, S. R., FOERSTER, J. M., DE LEON, N., KAI, Y., YEH, C.-T., LIU, S., JEDDELOH, J. A., SCHNABLE, P. S., KAEPPLER, S. M. & SPRINGER, N. M. 2011. B73-Mo17 Near-Isogenic Lines Demonstrate Dispersed Structural Variation in Maize. *Plant Physiology,* 156**,** 1679-1690.

EID, J., FEHR, A., GRAY, J., LUONG, K., LYLE, J., OTTO, G., PELUSO, P., RANK, D., BAYBAYAN, P., BETTMAN, B., BIBILLO, A., BJORNSON, K., CHAUDHURI, B., CHRISTIANS, F., CICERO, R., CLARK, S., DALAL, R., DEWINTER, A., DIXON, J., FOQUET, M., GAERTNER, A., HARDENBOL, P., HEINER, C., HESTER, K.,

HOLDEN, D., KEARNS, G., KONG, X., KUSE, R., LACROIX, Y., LIN, S., LUNDQUIST, P., MA, C., MARKS, P., MAXHAM, M., MURPHY, D., PARK, I., PHAM, T., PHILLIPS, M., ROY, J., SEBRA, R., SHEN, G., SORENSON, J., TOMANEY, A., TRAVERS, K., TRULSON, M., VIECELI, J., WEGENER, J., WU, D., YANG, A., ZACCARIN, D., ZHAO, P., ZHONG, F., KORLACH, J. & TURNER, S. 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science,* 323**,** 133-138.

EISENSTEIN, M. 2012. Oxford Nanopore announcement sets sequencing sector abuzz. *Nat Biotech,* 30**,** 295-296.

ELSIK, C. G., WORLEY, K. C., ZHANG, L., MILSHINA, N. V., JIANG, H., REESE, J. T., CHILDS, K. L., VENKATRAMAN, A., DICKENS, C. M. & WEINSTOCK, G. M. 2006. Community annotation: procedures, protocols and supporting tools. *Genome Res,* 16.

ENSEMBL 2017. Plant genomes.

EVENSON, R. E. & GOLLIN, D. 2003. Assessing the Impact of the Green Revolution, 1960 to 2000. *Science,* 300**,** 758-762.

FAN, L., ZHANG, M.-Y., LIU, Q.-Z., LI, L.-T., SONG, Y., WANG, L.-F., ZHANG, S.-L. & WU, J. 2013. Transferability of Newly Developed Pear SSR Markers to Other Rosaceae Species. *Plant Molecular Biology Reporter,* 31**,** 1271-1282.

FAO 2016. *Save and grow in practice. A guide to sustainable cereal production,* Rome, FAO.

FELDMAN, M. & KISLEV, M. E. 2007. Domestication of emmer wheat and evolution of free-threshing tetraploid wheat. *Israel Journal of Plant Sciences,* 55**,** 207-221.

FERREIRA, A., SILVA, M. F. D., SILVA, L. D. C. E. & CRUZ, C. D. 2006. Estimating the effects of population size and type on the accuracy of genetic maps. *Genetics and Molecular Biology,* 29**,** 187-192.

FEUILLET, C., LANGRIDGE, P. & WAUGH, R. 2008. Cereal breeding takes a walk on the wild side. *Trends Genet,* 24.

FEUK, L., CARSON, A. R. & SCHERER, S. W. 2006. Structural variation in the human genome. *Nat Rev Genet,* 7**,** 85-97.

FLAVELL, R., BENNETT, M., SMITH, J. & SMITH, D. 1974. Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochem Genet,* 12.

FORREST, K. L., PUJOL, V., BULLI, P., PUMPHREY, M., WELLINGS, C. & HERRERA-FOESSEL, S. 2014. Development of a SNP marker assay for the Lr67 gene of wheat using a genotyping by sequencing approach. *Mol Breed,* 34.

FREELING, M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu Rev Plant Biol,* 60**,** 433-53.

FRIEBE, B. & GILL, B. S. 1994. C-band polymorphism and structural rearrangements detected in common wheat (Triticum aestivum). *Euphytica,* 78**,** 1-5.

FROMMER, M., MCDONALD, L. E., MILLAR, D. S., COLLIS, C. M., WATT, F., GRIGG, G. W., MOLLOY, P. L. & PAUL, C. L. 1992. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences,* 89**,** 1827-1831.

FU, Y., LUO, G. Z., CHEN, K., DENG, X., YU, M. & HAN, D. 2015. N6-methyldeoxyadenosine marks active transcription start sites in chlamydomonas. *Cell,* 161.

GABALDON, T. & KOONIN, E. V. 2013. Functional and evolutionary implications of gene orthology. *Nat Rev Genet,* 14**,** 360-366.

GAN, X., STEGLE, O., BEHR, J., STEFFEN, J. G., DREWE, P., HILDEBRAND, K. L., LYNGSOE, R., SCHULTHEISS, S. J., OSBORNE, E. J., SREEDHARAN, V. T., KAHLES, A., BOHNERT, R., JEAN, G., DERWENT, P., KERSEY, P., BELFIELD, E. J., HARBERD, N. P., KEMEN, E., TOOMAJIAN, C., KOVER, P. X., CLARK, R. M., RATSCH, G. & MOTT, R. 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature,* 477**,** 419-423.

GARDINER, L.-J., QUINTON-TULLOCH, M., OLOHAN, L., PRICE, J., HALL, N. & HALL, A. 2015. A genome-wide survey of DNA methylation in hexaploid wheat. *Genome Biology,* 16**,** 273.

GILES, R. J. & BROWN, T. A. 2006. GluDy allele variations in Aegilops tauschii and Triticum aestivum: implications for the origins of hexaploid wheats. *Theor Appl Genet,* 112**,** 1563-72.

GILL, B. S., APPELS, R., BOTHA-OBERHOLSTER, A.-M., BUELL, C. R., BENNETZEN, J. L., CHALHOUB, B., CHUMLEY, F., DVOŘÁK, J., IWANAGA, M., KELLER, B., LI, W., MCCOMBIE, W. R., OGIHARA, Y., QUETIER, F. & SASAKI, T. 2004. A Workshop Report on Wheat Genome Sequencing: International Genome Research on Wheat Consortium. *Genetics,* 168**,** 1087-1096.

GILL, K. S., GILL, B. S., ENDO, T. R. & TAYLOR, T. 1996. Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. *Genetics,* 144.

GIOVANNI PARMIGIANI, E. S. G., RAFAEL A. IRIZARRY, SCOTT L. ZEGER 2003. *The Analysis of Gene Expression Data*, Springer New York.

GOFF, S. A., RICKE, D., LAN, T.-H., PRESTING, G., WANG, R., DUNN, M., GLAZEBROOK, J., SESSIONS, A., OELLER, P., VARMA, H., HADLEY, D., HUTCHISON, D., MARTIN, C., KATAGIRI, F., LANGE, B. M., MOUGHAMER, T., XIA, Y., BUDWORTH, P., ZHONG, J., MIGUEL, T., PASZKOWSKI, U., ZHANG, S., COLBERT, M., SUN, W.-L., CHEN, L., COOPER, B., PARK, S., WOOD, T. C., MAO, L., QUAIL, P., WING, R., DEAN, R., YU, Y., ZHARKIKH, A., SHEN, R., SAHASRABUDHE, S., THOMAS, A., CANNINGS, R., GUTIN, A., PRUSS, D., REID, J., TAVTIGIAN, S., MITCHELL, J., ELDREDGE, G., SCHOLL, T., MILLER, R. M., BHATNAGAR, S., ADEY, N., RUBANO, T., TUSNEEM, N., ROBINSON, R., FELDHAUS, J., MACALMA, T., OLIPHANT, A. & BRIGGS, S. 2002. A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *japonica*). *Science,* 296**,** 92-100.

GOLICZ, A. A. 2016. *Construction and analysis of the Brassica oleracea pangenome.* Phylosophical doctor PhD, The University of Queensland.

GOLICZ, A. A., BATLEY, J. & EDWARDS, D. 2016a. Towards plant pangenomics. *Plant Biotechnol J,* 14**,** 1099-105.

GOLICZ, A. A., BAYER, P. E., BARKER, G. C., EDGER, P. P., KIM, H., MARTINEZ, P. A., CHAN, C. K. K., SEVERN-ELLIS, A., MCCOMBIE, W. R., PARKIN, I. A. P., PATERSON, A. H., PIRES, J. C., SHARPE, A. G., TANG, H., TEAKLE, G. R., TOWN, C. D., BATLEY, J. & EDWARDS, D. 2016b. The pangenome of an agronomically important crop plant Brassica oleracea. *Nature Communications,* 7**,** 13390.

GOLICZ, A. A., MARTINEZ, P. A., ZANDER, M., PATEL, D. A., VAN DE WOUW, A. P., VISENDI, P., FITZGERALD, T. L., EDWARDS, D. & BATLEY, J. 2015a. Gene loss in the fungal canola pathogen Leptosphaeria maculans. *Functional & integrative genomics,* 15**,** 189-196.

GOLICZ, A. A., SCHLIEP, M., LEE, H. T., LARKUM, A. W. D., DOLFERUS, R., BATLEY, J., CHAN, C.-K. K., SABLOK, G., RALPH, P. J. & EDWARDS, D. 2015b. Genome-wide survey of the seagrass Zostera muelleri suggests modification of the ethylene signalling network. *Journal of Experimental Botany*.

GONZALEZ-GARAY, M. L. 2016. Introduction to Isoform Sequencing Using Pacific Biosciences Technology (Iso-Seq). *In:* WU, J. (ed.) *Transcriptomics and Gene Regulation.* Dordrecht: Springer Netherlands.

GOODWIN, S., GURTOWSKI, J., ETHE-SAYERS, S., DESHPANDE, P., SCHATZ, M. C. & MCCOMBIE, W. R. 2015. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research,* 25**,** 1750-1756.

GOODWIN, S., MCPHERSON, J. D. & MCCOMBIE, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet,* 17**,** 333-351.

GORDON, S. P., PRIEST, H., DES MARAIS, D. L., SCHACKWITZ, W., FIGUEROA, M., MARTIN, J., BRAGG, J. N., TYLER, L., LEE, C.-R., BRYANT, D., WANG, W., MESSING, J., MANZANEDA, A. J., BARRY, K., GARVIN, D. F., BUDAK, H., TUNA, M., MITCHELL-OLDS, T., PFENDER, W. F., JUENGER, T. E., MOCKLER, T. C. & VOGEL, J. P. 2014. Genome diversity in Brachypodium distachyon: deep sequencing of highly diverse inbred lines. *The Plant Journal,* 79**,** 361-374.

GORE, M. A., CHIA, J.-M., ELSHIRE, R. J., SUN, Q., ERSOZ, E. S. & HURWITZ, B. L. 2009. A first-generation haplotype map of maize. *Science,* 326.

GREER, E. L., BLANCO, M. A., GU, L., SENDINC, E., LIU, J. & ARISTIZÁBAL-CORRALES, D. 2015. DNA methylation on N6-adenine in C. elegans. *Cell,* 161.

GREGORY R. WARNES, B. B., LODEWIJK BONEBAKKER, ROBERT GENTLEMAN, WOLFGANG HUBER ANDY LIAW, THOMAS LUMLEY, MARTIN MAECHLER, ARNI MAGNUSSON, STEFFEN MOELLER, MARC SCHWARTZ AND BILL VENABLES 2015. gplots: Various R Programming Tools for Plotting Data. R package version 2.17.0 ed.

GRIFFITHS, S., SHARP, R., FOOTE, T. N., BERTIN, I., WANOUS, M., READER, S., COLAS, I. & MOORE, G. 2006. Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. *Nature,* 439**,** 749-752.

GROVER, C., GALLAGHER, J., SZADKOWSKI, E., YOO, M., FLAGEL, L. & WENDEL, J. 2012. Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytol.,* 196.

GUO, Y., CHEN, S., LI, Z. & COWLING, W. A. 2014. Center of Origin and Centers of Diversity in an Ancient Crop, Brassica rapa (Turnip Rape). *Journal of Heredity,* 105**,** 555-565.

GUPTA, P. K., KULWAL, P. L. & RUSTGI, S. 2005. Wheat cytogenetics in the genomics era and its relevance to breeding. *Cytogenetic and Genome Research,* 109**,** 315-327.

GUPTA, P. K., MIR, R. R., MOHAN, A. & KUMAR, J. 2008. Wheat Genomics: Present Status and Future Prospects. *International Journal of Plant Genomics,* 2008**,** 896451.

HARCOURT, R. L. & GALE, M. D. 1991. A chromosome-specific DNA sequence which reveals a high level of RFLP in wheat. *Theoretical and Applied Genetics,* 81**,** 397-400.

HARDIGAN, M. A., CRISOVAN, E., HAMILTON, J. P., KIM, J., LAIMBEER, P., LEISNER, C. P., MANRIQUE-CARPINTERO, N. C., NEWTON, L., PHAM, G. M., VAILLANCOURT, B., YANG, X., ZENG, Z., DOUCHES, D. S., JIANG, J., VEILLEUX, R. E. & BUELL, C. R. 2016. Genome Reduction Uncovers a Large Dispensable Genome and Adaptive Role for Copy Number Variation in Asexually Propagated Solanum tuberosum. *The Plant Cell,* 28**,** 388-405.

HAUDRY, A., CENCI, A., RAVEL, C., BATAILLON, T., BRUNEL, D., PONCET, C., HOCHU, I., POIRIER, S., SANTONI, S., GLÉMIN, S. & DAVID, J. 2007. Grinding up Wheat: A Massive Loss of Nucleotide Diversity Since Domestication. *Molecular Biology and Evolution,* 24**,** 1506-1517.

HAYATSU, H., WATAYA, Y. & KAI, K. 1970. Addition of sodium bisulfite to uracil and to cytosine. *Journal of the American Chemical Society,* 92**,** 724-726.

HAYWARD, A. M., ANNALIESE S.; DALTON-MORGAN, JESSICA; ZANDER, MANUEL; EDWARDS, DAVID; BATLEY, JACQUELINE 2012. SNP discovery and applications in Brassica napus. *Journal of Plant Biotechnology,* 39**,** 49-61.

HEDDEN, P. 2003. The genes of the Green Revolution. *Trends in Genetics,* 19**,** 5-9.

HEFFNER, E. L., LORENZ, A. J., JANNINK, J.-L. & SORRELLS, M. E. 2010. Plant Breeding with Genomic Selection: Gain per Unit Time and Cost All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher. *Crop Science,* 50**,** 1681-1690.

HESLOT, N., YANG, H.-P., SORRELLS, M. E. & JANNINK, J.-L. 2012. Genomic Selection in Plant Breeding: A Comparison of Models. *Crop Science,* 52**,** 146-160.

HEUN, M., SCHÄFER-PREGL, R., KLAWAN, D., CASTAGNA, R., ACCERBI, M., BORGHI, B. & SALAMINI, F. 1997. Site of Einkorn Wheat Domestication Identified by DNA Fingerprinting. *Science,* 278**,** 1312-1314.

HIRSCH, C. N., FOERSTER, J. M., JOHNSON, J. M., SEKHON, R. S., MUTTONI, G. & VAILLANCOURT, B. 2014. Insights into the maize pan-genome and pan-transcriptome. *Plant Cell.,* 26.

HOLT, C. & YANDELL, M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics,* 12**,** 491.

HUANG, S., SIRIKHACHORNKIT, A., SU, X., FARIS, J., GILL, B., HASELKORN, R. & GORNICKI, P. 2002a. Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 8133-8138.

HUANG, X., BÖRNER, A., RÖDER, M. & GANAL, M. 2002b. Assessing genetic diversity of wheat (Triticum aestivum L.) germplasm using microsatellite markers. *Theoretical and Applied Genetics,* 105**,** 699-707.

HUANG, X., WANG, J., ALURU, S., YANG, S.-P. & HILLIER, L. 2003. PCAP: A Whole-Genome Assembly Program. *Genome Research,* 13**,** 2164-2170.

HUNT, M., KIKUCHI, T., SANDERS, M., NEWBOLD, C., BERRIMAN, M. & OTTO, T. D. 2013. REAPR: a universal tool for genome assembly evaluation. *Genome Biology,* 14**,** R47.

HUO, N., VOGEL, J., LAZO, G., YOU, F., MA, Y., MCMAHON, S., DVORAK, J., ANDERSON, O., LUO, M.-C. & GU, Y. 2009. Structural characterization of Brachypodium genome and its syntenic relationship with rice and wheat. *Plant Molecular Biology,* 70**,** 47-61.

HUSE, S. M., HUBER, J. A., MORRISON, H. G., SOGIN, M. L. & WELCH, D. M. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol,* 8**,** R143.

IAFRATE, A. J., FEUK, L., RIVERA, M. N., LISTEWNIK, M. L., DONAHOE, P. K., QI, Y., SCHERER, S. W. & LEE, C. 2004. Detection of large-scale variation in the human genome. *Nat Genet,* 36**,** 949-951.

IAKOUBOV, L., MOSSAKOWSKA, M., SZWED, M., DUAN, Z., SESTI, F. & PUZIANOWSKA-KUZNICKA, M. 2013. A Common Copy Number Variation (CNV) Polymorphism in the CNTNAP4 Gene: Association with Aging in Females. *PLOS ONE,* 8**,** e79790.

IEHISA, J. C. M., OHNO, R., KIMURA, T., ENOKI, H., NISHIMURA, S., OKAMOTO, Y., NASUDA, S. & TAKUMI, S. 2014. A High-Density Genetic Map with Array-Based Markers Facilitates Structural and Quantitative Trait Locus Analyses of the Common Wheat Genome. *DNA Research,* 21**,** 555-567.

INTERNATIONAL BRACHYPODIUM, I. 2010. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature,* 463**,** 763-768.

IOVENE, M., ZHANG, T., LOU, Q., BUELL, C. R. & JIANG, J. 2013. Copy number variation in potato – an asexually propagated autotetraploid species. *The Plant Journal,* 75**,** 80-89.

IQBAL, Z., CACCAMO, M., TURNER, I., FLICEK, P. & MCVEAN, G. 2012. De novo assembly and genotyping of variants using colored de bruijn graphs. *Nat Genet,* 44.

IRISH, V. F. & LITT, A. 2005. Flower development and evolution: gene duplication, diversification and redeployment. *Current Opinion in Genetics & Development,* 15**,** 454-460.

ISHII, T., MORI, N. & OGIHARA, Y. 2001. Evaluation of allelic diversity at chloroplast microsatellite loci among common wheat and its ancestral species. *Theoretical and Applied Genetics,* 103**,** 896-904.

IWAKI, K., HARUNA, S., NIWA, T. & KATO, K. 2001. Adaptation and ecological differentiation in wheat with special reference to geographical variation of growth habit and Vrn genotype. *Plant Breeding,* 120**,** 107-114.

IWGSC 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science,* 345.

JACOBSEN, A., HENDRIKSEN, R. S., AARESTURP, F. M., USSERY, D. W. & FRIIS, C. 2011. The Salmonella enterica pan-genome. *Microb Ecol,* 62**,** 487-504.

JANNINK, J.-L., LORENZ, A. J. & IWATA, H. 2010. Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics,* 9**,** 166-177.

JIA, J., ZHAO, S., KONG, X., LI, Y., ZHAO, G., HE, W., APPELS, R., PFEIFER, M., TAO, Y., ZHANG, X., JING, R., ZHANG, C., MA, Y., GAO, L., GAO, C., SPANNAGL, M., MAYER, K. F. X., LI, D., PAN, S., ZHENG, F., HU, Q., XIA, X., LI, J., LIANG, Q., CHEN, J., WICKER, T., GOU, C., KUANG, H., HE, G., LUO, Y., KELLER, B., XIA, Q., LU, P., WANG, J., ZOU, H., ZHANG, R., XU, J., GAO, J., MIDDLETON, C., QUAN, Z., LIU, G., WANG, J., YANG, H., LIU, X., HE, Z., MAO, L. & WANG, J. 2013. Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature,* 496**,** 91-95.

JIN, M., LIU, H., HE, C., FU, J., XIAO, Y., WANG, Y., XIE, W., WANG, G. & YAN, J. 2016. Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. *Scientific Reports,* 6**,** 18936.

JORDAN, K. W., WANG, S., LUN, Y., GARDINER, L.-J., MACLACHLAN, R., HUCL, P., WIEBE, K., WONG, D., FORREST, K. L., SHARPE, A. G., SIDEBOTTOM, C. H.,

HALL, N., TOOMAJIAN, C., CLOSE, T., DUBCOVSKY, J., AKHUNOVA, A., TALBERT, L., BANSAL, U. K., BARIANA, H. S., HAYDEN, M. J., POZNIAK, C., JEDDELOH, J. A., HALL, A. & AKHUNOV, E. 2015. A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biology,* 16**,** 48.

JURKA, J., KAPITONOV, V. V., PAVLICEK, A., KLONOWSKI, P., KOHANY, O. & WALICHIEWICZ, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res,* 110**,** 462-7.

KAEPPLER, S. 2012. Heterosis: Many Genes, Many Mechanisms—End the Search for an Undiscovered Unifying Theory. *ISRN Botany,* 2012**,** 12.

KAM-MORGAN, L. N. W. 1988. DNA restriction fragment length polymorphisms as genetic markers in mapping the wheat genome. *Dissertation Abstracts International, B (Sciences and Engineering),* 48**,** 2201B-2202B.

KASHKUSH, K., FELDMAN, M. & LEVY, A. A. 2002. Gene Loss, Silencing and Activation in a Newly Synthesized Wheat Allotetraploid. *Genetics,* 160**,** 1651-1659.

KEANE, T. M., GOODSTADT, L., DANECEK, P., WHITE, M. A., WONG, K., YALCIN, B., HEGER, A., AGAM, A., SLATER, G., GOODSON, M., FURLOTTE, N. A., ESKIN, E., NELLAKER, C., WHITLEY, H., CLEAK, J., JANOWITZ, D., HERNANDEZ-PLIEGO, P., EDWARDS, A., BELGARD, T. G., OLIVER, P. L., MCINTYRE, R. E., BHOMRA, A., NICOD, J., GAN, X., YUAN, W., VAN DER WEYDEN, L., STEWARD, C. A., BALA, S., STALKER, J., MOTT, R., DURBIN, R., JACKSON, I. J., CZECHANSKI, A., GUERRA-ASSUNCAO, J. A., DONAHUE, L. R., REINHOLDT, L. G., PAYSEUR, B. A., PONTING, C. P., BIRNEY, E., FLINT, J. & ADAMS, D. J. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature,* 477**,** 289-294.

KELLER, O., KOLLMAR, M., STANKE, M. & WAACK, S. 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics,* 27**,** 757-763.

KENAN-EICHLER, M., LESHKOWITZ, D., TAL, L., NOOR, E., MELAMED-BESSUDO, C. & FELDMAN, M. 2011. Wheat hybridization and polyploidization results in deregulation of small RNAs. *Genetics.,* 188.

KERSEY, P. J., ALLEN, J. E., ARMEAN, I., BODDU, S., BOLT, B. J., CARVALHO-SILVA, D., CHRISTENSEN, M., DAVIS, P., FALIN, L. J., GRABMUELLER, C., HUMPHREY, J., KERHORNOU, A., KHOBOVA, J., ARANGANATHAN, N. K., LANGRIDGE, N., LOWY, E., MCDOWALL, M. D., MAHESWARI, U., NUHN, M.,

ONG, C. K., OVERDUIN, B., PAULINI, M., PEDRO, H., PERRY, E., SPUDICH, G., TAPANARI, E., WALTS, B., WILLIAMS, G., TELLO–RUIZ, M., STEIN, J., WEI, S., WARE, D., BOLSER, D. M., HOWE, K. L., KULESHA, E., LAWSON, D., MASLEN, G. & STAINES, D. M. 2016. Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Research,* 44**,** D574-D580.

KIHARA, H. 1919. Über cytologische Studien nei einige Getreidearten. I. Species-Bastarde des Weizens und Weizenroggen-Bastarde. *Botanical Magazine,* 32**,** 17-38.

KIHARA, H. 1966. Factors affecting the evolution of common wheat. *Indian J. Genet.,* 26A**,** 14-28.

KILIAN, B., OZKAN, H., WALTHER, A., KOHL, J., DAGAN, T., SALAMINI, F. & MARTIN, W. 2007. Molecular diversity at 18 loci in 321 wild and 92 domesticate lines reveal no reduction of nucleotide diversity during Triticum monococcum (Einkorn) domestication: implications for the origin of agriculture. *Mol Biol Evol,* 24**,** 2657-68.

KIM, D., LANGMEAD, B. & SALZBERG, S. L. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Meth,* 12**,** 357-360.

KIM, D., PERTEA, G., TRAPNELL, C., PIMENTEL, H., KELLEY, R. & SALZBERG, S. L. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology,* 14**,** 1-13.

KOREN, S. & PHILLIPPY, A. M. 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology,* 23**,** 110-120.

KORLACH, J., BJORNSON, K. P., CHAUDHURI, B. P., CICERO, R. L., FLUSBERG, B. A., GRAY, J. J., HOLDEN, D., SAXENA, R., WEGENER, J. & TURNER, S. W. 2010. Real-Time DNA Sequencing from Single Polymerase Molecules. *Methods in Enzymology,* 472**,** 431-455.

KOZIOL, M. J., BRADSHAW, C. R., ALLEN, G. E., COSTA, A. S., FREZZA, C. & GURDON, J. B. 2015. Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. *Nat Struct Mol Biol*.

LAI, J., LI, R., XU, X., JIN, W., XU, M., ZHAO, H., XIANG, Z., SONG, W., YING, K., ZHANG, M., JIAO, Y., NI, P., ZHANG, J., LI, D., GUO, X., YE, K., JIAN, M., WANG, B., ZHENG, H., LIANG, H., ZHANG, X., WANG, S., CHEN, S., LI, J., FU, Y., SPRINGER, N. M., YANG, H., WANG, J., DAI, J., SCHNABLE, P. S. & WANG, J. 2010. Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet,* 42**,** 1027-1030.

LAI, K. 2015. *Genome Diversity in Triticum aestivum.* Philosophical doctor Research, University of Queensland.

LAI, K., BERKMAN, P. J., LORENC, M. T., DURAN, C., SMITS, L., MANOLI, S., STILLER, J. & EDWARDS, D. 2012a. WheatGenome.info: An Integrated Database and Portal for Wheat Genome Information. *Plant and Cell Physiology,* 53**,** e2.

LAI, K., DURAN, C., BERKMAN, P. J., LORENC, M. T., STILLER, J., MANOLI, S., HAYDEN, M. J., FORREST, K. L., FLEURY, D., BAUMANN, U., ZANDER, M., MASON, A. S., BATLEY, J. & EDWARDS, D. 2012b. Single nucleotide polymorphism discovery from wheat next-generation sequence data. *Plant Biotechnology Journal,* 10**,** 743-749.

LAI, K., LORENC, M. T., LEE, H., BERKMAN, P. J., BAYER, P. E., VISENDI, P., RUPERAO, P., FITZGERALD, T. L., ZANDER, M., CHAN, C. K., MANOLI, S., STILLER, J., BATLEY, J. & EDWARDS, D. 2015a. Identification and characterisation of more than 4 million inter-varietal SNPs across the group 7 chromosomes of bread wheat. *Plant Biotechnology Journal* 13**,** 97-104.

LAING, C., BUCHANAN, C., TABOADA, E. N., ZHANG, Y., KROPINSKI, A., VILLEGAS, A., THOMAS, J. E. & GANNON, V. P. 2010. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics,* 11**,** 461.

LAM, H.-M., XU, X., LIU, X., CHEN, W., YANG, G., WONG, F.-L., LI, M.-W., HE, W., QIN, N., WANG, B., LI, J., JIAN, M., WANG, J., SHAO, G., WANG, J., SUN, S. S.-M. & ZHANG, G. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet,* 42**,** 1053-1059.

LAMOUREUX, D., PETERSON, D. G., LI, W., FELLERS, J. P. & GILL, B. S. 2005. The efficacy of Cot-based gene enrichment in wheat (Triticum aestivum L.). *Genome,* 48**,** 1120-1126.

LANDER, E. S. & WATERMAN, M. S. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics,* 2**,** 231-9.

LANGMEAD, B. & SALZBERG, S. L. 2012a. Fast gapped-read alignment with Bowtie 2. *Nat Meth,* 9**,** 357-359.

LAPIERRE, P. & GOGARTEN, J. P. 2009. Estimating the size of the bacterial pan-genome. *Trends in Genetics,* 25**,** 107-110.

LAZO, G. R., CHAO, S., HUMMEL, D. D., EDWARDS, H., CROSSMAN, C. C., LUI, N., MATTHEWS, D. E., CAROLLO, V. L., HANE, D. L., YOU, F. M., BUTLER, G. E., MILLER, R. E., CLOSE, T. J., PENG, J. H., LAPITAN, N. L. V., GUSTAFSON, J. P.,

QI, L. L., ECHALIER, B., GILL, B. S., DILBIRLIGI, M., RANDHAWA, H. S., GILL, K. S., GREENE, R. A., SORRELLS, M. E., AKHUNOV, E. D., DVOŘÁK, J., LINKIEWICZ, A. M., DUBCOVSKY, J., HOSSAIN, K. G., KALAVACHARLA, V., KIANIAN, S. F., MAHMOUD, A. A., MIFTAHUDIN, MA, X.-F., CONLEY, E. J., ANDERSON, J. A., PATHAN, M. S., NGUYEN, H. T., MCGUIRE, P. E., QUALSET, C. O. & ANDERSON, O. D. 2004. Development of an Expressed Sequence Tag (EST) Resource for Wheat (Triticum aestivum L.). *EST Generation, Unigene Analysis, Probe Selection and Bioinformatics for a 16,000-Locus Bin-Delineated Map,* 168**,** 585-593.

LEE, C., IAFRATE, A. J. & BROTHMAN, A. R. 2007. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat Genet*.

LEGGETT, R. M., CLAVIJO, B. J., CLISSOLD, L., CLARK, M. D. & CACCAMO, M. 2014. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics,* 30**,** 566-568.

LELLEY, T., STACHEL, M., GRAUSGRUBER, H. & VOLLMANN, J. 2000. Analysis of relationships between Aegilops tauschii and the D genome of wheat utilizing microsatellites. *Genome,* 43**,** 661-668.

LEV-YADUN, S., GOPHER, A. & ABBO, S. 2000. The Cradle of Agriculture. *Science,* 288**,** 1602-1603.

LEVENE, M. J., KORLACH, J., TURNER, S. W., FOQUET, M., CRAIGHEAD, H. G. & WEBB, W. W. 2003. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. *Science,* 299**,** 682-686.

LEVY, S., SUTTON, G., NG, P. C., FEUK, L., HALPERN, A. L., WALENZ, B. P., AXELROD, N., HUANG, J., KIRKNESS, E. F., DENISOV, G., LIN, Y., MACDONALD, J. R., PANG, A. W., SHAGO, M., STOCKWELL, T. B., TSIAMOURI, A., BAFNA, V., BANSAL, V., KRAVITZ, S. A., BUSAM, D. A., BEESON, K. Y., MCINTOSH, T. C., REMINGTON, K. A., ABRIL, J. F., GILL, J., BORMAN, J., ROGERS, Y. H., FRAZIER, M. E., SCHERER, S. W. & STRAUSBERG, R. L. 2007. The diploid genome sequence of an individual human. *PLoS Biol,* 5.

LI, A.-L., GENG, S.-F., ZHANG, L.-Q., LIU, D.-C. & MAO, L. 2015a. Making the Bread: Insights from Newly Synthesized Allohexaploid Wheat. *Molecular Plant,* 8**,** 847-859.

LI, C., BAI, G., CHAO, S. & WANG, Z. 2015b. A High-Density SNP and SSR Consensus Map Reveals Segregation Distortion Regions in Wheat. *BioMed Research International,* 2015**,** 830618.

LI, H. 2011. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics,* 27.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LI, H., VIKRAM, P., SINGH, R. P., KILIAN, A., CARLING, J., SONG, J., BURGUENO-FERREIRA, J. A., BHAVANI, S., HUERTA-ESPINO, J., PAYNE, T., SEHGAL, D., WENZL, P. & SINGH, S. 2015. A high density GBS map of bread wheat and its application for dissecting complex disease resistance traits. *BMC Genomics,* 16**,** 216.

LI, J. Y., WANG, J. & ZEIGLER, R. S. 2014. The 3,000 rice genomes project: new opportunities and challenges for future rice research. *Gigascience,* 3.

LI, L., STOECKERT, C. J., JR. & ROOS, D. S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res,* 13**,** 2178-89.

LI, R., YU, C., LI, Y., LAM, T.-W., YIU, S.-M., KRISTIANSEN, K. & WANG, J. 2009c. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics,* 25**,** 1966-1967.

LI, R., ZHU, H., RUAN, J., QIAN, W., FANG, X., SHI, Z., LI, Y., LI, S., SHAN, G., KRISTIANSEN, K., LI, S., YANG, H., WANG, J. & WANG, J. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research,* 20**,** 265-272.

LI, S., JIA, J., WEI, X., ZHANG, X., LI, L., CHEN, H., FAN, Y., SUN, H., ZHAO, X., LEI, T., XU, Y., JIANG, F., WANG, H. & LI, L. 2007. A intervarietal genetic map and QTL analysis for yield traits in wheat. *Molecular Breeding,* 20**,** 167-178.

LI, W., ZHANG, P., FELLERS, J. P., FRIEBE, B. & GILL, B. S. 2004. Sequence composition, organization, and evolution of the core Triticeae genome. *The Plant Journal,* 40**,** 500-511.

LI, Y.-H., ZHOU, G., MA, J., JIANG, W., JIN, L.-G., ZHANG, Z., GUO, Y., ZHANG, J., SUI, Y., ZHENG, L., ZHANG, S.-S., ZUO, Q., SHI, X.-H., LI, Y.-F., ZHANG, W.-K., HU, Y., KONG, G., HONG, H.-L., TAN, B., SONG, J., LIU, Z.-X., WANG, Y., RUAN, H., YEUNG, C. K. L., LIU, J., WANG, H., ZHANG, L.-J., GUAN, R.-X., WANG, K.-J., LI, W.-B., CHEN, S.-Y., CHANG, R.-Z., JIANG, Z., JACKSON, S. A., LI, R. & QIU, L.-J. 2014. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotech,* 32**,** 1045-1052.

LIAO, H.-M., CHAO, Y.-L., HUANG, A.-L., CHENG, M.-C., CHEN, Y.-J., LEE, K.-F., FANG, J.-S., HSU, C.-H. & CHEN, C.-H. 2012. Identification and characterization of three inherited genomic copy number variations associated with familial schizophrenia. *Schizophrenia Research,* 139**,** 229-236.

LIAO, Y.-C., LIN, S.-H. & LIN, H.-H. 2015. Completing bacterial genome assemblies: strategy and performance comparisons. 5**,** 8747.

LING, H.-Q., ZHAO, S., LIU, D., WANG, J., SUN, H., ZHANG, C., FAN, H., LI, D., DONG, L., TAO, Y., GAO, C., WU, H., LI, Y., CUI, Y., GUO, X., ZHENG, S., WANG, B., YU, K., LIANG, Q., YANG, W., LOU, X., CHEN, J., FENG, M., JIAN, J., ZHANG, X., LUO, G., JIANG, Y., LIU, J., WANG, Z., SHA, Y., ZHANG, B., WU, H., TANG, D., SHEN, Q., XUE, P., ZOU, S., WANG, X., LIU, X., WANG, F., YANG, Y., AN, X., DONG, Z., ZHANG, K., ZHANG, X., LUO, M.-C., DVORAK, J., TONG, Y., WANG, J., YANG, H., LI, Z., WANG, D., ZHANG, A. & WANG, J. 2013. Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature,* 496**,** 87-90.

LIU, M., STILLER, J., HOLUŠOVÁ, K., VRÁNA, J., LIU, D., DOLEŽEL, J. & LIU, C. 2016. Chromosome-specific sequencing reveals an extensive dispensable genome component in wheat. *Scientific Reports,* 6**,** 36398.

LOBELL, D. B., SCHLENKER, W. & COSTA-ROBERTS, J. 2011. Climate Trends and Global Crop Production Since 1980. *Science,* 333**,** 616-620.

LOMAN, N. J., QUICK, J. & SIMPSON, J. T. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Meth,* 12**,** 733-735.

LONGIN, C. F. H. & REIF, J. C. 2014. Redesigning the exploitation of wheat genetic resources. *Trends in Plant Science,* 19**,** 631-636.

LOPES, M. S., EL-BASYONI, I., BAENZIGER, P. S., SINGH, S., ROYO, C., OZBEK, K., AKTAS, H., OZER, E., OZDEMIR, F., MANICKAVELU, A., BAN, T. & VIKRAM, P. 2015. Exploiting genetic diversity from landraces in wheat breeding for adaptation to climate change. *Journal of Experimental Botany,* 66**,** 3477-3486.

LORENC, M. T., HAYASHI, S., STILLER, J., LEE, H., MANOLI, S., RUPERAO, P., VISENDI, P., BERKMAN, P. J., LAI, K., BATLEY, J. & EDWARDS, D. 2012. Discovery of Single Nucleotide Polymorphisms in Complex Genomes Using SGSautoSNP. *Biology,* 1**,** 370-382.

LOVE, R. R., WEISENFELD, N. I., JAFFE, D. B., BESANSKY, N. J. & NEAFSEY, D. E. 2016. Evaluation of DISCOVAR de novo using a mosquito sample for cost-effective short-read genome assembly. *BMC Genomics,* 17**,** 187.

LU, F., ROMAY, M. C., GLAUBITZ, J. C., BRADBURY, P. J., ELSHIRE, R. J. & WANG, T. 2015. High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat Commun.,* 6.

LU, J., TANG, T., TANG, H., HUANG, J., SHI, S. & WU, C.-I. 2006. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends in Genetics,* 22**,** 126-131.

LU, P., HAN, X., QI, J., YANG, J., WIJERATNE, A. J., LI, T. & MA, H. 2012. Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. *Genome Research,* 22**,** 508-518.

LUKENS, L. N., PIRES, J. C., LEON, E., VOGELZANG, R., OSLACH, L. & OSBORN, T. 2006a. Patterns of Sequence Loss and Cytosine Methylation within a Population of Newly Resynthesized Brassica napus Allopolyploids. *Plant Physiology,* 140**,** 336-348.

LUKENS, L. N., PIRES, J. C., LEON, E. J., VOGELZANG, R., OSLACH, L. & OSBORN, T. C. 2006b. Patterns of sequence loss and cytosine methylation within a population of newly resynthesized Brassica napusallopolyploids. *Plant Physiology,* 140.

LUO, C., TSEMENTZI, D., KYRPIDES, N., READ, T. & KONSTANTINIDIS, K. T. 2012a. Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. *PLoS ONE,* 7**,** e30087.

LUO, M.-C., GU, Y. Q., YOU, F. M., DEAL, K. R., MA, Y. & HU, Y. 2013. A 4-gigabase physical map unlocks the structure and evolution of the complex genome of Aegilops tauschii, the wheat D-genome progenitor. *Proc Natl Acad Sci U S A,* 110.

LUO, M.-C., YANG, Z.-L., YOU, F. M., KAWAHARA, T., WAINES, J. G. & DVORAK, J. 2007. The structure of wild and domesticated emmer wheat populations, gene flow between them, and the site of emmer domestication. *Theoretical and Applied Genetics,* 114**,** 947-959.

LUO, R., LIU, B., XIE, Y., LI, Z., HUANG, W., YUAN, J., HE, G., CHEN, Y., PAN, Q., LIU, Y., TANG, J., WU, G., ZHANG, H., SHI, Y., LIU, Y., YU, C., WANG, B., LU, Y., HAN, C., CHEUNG, D., YIU, S.-M., PENG, S., XIAOQIAN, Z., LIU, G., LIAO, X., LI, Y., YANG, H., WANG, J., LAM, T.-W. & WANG, J. 2012b. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience,* 1.

MARCAIS, G. & KINGSFORD, C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics,* 27**,** 764-70.

MARÇAIS, G., YORKE, J. A. & ZIMIN, A. 2015. QuorUM: An Error Corrector for Illumina Reads. *PLOS ONE,* 10**,** e0130821.

MARCUS, S., LEE, H. & SCHATZ, M. C. 2014. SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips. *Bioinformatics,* 30**,** 3476-3483.

MARCUSSEN, T., SANDVE, S. R., HEIER, L., SPANNAGL, M., PFEIFER, M. & JAKOBSEN, K. S. 2014. Ancient hybridizations among the ancestral genomes of bread wheat. *Science,* 345.

MARIETTE, J., NOIROT, C. & KLOPP, C. 2011. Assessment of replicate bias in 454 pyrosequencing and a multi-purpose read-filtering tool. *BMC Research Notes,* 4**,** 149.

MARTIN, J. M., TALBERT, L. E., LANNING, S. P. & BLAKE, N. K. 1995. Hybrid Performance in Wheat as Related to Parental Diversity. *Crop Science,* 35**,** 104-108.

MARTINEZ-PEREZ, E., SHAW, P. & MOORE, G. 2001. The Ph1 locus is needed to ensure specific somatic and meiotic centromere association. *Nature,* 411**,** 204-207.

MASCHER, M., MUEHLBAUER, G. J., ROKHSAR, D. S., CHAPMAN, J., SCHMUTZ, J., BARRY, K., MUNOZ-AMATRIAIN, M., CLOSE, T. J., WISE, R. P., SCHULMAN, A. H., HIMMELBACH, A., MAYER, K. F., SCHOLZ, U., POLAND, J. A., STEIN, N. & WAUGH, R. 2013. Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *Plant J,* 76**,** 718-27.

MAYER, K. F., ROGERS, J., DOLEŽEL, J., POZNIAK, C., EVERSOLE, K. & FEUILLET, C. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science.,* 345.

MAYER, K. F., WAUGH, R., BROWN, J. W., SCHULMAN, A. & LANGRIDGE, P. 2012. A physical, genetic and functional sequence assembly of the barley genome. *Nature,* 491.

MAYER, K. F. X., MARTIS, M., HEDLEY, P. E., ŠIMKOVÁ, H., LIU, H., MORRIS, J. A., STEUERNAGEL, B., TAUDIEN, S., ROESSNER, S., GUNDLACH, H., KUBALÁKOVÁ, M., SUCHÁNKOVÁ, P., MURAT, F., FELDER, M., NUSSBAUMER, T., GRANER, A., SALSE, J., ENDO, T., SAKAI, H., TANAKA, T., ITOH, T., SATO, K., PLATZER, M., MATSUMOTO, T., SCHOLZ, U., DOLEŽEL, J., WAUGH, R. & STEIN, N. 2011. Unlocking the Barley Genome by Chromosomal and Comparative Genomics. *The Plant Cell,* 23**,** 1249-1263.

MAYER, K. F. X., TAUDIEN, S., MARTIS, M., SIMKOVÁ, H., SUCHÁNKOVÁ, P. & GUNDLACH, H. 2009. Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol,* 151.

MCCARROLL, S. A. & ALTSHULER, D. M. 2007. Copy-number variation and association studies of human disease. *Nat Genet*.

MCCARTHY, A. 2010. Third Generation DNA Sequencing: Pacific Biosciences' Single Molecule Real Time Technology. *Chemistry & Biology,* 17**,** 675-676.

MCFADDEN, E. S. & SEARS, E. R. 1946. The origin of Triticum spelta and its free-threshing hexaploid relatives. *J Hered,* 37.

MEDINI, D., DONATI, C., TETTELIN, H., MASIGNANI, V. & RAPPUOLI, R. 2005. The microbial pan-genome. *Current Opinion in Genetics & Development,* 15**,** 589-594.

MEISSNER, A., GNIRKE, A., BELL, G. W., RAMSAHOYE, B., LANDER, E. S. & JAENISCH, R. 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research,* 33**,** 5868-5877.

MENGISTU, D. K., KIDANE, Y. G., CATELLANI, M., FRASCAROLI, E., FADDA, C., PÈ, M. E. & DELL'ACQUA, M. 2016. High-density molecular characterization and association mapping in Ethiopian durum wheat landraces reveals high diversity and potential for wheat breeding. *Plant Biotechnology Journal,* 14**,** 1800-1812.

METTE, M. F., GILS, M., LONGIN, C. F. H. & REIF, J. C. 2015. Hybrid Breeding in Wheat. *In:* OGIHARA, Y., TAKUMI, S. & HANDA, H. (eds.) *Advances in Wheat Genetics: From Genome to Field: Proceedings of the 12th International Wheat Genetics Symposium.* Tokyo: Springer Japan.

MEYER, K. D. & JAFFREY, S. R. 2016. Expanding the diversity of DNA base modifications with N 6-methyldeoxyadenosine. *Genome Biology,* 17**,** 5.

MICHAEL, T. P. & JACKSON, S. 2013. The First 50 Plant Genomes. *The Plant Genome,* 6.

MIDDLETON, C. P., SENERCHIA, N., STEIN, N., AKHUNOV, E. D., KELLER, B., WICKER, T. & KILIAN, B. 2014. Sequencing of chloroplast genomes from wheat, barley, rye and their relatives provides a detailed insight into the evolution of the Triticeae tribe. *PLoS One,* 9**,** e85761.

MIKHEYEV, A. S. & TIN, M. M. Y. 2014. A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources,* 14**,** 1097-1102.

MOCHIDA, K., SAKURAI, T., SEKI, H., YOSHIDA, T., TAKAHAGI, K., SAWAI, S., UCHIYAMA, H., MURANAKA, T. & SAITO, K. 2017. Draft genome assembly and annotation of Glycyrrhiza uralensis, a medicinal legume. *The Plant Journal,* 89**,** 181-194.

MOORE, R. C. & PURUGGANAN, M. D. 2005. The evolutionary dynamics of plant duplicate genes. *Current Opinion in Plant Biology,* 8**,** 122-128.

MOVAHEDI, N. S., EMBREE, M., NAGARAJAN, H., ZENGLER, K. & CHITSAZ, H. 2016. Efficient Synergistic Single-Cell Genome Assembly. *Frontiers in Bioengineering and Biotechnology,* 4.

MUHINDIRA, P. V. 2016. A novel approach for the assembly of complex genomic DNA cloned into bacterial artificial chromosome vectors: assembly and analysis of Triticum aestivum chromosome arm 7DS. The University of Queensland, School of Agriculture and Food Sciences.

MUJEEB-KAZI, A., GUL, A., FAROOQ, M., RIZWAN, S. & AHMAD, I. 2008. Rebirth of synthetic hexaploids with global implications for wheat improvement. *Australian Journal of Agricultural Research,* 59**,** 391-398.

MURAT, F., XU, J.-H., TANNIER, E., ABROUK, M., GUILHOT, N., PONT, C., MESSING, J. & SALSE, J. 2010. Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Research,* 20**,** 1545-1557.

MYERS, E. W., SUTTON, G. G., DELCHER, A. L., DEW, I. M., FASULO, D. P. & FLANIGAN, M. J. 2000. A whole-genome assembly of Drosophila. *Science,* 287.

NEALE, D. B., WEGRZYN, J. L., STEVENS, K. A., ZIMIN, A. V., PUIU, D., CREPEAU, M. W., CARDENO, C., KORIABINE, M., HOLTZ-MORRIS, A. E., LIECHTY, J. D., MARTÍNEZ-GARCÍA, P. J., VASQUEZ-GROSS, H. A., LIN, B. Y., ZIEVE, J. J., DOUGHERTY, W. M., FUENTES-SORIANO, S., WU, L.-S., GILBERT, D., MARÇAIS, G., ROBERTS, M., HOLT, C., YANDELL, M., DAVIS, J. M., SMITH, K. E., DEAN, J. F., LORENZ, W. W., WHETTEN, R. W., SEDEROFF, R., WHEELER, N., MCGUIRE, P. E., MAIN, D., LOOPSTRA, C. A., MOCKAITIS, K., DEJONG, P. J., YORKE, J. A., SALZBERG, S. L. & LANGLEY, C. H. 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology,* 15**,** R59.

NEPH, S., KUEHN, M. S., REYNOLDS, A. P., HAUGEN, E., THURMAN, R. E., JOHNSON, A. K., RYNES, E., MAURANO, M. T., VIERSTRA, J., THOMAS, S., SANDSTROM, R., HUMBERT, R. & STAMATOYANNOPOULOS, J. A. 2012. BEDOPS: high-performance genomic feature operations. *Bioinformatics,* 28**,** 1919-1920.

NEWELL, M. A. & JANNINK, J.-L. 2014. Genomic Selection in Plant Breeding. *In:* FLEURY, D. & WHITFORD, R. (eds.) *Crop Breeding: Methods and Protocols.* New York, NY: Springer New York.

NRGENE. 2016. *DeNovoMAGIC2.0* [Online]. Available: http://nrgene.com/products-technology/denovomagic/ [Accessed 11-11 2016].

NRGENE. 2017. *DeNovoMagic* [Online]. Available: http://nrgene.com/ [Accessed 2017 2017].

NURK, S. M., DMITRY; KOROBEYNIKOV, ANTON; PEVZNER, PAVEL 2016. metaSPAdes: a new versatile de novo metagenomics assembler. *bioarXiv*.

O'MARA, J. G. 1951. Cytogenetic Studies on Triticale II

The kinds of intergeneric chromosome addition. *CYTOLOGIA,* 16**,** 225-232.

O'MARA, J. G. 1953. The cytogenetics of Triticale. *The Botanical Review,* 19**,** 587-605.

OGIHARA, Y. & TSUNEWAKI, K. 1988. Diversity and evolution of chloroplast DNA in Triticum and Aegilops as revealed by restriction fragment analysis. *Theoretical and Applied Genetics,* 76**,** 321-332.

OSSOWSKI, S., SCHNEEBERGER, K., CLARK, R. M., LANZ, C., WARTHMANN, N. & WEIGEL, D. 2008. Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res,* 18**,** 2024-33.

OZKAN, H., BRANDOLINI, A., SCHAFER-PREGL, R. & SALAMINI, F. 2002. AFLP analysis of a collection of tetraploid wheats indicates the origin of emmer and hard wheat domestication in southeast Turkey. *Mol Biol Evol,* 19**,** 1797-801.

PANKRATZ, N., DUMITRIU, A., HETRICK, K. N., SUN, M., LATOURELLE, J. C., WILK, J. B., HALTER, C., DOHENY, K. F., GUSELLA, J. F., NICHOLS, W. C., MYERS, R. H., FOROUD, T., DESTEFANO, A. L., THE, P. P., GENEPD INVESTIGATORS, C. & MOLECULAR GENETIC, L. 2011. Copy Number Variation in Familial Parkinson Disease. *PLOS ONE,* 6**,** e20988.

PARKIN, I. A., KOH, C., TANG, H., ROBINSON, S. J., KAGALE, S., CLARKE, W. E., TOWN, C. D., NIXON, J., KRISHNAKUMAR, V., BIDWELL, S. L., DENOEUD, F., BELCRAM, H., LINKS, M. G., JUST, J., CLARKE, C., BENDER, T., HUEBERT, T., MASON, A. S., PIRES, C. J., BARKER, G., MOORE, J., WALLEY, P. G., MANOLI, S., BATLEY, J., EDWARDS, D., NELSON, M. N., WANG, X., PATERSON, A. H., KING, G., BANCROFT, I., CHALHOUB, B. & SHARPE, A. G. 2014. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. *Genome Biol,* 15**,** R77.

PARRA, G., BRADNAM, K. & KORF, I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics,* 23**,** 1061-7.

PARRA, G., BRADNAM, K., NING, Z., KEANE, T. & KORF, I. 2009. Assessing the gene space in draft genomes. *Nucleic Acids Research,* 37**,** 289-297.

PATERSON, A. H., BOWERS, J. E., BRUGGMANN, R., DUBCHAK, I., GRIMWOOD, J., GUNDLACH, H., HABERER, G., HELLSTEN, U., MITROS, T., POLIAKOV, A., SCHMUTZ, J., SPANNAGL, M., TANG, H., WANG, X., WICKER, T., BHARTI, A. K., CHAPMAN, J., FELTUS, F. A., GOWIK, U., GRIGORIEV, I. V., LYONS, E., MAHER, C. A., MARTIS, M., NARECHANIA, A., OTILLAR, R. P., PENNING, B. W., SALAMOV, A. A., WANG, Y., ZHANG, L., CARPITA, N. C., FREELING, M., GINGLE, A. R., HASH, C. T., KELLER, B., KLEIN, P., KRESOVICH, S., MCCANN, M. C., MING, R., PETERSON, D. G., MEHBOOB UR, R., WARE, D., WESTHOFF, P., MAYER, K. F. X., MESSING, J. & ROKHSAR, D. S. 2009. The Sorghum bicolor genome and the diversification of grasses. *Nature,* 457**,** 551-556.

PATERSON AH, W. X., LI J, TANG H. 2012. Ancient and recent polyploidy in monocots. *In:* SOLTIS P, S. D. (ed.) *Polyploidy and genome evolution.* Berlin: Springer.

PAUX, E., ROGER, D., BADAEVA, E., GAY, G., BERNARD, M., SOURDILLE, P. & FEUILLET, C. 2006. Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. *Plant J,* 48**,** 463-74.

PAUX, E., SOURDILLE, P., SALSE, J., SAINTENAC, C., CHOULET, F., LEROY, P., KOROL, A., MICHALAK, M., KIANIAN, S., SPIELMEYER, W., LAGUDAH, E., SOMERS, D., KILIAN, A., ALAUX, M., VAUTRIN, S., BERGÈS, H., EVERSOLE, K., APPELS, R., SAFAR, J., SIMKOVA, H., DOLEZEL, J., BERNARD, M. & FEUILLET, C. 2008. A Physical Map of the 1-Gigabase Bread Wheat Chromosome 3B. *Science,* 322**,** 101-104.

PENG, J., SUN, D. & NEVO, E. 2011. Domestication evolution, genetics and genomics in wheat. *Molecular Breeding,* 28**,** 281-301.

PENG, J., ZADEH, H., LAZO, G., GUSTAFSON, J., CHAO, S., ANDERSON, O., QI, L., ECHALIER, B., GILL, B. & DILBIRLIGI, M. 2004. Chromosome bin map of expressed sequence tags in homoeologous group 1 of hexaploid wheat and homoeology with rice and Arabidopsis. *Genetics,* 168.

PENG, Y., LEUNG, H. C. M., YIU, S. M. & CHIN, F. Y. L. 2010. IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler. *In:* BERGER, B. (ed.) *Research in Computational Molecular Biology: 14th Annual International Conference, RECOMB 2010, Lisbon, Portugal, April 25-28, 2010. Proceedings.* Berlin, Heidelberg: Springer Berlin Heidelberg.

PENG, Y., LEUNG, H. C. M., YIU, S. M. & CHIN, F. Y. L. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics,* 28**,** 1420-1428.

PESTSOVA, E. G., BORNER, A. & RODER, M. S. 2001. Development of a set of Triticum aestivum-Aegilops tauschii introgression lines. *Hereditas,* 135**,** 139-43.

PETERSON, D. G., SCHULZE, S. R., SCIARA, E. B., LEE, S. A., BOWERS, J. E., NAGEL, A., JIANG, N., TIBBITTS, D. C., WESSLER, S. R. & PATERSON, A. H. 2002. Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Res,* 12**,** 795-807.

PEVZNER, P. A., TANG, H. & WATERMAN, M. S. 2001. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci USA,* 98.

PFEIFER, M., KUGLER, K. G., SANDVE, S. R., ZHAN, B., RUDI, H. & HVIDSTEN, T. R. 2014. Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science.,* 345.

PFEIFER, M., MARTIS, M., ASP, T., MAYER, K. F., LUBBERSTEDT, T., BYRNE, S., FREI, U. & STUDER, B. 2013. The perennial ryegrass GenomeZipper: targeted use of genome resources for comparative grass genomics. *Plant Physiol,* 161**,** 571-82.

The Potato Genome Sequenecing Consortium. 2011. Genome sequence and analysis of the tuber crop potato. Nature 475, 189–195. doi:10.1038/nature10158.

PHANSTIEL, D. H., BOYLE, A. P., ARAYA, C. L. & SNYDER, M. P. 2014. Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics,* 30**,** 2808-10.

PHILLIPPY, A. M. 2017. New advances in sequence assembly. *Genome Research,* 27**,** xi-xiii.

PLASCHKE, J., GANAL, M. W. & RODER, M. S. 1995a. Detection of genetics diversity in closely-related bread wheat using microsatellite markers. *Theoretical and Applied Genetics,* 91**,** 1001-1007.

POLAND, J., ENDELMAN, J., DAWSON, J., RUTKOSKI, J., WU, S., MANES, Y., DREISIGACKER, S., CROSSA, J., SÁNCHEZ-VILLEDA, H., SORRELLS, M. & JANNINK, J.-L. 2012a. Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *The Plant Genome,* 5**,** 103-113.

POLAND, J. A., BROWN, P. J., SORRELLS, M. E. & JANNINK, J.-L. 2012b. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLOS ONE,* 7**,** e32253.

PONT, C., MURAT, F., GUIZARD, S., FLORES, R., FOUCRIER, S., BIDET, Y., QURAISHI, U. M., ALAUX, M., DOLEŽEL, J., FAHIMA, T., BUDAK, H., KELLER, B., SALVI, S., MACCAFERRI, M., STEINBACH, D., FEUILLET, C., QUESNEVILLE, H. & SALSE, J. 2013. Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. *The Plant Journal,* 76**,** 1030-1044.

POP, M., KOSACK, D. S. & SALZBERG, S. L. 2004. Hierarchical Scaffolding With Bambus. *Genome Research,* 14**,** 149-159.

PRASAD, M., VARSHNEY, R. K., ROY, J. K., BALYAN, H. S. & GUPTA, P. K. 2000. The use of microsatellites for detecting DNA polymorphism, genotype identification and genetic diversity in wheat. *Theoretical and Applied Genetics,* 100**,** 584-592.

PRINCE, V. E. & PICKETT, F. B. 2002. Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet,* 3**,** 827-837.

QI, L. L., ECHALIER, B., CHAO, S., LAZO, G. R., BUTLER, G. E., ANDERSON, O. D., AKHUNOV, E. D., DVOŘÁK, J., LINKIEWICZ, A. M., RATNASIRI, A., DUBCOVSKY, J., BERMUDEZ-KANDIANIS, C. E., GREENE, R. A., KANTETY, R., LA ROTA, C. M., MUNKVOLD, J. D., SORRELLS, S. F., SORRELLS, M. E., DILBIRLIGI, M., SIDHU, D., ERAYMAN, M., RANDHAWA, H. S., SANDHU, D., BONDAREVA, S. N., GILL, K. S., MAHMOUD, A. A., MA, X.-F., MIFTAHUDIN, GUSTAFSON, J. P., CONLEY, E. J., NDUATI, V., GONZALEZ-HERNANDEZ, J. L., ANDERSON, J. A., PENG, J. H., LAPITAN, N. L. V., HOSSAIN, K. G., KALAVACHARLA, V., KIANIAN, S. F., PATHAN, M. S., ZHANG, D. S., NGUYEN, H. T., CHOI, D.-W., FENTON, R. D., CLOSE, T. J., MCGUIRE, P. E., QUALSET, C. O. & GILL, B. S. 2004. A Chromosome Bin Map of 16,000 Expressed Sequence Tag Loci and Distribution of Genes Among the Three Genomes of Polyploid Wheat. *Genetics,* 168**,** 701-712.

QUAIL, M. A., SMITH, M., COUPLAND, P., OTTO, T. D., HARRIS, S. R., CONNOR, T. R., BERTONI, A., SWERDLOW, H. P. & GU, Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics,* 13**,** 341.

QUARRIE, S. A., STEED, A., CALESTANI, C., SEMIKHODSKII, A., LEBRETON, C., CHINOY, C., STEELE, N., PLJEVLJAKUSIĆ, D., WATERMAN, E., WEYEN, J., SCHONDELMAIER, J., HABASH, D. Z., FARMER, P., SAKER, L., CLARKSON, D. T., ABUGALIEVA, A., YESSIMBEKOVA, M., TURUSPEKOV, Y., ABUGALIEVA, S., TUBEROSA, R., SANGUINETI, M.-C., HOLLINGTON, P. A., ARAGUÉS, R.,

ROYO, A. & DODIG, D. 2005. A high-density genetic map of hexaploid wheat (Triticum aestivum L.) from the cross Chinese Spring × SQ1 and its use to compare QTLs for grain yield across a range of environments. *Theoretical and Applied Genetics,* 110**,** 865-880.

R CORE TEAM 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing.

RABINOWICZ, P. D., SCHUTZ, K., DEDHIA, N., YORDAN, C., PARNELL, L. D., STEIN, L., MCCOMBIE, W. R. & MARTIENSSEN, R. A. 1999. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat Genet,* 23**,** 305-8.

RAMAN, H., DALTON-MORGAN, J., DIFFEY, S., RAMAN, R., ALAMERY, S., EDWARDS, D. & BATLEY, J. 2014. SNP markers-based map construction and genome-wide linkage analysis in Brassica napus. *Plant Biotechnology Journal,* 12**,** 851-860.

RAMSEY, J. & SCHEMSKE, D. W. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu Rev Ecol Syst.,* 29.

RAND, A. C., JAIN, M., EIZENGA, J. M., MUSSELMAN-BROWN, A., OLSEN, H. E., AKESON, M. & PATEN, B. 2017. Mapping DNA methylation with high-throughput nanopore sequencing. *Nat Meth,* 14**,** 411-413.

RASHEED, A., XIA, X., OGBONNAYA, F., MAHMOOD, T., ZHANG, Z. & MUJEEB-KAZI, A. 2014. Genome-wide association for grain morphology in synthetic hexaploid wheats using digital imaging analysis. *BMC Plant Biol,* 14.

RATEL, D., RAVANAT, J. L., CHARLES, M. P., PLATET, N., BREUILLAUD, L. & LUNARDI, J. 2006. Undetectable levels of N6-methyl adenine in mouse DNA: cloning and analysis of PRED28, a gene coding for a putative mammalian DNA adenine methyltransferase. *FEBS Lett,* 580.

RAVEL, C., PRAUD, S., MURIGNEUX, A., CANAGUIER, A., SAPET, F., SAMSON, D., BALFOURIER, F., DUFOUR, P., CHALHOUB, B., BRUNEL, D., BECKERT, M. & CHARMET, G. 2006. Single-nucleotide polymorphism frequency in a set of selected lines of bread wheat (Triticum aestivum L.). *Genome,* 49**,** 1131-1139.

REBETZKE, G. J., CHAPMAN, S. C., MCINTYRE, C. L., RICHARDS, R. A., CONDON, A. G., WATT, M. & VAN HERWAARDEN, A. F. 2009. Grain Yield Improvement in Water-Limited Environments. *Wheat Science and Trade.* Wiley-Blackwell.

REIF, J. C., ZHANG, P., DREISIGACKER, S., WARBURTON, M. L., VAN GINKEL, M., HOISINGTON, D., BOHN, M. & MELCHINGER, A. E. 2005. Wheat genetic diversity trends during domestication and breeding. *Theor Appl Genet,* 110**,** 859-64.

REYNOLDS, M., BONNETT, D., CHAPMAN, S. C., FURBANK, R. T., MANÈS, Y., MATHER, D. E. & PARRY, M. A. J. 2011. Raising yield potential of wheat. I. Overview of a consortium approach and breeding strategies. *Journal of Experimental Botany,* 62**,** 439-452.

RHOADS, A. & AU, K. F. 2015. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics,* 13**,** 278-289.

RIZZON, C., PONGER, L. & GAUT, B. S. 2006. Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice. *PLoS Comput Biol,* 2**,** e115.

ROBBINS, A. M. 2009. *Dwarfing genes in spring wheat: An agronomic comparison of Rht-B1, Rht-D1, and Rht8.* Montana State University-Bozeman, College of Agriculture.

ROBERTS, R. J., CARNEIRO, M. O. & SCHATZ, M. C. 2013. The advantages of SMRT sequencing. *Genome Biol,* 14.

RÖDER, M. S., KORZUN, V., WENDEHAKE, K., PLASCHKE, J., TIXIER, M.-H., LEROY, P. & GANAL, M. W. 1998. A Microsatellite Map of Wheat. *Genetics,* 149**,** 2007-2023.

ROTHBERG, J. M. & LEAMON, J. H. 2008. The development and impact of 454 sequencing. *Nat Biotech,* 26**,** 1117-1124.

ROUGEMONT, J., AMZALLAG, A., ISELI, C., FARINELLI, L., XENARIOS, I. & NAEF, F. 2008. Probabilistic base calling of Solexa sequencing data. *BMC Bioinformatics,* 9**,** 431.

ROULI, L., MERHEJ, V., FOURNIER, P. E. & RAOULT, D. 2015. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections,* 7**,** 72-85.

RUPERAO, P., CHAN, C.-K. K., AZAM, S., KARAFIÁTOVÁ, M., HAYASHI, S., ČÍŽKOVÁ, J., SAXENA, R. K., ŠIMKOVÁ, H., SONG, C., VRÁNA, J., CHITIKINENI, A., VISENDI, P., GAUR, P. M., MILLÁN, T., SINGH, K. B., TARAN, B., WANG, J., BATLEY, J., DOLEŽEL, J., VARSHNEY, R. K. & EDWARDS, D. 2014. A chromosomal genomics approach to assess and validate the desi and kabuli draft chickpea genome assemblies. *Plant Biotechnology Journal,* 12**,** 778-786.

ŠAFÁŘ, J., BARTOŠ, J., JANDA, J., BELLEC, A., KUBALÁKOVÁ, M., VALARIK, M., PATEYRON, S., WEISEROVA, J., TUSKOVA, R. & ČÍHALÍKOVÁ, J. 2004. Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. *Plant J,* 39.

ŠAFÁŘ, J., ŠIMKOVÁ, H., KUBALÁKOVÁ, M., ČÍHALÍKOVÁ, J., SUCHÁNKOVÁ, P., BARTOŠ, J. & DOLEŽEL, J. 2010. Development of Chromosome-Specific BAC Resources for Genomics of Bread Wheat. *Cytogenetic and Genome Research,* 129**,** 211-223.

SAINTENAC, C., JIANG, D. & AKHUNOV, E. D. 2011. Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Genome Biol.,* 12.

SAKAMURA, T. 1918. Kurze Mitteilung über die Chromosomenzahlen und die Verwandtschaftsverhältnisse der Triticum-Arten. *Botanical Magazine,* 32**,** 151-154.

SALLAM, A. H., ENDELMAN, J. B., JANNINK, J. L. & SMITH, K. P. 2015. Assessing genomic selection prediction accuracy in a dynamic barley breeding population. *Plant Genome,* 8.

SALMELA, L., WALVE, R., RIVALS, E. & UKKONEN, E. 2017. Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics,* 33**,** 799-806.

SALMON, A., AINOUCHE, M. L. & WENDEL, J. F. 2005. Genetic and epigenetic consequences of recent hybridization and polyploidy in Spartina (Poaceae). *Molecular Ecology,* 14**,** 1163-1175.

SALZBERG, S. L. & YORKE, J. A. 2005. Beware of mis-assembled genomes. *Bioinformatics,* 21**,** 4320-4321.

SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences,* 74**,** 5463-5467.

SANTUARI, L., PRADERVAND, S., AMIGUET-VERCHER, A.-M., THOMAS, J., DORCEY, E., HARSHMAN, K., XENARIOS, I., JUENGER, T. E. & HARDTKE, C. S. 2010. Substantial deletion overlap among divergent Arabidopsis genomes revealed by intersection of short reads and tiling arrays. *Genome Biology,* 11**,** R4-R4.

SARKAR, P. & STEBBINS, G. L. 1956. Morphological Evidence Concerning the Origin of the B Genome in Wheat. *American Journal of Botany,* 43**,** 297-304.

SATYA, P., PASWAN, P. K., GHOSH, S., MAJUMDAR, S. & ALI, N. 2016. Confamiliar transferability of simple sequence repeat (SSR) markers from cotton (Gossypium hirsutum L.) and jute (Corchorus olitorius L.) to twenty two Malvaceous species. *3 Biotech,* 6**,** 65.

SAXENA, R. K., EDWARDS, D. & VARSHNEY, R. K. 2014. Structural variations in plant genomes. *Briefings in functional genomics,* 13**,** 296-307.

SCHATZ, M. C., MARON, L. G., STEIN, J. C., WENCES, A. H., GURTOWSKI, J., BIGGERS, E., LEE, H., KRAMER, M., ANTONIOU, E., GHIBAN, E., WRIGHT, M. H., CHIA, J.-M., WARE, D., MCCOUCH, S. R. & MCCOMBIE, W. R. 2014. Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. *Genome Biology,* 15**,** 1-16.

SCHIRRMEISTER, B. E., DALQUEN, D. A., ANISIMOVA, M. & BAGHERI, H. C. 2012. Gene copy number variation and its significance in cyanobacterial phylogeny. *BMC Microbiology,* 12**,** 177.

SCHNABLE, J. C., SPRINGER, N. M. & FREELING, M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad Sci U S A.,* 108.

SCHOEN, C., BLOM, J., CLAUS, H., SCHRAMM-GLUCK, A., BRANDT, P., MULLER, T., GOESMANN, A., JOSEPH, B., KONIETZNY, S., KURZAI, O., SCHMITT, C., FRIEDRICH, T., LINKE, B., VOGEL, U. & FROSCH, M. 2008. Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in Neisseria meningitidis. *Proc Natl Acad Sci U S A,* 105**,** 3473-8.

SCOTT, D. & ELY, B. 2015. Comparison of Genome Sequencing Technology and Assembly Methods for the Analysis of a GC-Rich Bacterial Genome. *Current Microbiology,* 70**,** 338-344.

SEARS, E. 1966. Nullisomic-Tetrasomic Combinations in Hexaploid Wheat. *In:* RILEY, R. & LEWIS, K. R. (eds.) *Chromosome manipulations and plant genetics.*

SEARS, E. R. 1969. Wheat cytogenetics. *Annual Review of Genetics,* 3**,** 451-468.

SEARS, E. R. & MILLER, T. E. 1985. THE HISTORY OF CHINESE SPRING WHEAT. *Cereal Research Communications,* 13**,** 261-263.

SEBAT, J., LAKSHMI, B., TROGE, J., ALEXANDER, J., YOUNG, J., LUNDIN, P., MÅNÉR, S., MASSA, H., WALKER, M., CHI, M., NAVIN, N., LUCITO, R., HEALY, J., HICKS, J., YE, K., REINER, A., GILLIAM, T. C., TRASK, B., PATTERSON, N., ZETTERBERG, A. & WIGLER, M. 2004. Large-Scale Copy Number Polymorphism in the Human Genome. *Science,* 305**,** 525-528.

SEEMANN, S. G. A. T. 2012. VelvetOptimiser is a multi-threaded Perl script for automatically optimising the three primary parameter options (K, -exp_cov, -cov_cutoff) for the Velvet de novo sequence assembler. *Victorian Bioinformatics Consortium.*

SEHGAL, S. K., LI, W., RABINOWICZ, P. D., CHAN, A., SIMKOVÁ, H. & DOLEŽEL, J. 2012. Chromosome arm-specific BAC end sequences permit comparative analysis

of homoeologous chromosomes and genomes of polyploid wheat. *BMC Plant Biol,* 12.

SEMAGN, K., BJØRNSTAD, Å., SKINNES, H., MARØY, A. G., TARKEGNE, Y. & WILLIAM, M. 2006. Distribution of DArT, AFLP, and SSR markers in a genetic linkage map of a doubled-haploid hexaploid wheat population. *Genome,* 49**,** 545-555.

INTERNATIONAL RICE GENOME SEUQENCING PROJECT (IRGSP). 2005. The map-based sequence of the rice genome. *Nature,* 436**,** 793-800.

SHAKED, H., KASHKUSH, K., OZKAN, H., FELDMAN, M. & LEVY, A. A. 2001. Sequence Elimination and Cytosine Methylation Are Rapid and Reproducible Responses of the Genome to Wide Hybridization and Allopolyploidy in Wheat. *The Plant Cell,* 13**,** 1749-1759.

SHAPIRO, R., SERVIS, R. E. & WELCHER, M. 1970. Reactions of Uracil and Cytosine Derivatives with Sodium Bisulfite. *Journal of the American Chemical Society,* 92**,** 422-424.

SHARP, P. J., KREIS, M., SHEWRY, P. R. & GALE, M. D. 1988. Location of β-amylase sequences in wheat and its relatives. *Theoretical and Applied Genetics.,* 75**,** 286-290.

SHEIKHIZADEH, S., SCHRANZ, M. E., AKDEL, M., DE RIDDER, D. & SMIT, S. 2016. PanTools: representation, storage and exploration of pan-genomic data. *Bioinformatics,* 32**,** i487-i493.

SIEDLER, H., MESSMER, M. M., SCHACHERMAYR, G. M., WINZELER, H., WINZELER, M. & KELLER, B. 1994. Genetic diversity in European wheat and spelt breeding material based on RFLP data. *Theoretical and Applied Genetics,* 88**,** 994-1003.

SIMÃO, F. A., WATERHOUSE, R. M., IOANNIDIS, P., KRIVENTSEVA, E. V. & ZDOBNOV, E. M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*.

SIMEÃO RESENDE, R. M., CASLER, M. D. & VILELA DE RESENDE, M. D. 2014. Genomic Selection in Forage Breeding: Accuracy and Methods. *Crop Science,* 54**,** 143-156.

ŠIMKOVÁ, H., JANDA, J., HŘIBOVÁ, E., ŠAFÁŘ, J., DOLEŽEL, J. 2007. Cot-based cloning and sequencing of the short arm of wheat chromosome 1B *Plant , soil and environment,* 53**,** 437-441.

SIMPSON, J. T., WONG, K., JACKMAN, S. D., SCHEIN, J. E., JONES, S. J. M. & BIROL, I. 2009. ABySS: A parallel assembler for short read sequence data. *Genome Research,* 19**,** 1117-1123.

SIMPSON, J. T., WORKMAN, R. E., ZUZARTE, P. C., DAVID, M., DURSI, L. J. & TIMP, W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Meth,* 14**,** 407-410.

SINGH, R., SHEORAN, S., SHARMA, P. & CHATRATH, R. 2011. Analysis of simple sequence repeats (SSRs)dynamics in fungus Fusarium graminearum. *Bioinformation,* 5**,** 402-404.

SMET, R., ADAMS, K. L., VANDEPOELE, K., MONTAGU, M. C., MAERE, S. & PEER, Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci U S A.,* 110.

SMIT, A., HUBLEY, R. & GREEN, P. 2015. RepeatMasker Open-4.0. *http://www.repeatmasker.org*.

SMITH, D. B. & FLAVELL, R. B. 1975. Characterisation of the wheat genome by renaturation kinetics. *Chromosoma,* 50**,** 223-242.

SMITH, D. B., RIMPAU, J. & FLAVELL, R. B. 1976. Interspersion of different repeated sequences in the wheat genome revealed by interspecies DNA/DNA hybridisation. *Nucleic Acids Research,* 3**,** 2811-2825.

SMITH, R. B. F. A. D. B. 1976. Nucleotide sequence organisation in the wheat genome. *Heredity,* 37**,** 231-252.

SOLTIS, P. S. & SOLTIS, D. E. 2012. *Polyploidy and genome evolution,* Berlin, Springer.

SOMERS, D. J., ISAAC, P. & EDWARDS, K. 2004. A high-density microsatellite consensus map for bread wheat (Triticum aestivum L.). *Theoretical and Applied Genetics,* 109**,** 1105-1114.

SONAH, H., BASTIEN, M., IQUIRA, E., TARDIVEL, A., LÉGARÉ, G., BOYLE, B., NORMANDEAU, É., LAROCHE, J., LAROSE, S., JEAN, M. & BELZILE, F. 2013. An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. *PLOS ONE,* 8**,** e54603.

SONG, Q. J., SHI, J. R., SINGH, S., FICKUS, E. W., COSTA, J. M., LEWIS, J., GILL, B. S., WARD, R. & CREGAN, P. B. 2005. Development and mapping of microsatellite (SSR) markers in wheat. *Theoretical and Applied Genetics,* 110**,** 550-560.

SORRELLS, M. E., GUSTAFSON, J. P., SOMERS, D., CHAO, S., BENSCHER, D., GUEDIRA-BROWN, G., HUTTNER, E., KILIAN, A., MCGUIRE, P. E., ROSS, K.,

TANAKA, J., WENZL, P., WILLIAMS, K. & QUALSET, C. O. 2011. Reconstruction of the Synthetic W7984 × Opata M85 wheat reference population. *Genome,* 54**,** 875-882.

SOURDILLE, P., SINGH, S., CADALEN, T., BROWN-GUEDIRA, G., GAY, G., QI, L., QI, L. L., DUFOUR, P., MURIGNEUX, A. & BERNARD, M. 2004. Microsatellite-based deletion bin system for the establishment of genetic-physical map relationships in wheat (Triticum aestivum L.). *Funct Integr Genomics,* 4.

SPRINGER, N. M., YING, K., FU, Y., JI, T., YEH, C.-T., JIA, Y., WU, W., RICHMOND, T., KITZMAN, J., ROSENBAUM, H., INIGUEZ, A. L., BARBAZUK, W. B., JEDDELOH, J. A., NETTLETON, D. & SCHNABLE, P. S. 2009. Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. *PLoS Genet,* 5**,** e1000734.

STAJICH, J. E., BLOCK, D., BOULEZ, K., BRENNER, S. E., CHERVITZ, S. A., DAGDIGIAN, C., FUELLEN, G., GILBERT, J. G., KORF, I. & LAPP, H. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res,* 12.

STAMATAKIS, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics,* 22**,** 2688-2690.

STANKE, M. & MORGENSTERN, B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research,* 33**,** W465-W467.

STEBBINS, G. L. 1947. Types of polyploids: their classification and significance. *Adv Genet.,* 1.

STEPHENSON, P., BRYAN, G., KIRBY, J., COLLINS, A., DEVOS, K., BUSSO, C. & GALE, M. 1998. Fifty new microsatellite loci for the wheat genetic map. *Theoretical and Applied Genetics,* 97**,** 946-949.

STRNADOVA V, B. A., GONZALES J, JEGEKLA S, CHAPMAN J, GILBERT JR, ET AL. 2014. Efficient and accurate clustering for large-scale genetic mapping.

SUN, C., HU, Z., ZHENG, T., LU, K., ZHAO, Y., WANG, W., SHI, J., WANG, C., LU, J., ZHANG, D., LI, Z. & WEI, C. 2017. RPAN: rice pan-genome browser for ~3000 rice genomes. *Nucleic Acids Research,* 45**,** 597-605.

SUZUKI, R. & SHIMODAIRA, H. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics,* 22**,** 1540-1542.

SWANSON-WAGNER, R. A., EICHTEN, S. R., KUMARI, S., TIFFIN, P., STEIN, J. C., WARE, D. & SPRINGER, N. M. 2010. Pervasive gene content variation and copy

number variation in maize and its undomesticated progenitor. *Genome Research,* 20**,** 1689-1699.

TAE, H., KARUNASENA, E., BAVARVA, J. H., MCIVER, L. J. & GARNER, H. R. 2014. Large scale comparison of non-human sequences in human sequencing data. *Genomics,* 104**,** 453-458.

TAHILIANI, M., KOH, K. P., SHEN, Y., PASTOR, W. A., BANDUKWALA, H. & BRUDNO, Y. 2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science,* 324.

TALBERT, E.L., BLAKE, N.K., STORLIE, E.W. & LAVIN, M. 1995. Variability in wheat based on low-copy DNA sequence comparisons. Genome 38(5):951-7.

TAN, S., ZHONG, Y., HOU, H., YANG, S. & TIAN, D. 2012. Variation of presence/absence genes among Arabidopsis populations. *BMC Evolutionary Biology,* 12**,** 86.

TANG, N., JIANG, Y., HE, B.-R. & HU, Y.-G. 2009. The Effects of Dwarfing Genes (Rht-B1b, Rht-D1b, and Rht8) with Different Sensitivity to GA3 on the Coleoptile Length and Plant Height of Wheat. *Agricultural Sciences in China,* 8**,** 1028-1038.

TATARINOVA, T. V., CHEKALIN, E., NIKOLSKY, Y., BRUSKIN, S., CHEBOTAROV, D., MCNALLY, K. L. & ALEXANDROV, N. 2016. Nucleotide diversity analysis highlights functionally important genomic regions. *Scientific Reports,* 6**,** 35730.

TATE, J. A., JOSHI, P., SOLTIS, K. A., SOLTIS, P. S. & SOLTIS, D. E. 2009. On the road to diploidization? Homoeolog loss in independently formed populations of the allopolyploid Tragopogon miscellus (Asteraceae). *BMC Plant Biology,* 9**,** 80.

TAUTZ, D. & RENZ, M. 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Research,* 12**,** 4127-4138.

TAUTZ, D. & SCHLÖTTERER, C. 1994. Simple sequences. *Current Opinion in Genetics & Development,* 4**,** 832-837.

TESTER, M. & LANGRIDGE, P. 2010. Breeding Technologies to Increase Crop Production in a Changing World. *Science,* 327**,** 818-822.

TETTELIN, H., MASIGNANI, V., CIESLEWICZ, M. J., DONATI, C., MEDINI, D., WARD, N. L., ANGIUOLI, S. V., CRABTREE, J., JONES, A. L., DURKIN, A. S., DEBOY, R. T., DAVIDSEN, T. M., MORA, M., SCARSELLI, M., MARGARIT Y ROS, I., PETERSON, J. D., HAUSER, C. R., SUNDARAM, J. P., NELSON, W. C., MADUPU, R., BRINKAC, L. M., DODSON, R. J., ROSOVITZ, M. J., SULLIVAN, S. A., DAUGHERTY, S. C., HAFT, D. H., SELENGUT, J., GWINN, M. L., ZHOU, L., ZAFAR, N., KHOURI, H., RADUNE, D., DIMITROV, G., WATKINS, K., O'CONNOR, K. J. B., SMITH, S., UTTERBACK, T. R., WHITE, O., RUBENS, C. E., GRANDI, G.,

MADOFF, L. C., KASPER, D. L., TELFORD, J. L., WESSELS, M. R., RAPPUOLI, R. & FRASER, C. M. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pan-genome”. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 13950-13955.

TETTELIN, H., RILEY, D., CATTUTO, C. & MEDINI, D. 2008. Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology,* 11**,** 472-477.

THE GENOMES PROJECT, C. 2015. A global reference for human genetic variation. *Nature,* 526**,** 68-74.

THE 3,000 GENOMES PROJECT. 2014. The 3,000 rice genomes project. Gigascience. 3:7. doi: 10.1186/2047-217X-3-7.

TILMAN, D., BALZER, C., HILL, J. & BEFORT, B. L. 2011. Global food demand and the sustainable intensification of agriculture. *Proc Natl Acad Sci U S A.,* 108.

TISCHLER, G. & MYERS, E. W. 2017. Non Hybrid Long Read Consensus Using Local De Bruijn Graph Assembly. *bioRxiv*.

TORADA, A., KOIKE, M., MOCHIDA, K. & OGIHARA, Y. 2006. SSR-based linkage map with new markers using an intraspecific population of common wheat. *Theoretical and Applied Genetics,* 112**,** 1042-1051.

TRAPNELL, C., PACHTER, L. & SALZBERG, S. L. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics,* 25.

TRAPNELL, C., ROBERTS, A., GOFF, L., PERTEA, G., KIM, D., KELLEY, D. R., PIMENTEL, H., SALZBERG, S. L., RINN, J. L. & PACHTER, L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protocols,* 7**,** 562-578.

TRICK, M., ADAMSKI, N. M., MUGFORD, S. G., JIANG, C. C., FEBRER, M. & UAUY, C. 2012. Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat. *BMC Plant Biol,* 12.

TSUNEWAKI, K. 2009. Plasmon analysis in the Triticum-Aegilops complex. *Breeding Science,* 59**,** 455-470.

UCHIMURA, Y., WYSS, M., BRUGIROUX, S., LIMENITAKIS, J. P., STECHER, B., MCCOY, K. D. & MACPHERSON, A. J. 2016. Complete Genome Sequences of 12 Species of Stable Defined Moderately Diverse Mouse Microbiota 2. *Genome Announcements,* 4.

VALOUEV, A., ICHIKAWA, J., TONTHAT, T., STUART, J., RANADE, S., PECKHAM, H., ZENG, K., MALEK, J. A., COSTA, G., MCKERNAN, K., SIDOW, A., FIRE, A. &

JOHNSON, S. M. 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res,* 18**,** 1051-63.

VAN BELKUM, A., SCHERER, S., VAN ALPHEN, L. & VERBRUGH, H. 1998. Short-Sequence DNA Repeats in Prokaryotic Genomes. *Microbiology and Molecular Biology Reviews,* 62**,** 275-293.

VAN HEESCH, S., KLOOSTERMAN, W. P., LANSU, N., RUZIUS, F.-P., LEVANDOWSKY, E., LEE, C. C., ZHOU, S., GOLDSTEIN, S., SCHWARTZ, D. C., HARKINS, T. T., GURYEV, V. & CUPPEN, E. 2013. Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing. *BMC Genomics,* 14**,** 257.

VASER, R., SOVIĆ, I., NAGARAJAN, N. & ŠIKIĆ, M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research,* 27**,** 737-746.

VERNIKOS, G., MEDINI, D., RILEY, D. R. & TETTELIN, H. 2015. Ten years of pan-genome analyses. *Current Opinion in Microbiology,* 23**,** 148-154.

VILLA, T. C. C., MAXTED, N., SCHOLTEN, M. & FORD-LLOYD, B. 2005. Defining and identifying crop landraces. *Plant genetic resources: characterization and utilization,* 3**,** 373-384.

VISENDI, P., BATLEY, J. & EDWARDS, D. 2013. Next generation characterisation of cereal genomes for marker discovery. *Biology,* 2**,** 1357-1377.

VOS, P., HOGERS, R., BLEEKER, M., REIJANS, M., LEE, T. V. D., HORNES, M., FRITERS, A., POT, J., PALEMAN, J., KUIPER, M. & ZABEAU, M. 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research,* 23**,** 4407-4414.

VYAHHI, N., PYSHKIN, A., PHAM, S. & PEVZNER, P. A. 2012. From de Bruijn Graphs to Rectangle Graphs for Genome Assembly. *In:* RAPHAEL, B. & TANG, J. (eds.) *Algorithms in Bioinformatics: 12th International Workshop, WABI 2012, Ljubljana, Slovenia, September 10-12, 2012. Proceedings.* Berlin, Heidelberg: Springer Berlin Heidelberg.

WANG, F., JIANG, L., CHEN, Y., HAELTERMAN, N. A., BELLEN, H. J. & CHEN, R. 2015. FlyVar: a database for genetic variation in Drosophila melanogaster. *Database: The Journal of Biological Databases and Curation,* 2015**,** bav079.

WANG, J., LUO, M.-C., CHEN, Z., YOU, F. M., WEI, Y., ZHENG, Y. & DVORAK, J. 2013. Aegilops tauschii single nucleotide polymorphisms shed light on the origins of wheat D-genome genetic diversity and pinpoint the geographic origin of hexaploid wheat. *New Phytologist,* 198**,** 925-937.

WANG, S., WONG, D., FORREST, K., ALLEN, A., CHAO, S., HUANG, B. E., MACCAFERRI, M., SALVI, S., MILNER, S. G., CATTIVELLI, L., MASTRANGELO, A. M., WHAN, A., STEPHEN, S., BARKER, G., WIESEKE, R., PLIESKE, J., INTERNATIONAL WHEAT GENOME SEQUENCING, C., LILLEMO, M., MATHER, D., APPELS, R., DOLFERUS, R., BROWN-GUEDIRA, G., KOROL, A., AKHUNOVA, A. R., FEUILLET, C., SALSE, J., MORGANTE, M., POZNIAK, C., LUO, M.-C., DVORAK, J., MORELL, M., DUBCOVSKY, J., GANAL, M., TUBEROSA, R., LAWLEY, C., MIKOULITCH, I., CAVANAGH, C., EDWARDS, K. J., HAYDEN, M. & AKHUNOV, E. 2014. Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnology Journal,* 12**,** 787-796.

WANG, Y., WANG, X. & PATERSON, A. H. 2012. Genome and gene duplications and gene expression divergence: a view from plants. *Annals of the New York Academy of Sciences,* 1256**,** 1-14.

WANJUGI, H., COLEMAN-DERR, D., HUO, N., KIANIAN, S. F., LUO, M.-C., WU, J., ANDERSON, O. & GU, Y. Q. 2009. Rapid development of PCR-based genome-specific repetitive DNA junction markers in wheat. *Genome,* 52**,** 576-587.

WARBURTON, M. L., CROSSA, J., FRANCO, J., KAZI, M., TRETHOWAN, R., RAJARAM, S., PFEIFFER, W., ZHANG, P., DREISIGACKER, S. & GINKEL, M. V. 2006. Bringing wild relatives back into the family: recovering genetic diversity in CIMMYT improved wheat germplasm. *Euphytica,* 149**,** 289-301.

WEBER, J. L. & WONG, C. 1993. Mutation of human short tandem repeats. *Human Molecular Genetics,* 2**,** 1123-1128.

WEISENFELD, N. I., YIN, S., SHARPE, T., LAU, B., HEGARTY, R., HOLMES, L., SOGOLOFF, B., TABBAA, D., WILLIAMS, L., RUSS, C., NUSBAUM, C., LANDER, E. S., MACCALLUM, I. & JAFFE, D. B. 2014. Comprehensive variation discovery in single human genomes. *Nat Genet,* 46**,** 1350-1355.

WENDEL, J. F. 2000. Genome evolution in polyploids. *Plant Mol Biol.,* 42.

WENDEL, J. F. & DOYLE, J. J. 2005. Polyploidy and evolution in plants. *In:* HENRY, R. J. (ed.) *Plant diversity and evolution.* Wallingford, UK: CABI Publishing.

WENDEL, J. F., JACKSON, S. A., MEYERS, B. C. & WING, R. A. 2016. Evolution of plant genome architecture. *Genome Biology,* 17.

WILLIAMS, R. C. 1989. Restriction fragment length polymorphism (RFLP). *American Journal of Physical Anthropology,* 32**,** 159-184.

WINFIELD, M. O., ALLEN, A. M., BURRIDGE, A. J., BARKER, G. L. A., BENBOW, H. R., WILKINSON, P. A., COGHILL, J., WATERFALL, C., DAVASSI, A., SCOPES, G., PIRANI, A., WEBSTER, T., BREW, F., BLOOR, C., KING, J., WEST, C., GRIFFITHS, S., KING, I., BENTLEY, A. R. & EDWARDS, K. J. 2015. High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnology Journal,* 14**,** 1195-1206.

WINFIELD, M. O., WILKINSON, P. A., ALLEN, A. M., BARKER, G. L. A., COGHILL, J. A. & BURRIDGE, A. 2012. Targeted re-sequencing of the allohexaploid wheat exome. *Plant Biotechnol J,* 10.

WOODHOUSE, M. R., CHENG, F., PIRES, J. C., LISCH, D., FREELING, M. & WANG, X. 2014. Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proc Natl Acad Sci U S A.,* 111.

WU, Q.-H., CHEN, Y.-X., ZHOU, S.-H., FU, L., CHEN, J.-J., XIAO, Y., ZHANG, D., OUYANG, S.-H., ZHAO, X.-J., CUI, Y., ZHANG, D.-Y., LIANG, Y., WANG, Z.-Z., XIE, J.-Z., QIN, J.-X., WANG, G.-X., LI, D.-L., HUANG, Y.-L., YU, M.-H., LU, P., WANG, L.-L., WANG, L., WANG, H., DANG, C., LI, J., ZHANG, Y., PENG, H.-R., YUAN, C.-G., YOU, M.-S., SUN, Q.-X., WANG, J.-R., WANG, L.-X., LUO, M.-C., HAN, J. & LIU, Z.-Y. 2015. High-Density Genetic Linkage Map Construction and QTL Mapping of Grain Shape and Size in the Wheat Population Yanda1817 × Beinong6. *PLOS ONE,* 10**,** e0118144.

WU, Y., BHAT, P. R., CLOSE, T. J. & LONARDI, S. 2008. Efficient and Accurate Construction of Genetic Linkage Maps from the Minimum Spanning Tree of a Graph. *PLoS Genetics,* 4**,** e1000212.

XU, X., LIU, X., GE, S., JENSEN, J. D., HU, F., LI, X., DONG, Y., GUTENKUNST, R. N., FANG, L., HUANG, L., LI, J., HE, W., ZHANG, G., ZHENG, X., ZHANG, F., LI, Y., YU, C., KRISTIANSEN, K., ZHANG, X., WANG, J., WRIGHT, M., MCCOUCH, S., NIELSEN, R., WANG, J. & WANG, W. 2012. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotech,* 30**,** 105-111.

XU, Y. & CROUCH, J. H. 2008. Marker-Assisted Selection in Plant Breeding: From Publications to Practice All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for

reprinting the material contained herein has been obtained by the publisher. *Crop Science,* 48**,** 391-407.

XU, Y., ZHONG, L., WU, X., FANG, X. & WANG, J. 2009. Rapid alterations of gene expression and cytosine methylation in newly synthesized Brassica napus allopolyploids. *Planta,* 229**,** 471-83.

XUE, S., ZHANG, Z., LIN, F., KONG, Z., CAO, Y., LI, C., YI, H., MEI, M., ZHU, H., WU, J., XU, H., ZHAO, D., TIAN, D., ZHANG, C. & MA, Z. 2008. A high-density intervarietal map of the wheat genome enriched with markers derived from expressed sequence tags. *Theoretical and Applied Genetics,* 117**,** 181-189.

YANDELL, M. & ENCE, D. 2012. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet,* 13.

YAO, W., LI, G., ZHAO, H., WANG, G., LIAN, X. & XIE, W. 2015. Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biology,* 16**,** 1-20.

YOSHIHARA, K., TAJIMA, A., ADACHI, S., QUAN, J., SEKINE, M., KASE, H., YAHATA, T., INOUE, I. & TANAKA, K. 2011. Germline copy number variations in BRCA1-associated ovarian cancer patients. *Genes, Chromosomes and Cancer,* 50**,** 167-177.

YU, J.-K., LA ROTA, M., KANTETY, R. V. & SORRELLS, M. E. 2004. EST derived SSR markers for comparative mapping in wheat and rice. *Molecular Genetics and Genomics,* 271**,** 742-751.

YU, P., WANG, C., XU, Q., FENG, Y., YUAN, X., YU, H., WANG, Y., TANG, S. & WEI, X. 2011. Detection of copy number variations in rice using array-based comparative genomic hybridization. *BMC Genomics,* 12**,** 372.

ZEGEYE, H., RASHEED, A., MAKDIS, F., BADEBO, A. & OGBONNAYA, F. C. 2014. Genome-wide association mapping for seedling and adult plant resistance to stripe rust in synthetic hexaploid wheat. *PLoS One,* 9.

ZERBINO, D. R. & BIRNEY, E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research,* 18**,** 821-829.

ZERBINO, D. R., MCEWEN, G. K., MARGULIES, E. H. & BIRNEY, E. 2009. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PLoS One,* 4**,** e8407.

ZEVEN, A. C. 1998. Landraces: A review of definitions and classifications. *Euphytica,* 104**,** 127-139.

ZHANG, L.-M., LUO, H., LIU, Z.-Q., ZHAO, Y., LUO, J.-C., HAO, D.-Y. & JING, H.-C. 2014. Genome-wide patterns of large-size presence/absence variants in sorghum. *Journal of Integrative Plant Biology,* 56**,** 24-37.

ZHAO, Y., JIA, X., YANG, J., LING, Y., ZHANG, Z., YU, J., WU, J. & XIAO, J. 2014. PanGP: A tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics,* 30**,** 1297-1299.

ZHAO, Y., LI, Z., LIU, G., JIANG, Y., MAURER, H. P., WÜRSCHUM, T., MOCK, H.-P., MATROS, A., EBMEYER, E., SCHACHSCHNEIDER, R., KAZMAN, E., SCHACHT, J., GOWDA, M., LONGIN, C. F. H. & REIF, J. C. 2015. Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. *Proceedings of the National Academy of Sciences,* 112**,** 15624-15629.

ZHENG, X., LEVINE, D., SHEN, J., GOGARTEN, S. M., LAURIE, C. & WEIR, B. S. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics,* 28**,** 3326-3328.

ZIETKIEWICZ, E., RAFALSKI, A. & LABUDA, D. 1994. Genome Fingerprinting by Simple Sequence Repeat (SSR)-Anchored Polymerase Chain Reaction Amplification. *Genomics,* 20**,** 176-183.

ZIMIN, A., STEVENS, K. A., CREPEAU, M. W., HOLTZ-MORRIS, A., KORIABINE, M., MARÇAIS, G., PUIU, D., ROBERTS, M., WEGRZYN, J. L., DE JONG, P. J., NEALE, D. B., SALZBERG, S. L., YORKE, J. A. & LANGLEY, C. H. 2014. Sequencing and Assembly of the 22-Gb Loblolly Pine Genome. *Genetics,* 196**,** 875-890.

ZIMIN, A. V., PUIU, D., LUO, M.-C., ZHU, T., KOREN, S., MARÇAIS, G., YORKE, J. A., DVOŘÁK, J. & SALZBERG, S. L. 2017a. 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**,** 787-792.

ZIMIN, A. V., STEVENS, K. A., CREPEAU, M. W., PUIU, D., WEGRZYN, J. L., YORKE, J. A., LANGLEY, C. H., NEALE, D. B. & SALZBERG, S. L. 2017b. An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *GigaScience,* 6**,** 1-4.

ŻMIEŃKO, A., SAMELAK, A., KOZŁOWSKI, P. & FIGLEROWICZ, M. 2014. Copy number polymorphism in plant genomes. *Theoretical and Applied Genetics,* 127**,** 1-18.

7 Appendix I

7.1 Raw data

Flow sorted chromosome arms raw data was downloaded from the Sequence Read Archive (SRA) at NCBI:

Experiment Accession	Experiment Title	Study Accession	Sample Accession	Total Size, Mb	Total Bases
ERX391147	3AS EXP 001	ERP003210	ERS400219	8396	12,997,603,200
ERX391146	3AL EXP 001	ERP003210	ERS400218	9256	14,170,344,800
ERX311327	3DS EXP 002	ERP003210	ERS250078	15821	27,238,800,900
ERX250537	6AS EXP 002	ERP003210	ERS250092	16316	25,361,835,200
ERX311326	3DL EXP 002	ERP003210	ERS250077	9527	14,469,137,520
ERX311325	2AS EXP 002	ERP003210	ERS345887	50375	81,250,688,600
ERX311324	2AS EXP 001	ERP003210	ERS345887	5727	8,923,582,720
ERX311323	2AL EXP 002	ERP003210	ERS345886	13894	24,486,972,720
ERX311322	2AL EXP 001	ERP003210	ERS345886	16972	24,567,068,800
ERX250539	6DS EXP 001	ERP003210	ERS250094	14613	21,914,534,448
ERX250538	6DL EXP 001	ERP003210	ERS250093	19321	29,436,580,224
ERX250536	6AS EXP 001	ERP003210	ERS250092	6476	10,327,628,160
ERX250535	6AL EXP 002	ERP003210	ERS250091	19108	31,656,656,000
ERX250534	6AL EXP 001	ERP003210	ERS250091	1400	2,344,984,640

ERX250533	5DS EXP 001	ERP003210	ERS250090	36290	55,346,149,080
ERX250532	5DL EXP 002	ERP003210	ERS250089	60029	92,436,506,000
ERX250531	5DL EXP 001	ERP003210	ERS250089	7952	12,004,334,400
ERX250528	5AS EXP 002	ERP003210	ERS250086	4035	6,438,617,200
ERX250529	5BL EXP 001	ERP003210	ERS250087	39024	60,583,601,728
ERX250530	5BS EXP 001	ERP003210	ERS250088	25510	42,200,649,200
ERX250527	5AS EXP 001	ERP003210	ERS250086	8734	13,248,494,400
ERX250526	5AL EXP 002	ERP003210	ERS250085	2567	3,642,008,000
ERX250525	5AL EXP 001	ERP003210	ERS250085	15432	22,080,901,500
ERX250524	4DS EXP 001	ERP003210	ERS250084	28459	40,038,268,600
ERX250523	4DL EXP 001	ERP003210	ERS250083	45035	64,213,999,600
ERX250522	4BS EXP 001	ERP003210	ERS250082	38616	60,746,860,864
ERX250521	4BL EXP 001	ERP003210	ERS250081	15630	23,828,691,000
ERX250520	4AS EXP 002	ERP003210	ERS250080	37641	56,995,583,600
ERX250519	4AS EXP 001	ERP003210	ERS250080	11332	19,474,296,600
ERX250518	4AL EXP 002	ERP003210	ERS250079	19919	31,865,134,200
ERX250517	4AL EXP 001	ERP003210	ERS250079	17861	30,504,178,200
ERX250516	3DS EXP 001	ERP003210	ERS250078	45720	63,719,600,534
ERX250515	3DL EXP 001	ERP003210	ERS250077	50473	74,533,213,610
ERX250514	2DS EXP 002	ERP003210	ERS250076	5510	853,268,9920
ERX250513	2DS EXP 001	ERP003210	ERS250076	23266	37,980,363,200

ERX250512	2DL EXP 002	ERP003210	ERS250075	13843	22,566,127,800
ERX250511	2DL EXP 001	ERP003210	ERS250075	4293	6,841,044,160
ERX250510	2BS EXP 001	ERP003210	ERS250074	30834	49,766,848,804
ERX250509	2BL EXP 001	ERP003210	ERS250073	52021	71,807,340,800
ERX250508	1DS EXP 002	ERP003210	ERS250072	5662	9,868,160,880
ERX250507	1DS EXP 001	ERP003210	ERS250072	17329	25,429,420,000
ERX250506	1DL EXP 001	ERP003210	ERS250071	19745	27,903,660,600
ERX250505	1BS EXP 002	ERP003210	ERS250070	4433	7,049,893,760
ERX250504	1BS EXP 001	ERP003210	ERS250070	22596	37,998,068,200
ERX250503	1BL EXP 002	ERP003210	ERS250069	42172	68,238,011,600
ERX250502	1BL EXP 001	ERP003210	ERS250069	4909	7,713,226,240
ERX250500	1AL EXP 002	ERP003210	ERS250067	52302	86,144,072,800
ERX250501	1AS EXP 001	ERP003210	ERS250068	23407	39,289,281,200
ERX250499	1AL EXP 001	ERP003210	ERS250067	5075	7,942,652,160
SRX232100	7AS - 100bp	SRP018533	SRS392985	5977	11,380,741,000
SRX232096	7AL - 100bp	SRP018533	SRS392983	10162	18,738,325,600
SRX232065	7BL - 100bp	SRP018533	SRS392969	7841	15,104,269,600
SRX232061	7DL - 100bp	SRP018533	SRS392964	15205	26,526,560,800
SRX036849	7BS - 100bp	SRP005092	SRS150933	8593	16,134,219,800
SRX040744	7DS - 100bp	SRP004476	SRS121460	11674	20,483,498,200

Australian modern wheat cultivars:

URL: https://downloads.bioplatforms.com/wheat_cultivars/samples

Download date: January 2016

RNA-seq data:

FTP: https://urgi.versailles.inra.fr/files/RNASeqWheat/

Download date: December 2014

454 whole genome shotgun data:

URL: https://www.ebi.ac.uk/ebisearch/search.ebi?db=allebi&query=ERP000319&submit1=1&requestFrom=ebi_index

Download date: January 2016

8 Appendix II

Contamination of libraries from cultivar BX-1. Three libraries produced from cultivar BX-1 that had mapping efficiencies below 35% were further analysed. One thousand reads were randomly selected from each of the libraries and aligned to the nucleotide database in NCBI. All blast results were merged in a single table and sorted by the genus of the aligned record. **Table 8-1.** Top ten most frequent blast hits from the 3 BX-1libraries**Table 8-1** shows the top ten most frequent genera and the count of reads that were aligned to them.

Table 8-1. Top ten most frequent blast hits from the 3 BX-1libraries with mapping efficiency below 35%. The count corresponds to the number of reads whose best blast hit was the corresponding genus.

Genus	Count
Mezorhizobium	227
Penicillium	215
Acinetobacter	169
Triticum	134
Actinomyces	108
Chlorella	93
Serratia	91
Bacillus	82
Setaria	64
Trifolium	61

9 Appendix III

Manuscript which presents the T. aestiivum pangenome that was published in the Plant Journal

Title

The pangenome of hexaploid bread wheat

Authors

Juan D. Montenegro[1*], Agnieszka A. Golicz[1,2*], Philipp E. Bayer[2*], Bhavna Hurgobin[1,2], HueyTyng Lee[1,2], Chon-Kit Kenneth Chan[2], Paul Visendi[1], Kaitao Lai[3] Jaroslav Doležel[4], Jacqueline Batley[1,2,5], David Edwards[1,2,5‡].

1, School of Agriculture and Food Sciences, University of Queensland, Brisbane, Australia

2, School of Plant Biology, University of Western Australia, WA, 6009, Australia

3, Hawkesbury Institute for the Environment, Western Sydney University, NSW, 2751, Australia

4, Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, Šlechtitelů 31, CZ-783 71 Olomouc, Czech Republic

5, Institute of Agriculture, University of Western Australia, WA, 6009, Australia

*These authors contributed equally to the manuscript.

‡ Address correspondence to Dave.Edwards@uwa.edu.au

Summary

There is an increasing understanding that gene presence absence variation plays an important role in the heritability of agronomic traits, however there have been relatively few studies on gene presence absence variation in crop species. Hexaploid wheat is one of the most important food crops in the world and intensive breeding has reduced the genetic diversity of elite cultivars. Major efforts have produced draft genome assemblies for the cultivar Chinese Spring, but it is unknown how well this represents the genome diversity found in current modern elite cultivars. In this study we build an improved reference for Chinese Spring and explore gene diversity across 18 wheat cultivars. We predict a pangenome size of 140,500 +/- 102 genes, a core genome of 81,070 +/- 1,631 genes, and an average of 128,656 genes in each cultivar. Functional annotation of the variable gene set suggests that it is enriched for genes that may be associated with important agronomic traits. In addition to gene presence variation, more than 36 million intervarietal SNPs were identified across the pangenome. This study of the wheat pangenome provides insight into elite wheat genome diversity as a basis for genomics based improvement of this important crop. A wheat pangenome Gbrowse is available at http://appliedbioinformatics.com.au/cgi-bin/gb2/gbrowse/WheatPan/, and data is available for download from http://wheatgenome.info/wheat_genome_databases.php.

Significance statement

We have assembled a wheat pangenome, identified and functionally annotated the core and variable genes and constructed the most comprehensive SNP database available for wheat. These resources can be applied for the wheat genomics and breeding communities as understanding the presence and diversity of genes is essential for their association with agronomic traits.

Introduction

Wheat is one of the most important food crops in the world, and its continued improvement is essential to maintain food security in the face of a growing human population and disturbance of agricultural production due to climate change (Abberton et al., 2015, Batley and Edwards, 2016). Wheat was domesticated 8,000 – 10,000 years ago (Dubcovsky and Dvorak, 2007), and today bread wheat (*Triticum aestivum*) provides roughly a fifth of the world's food. Genome analysis in bread wheat is a challenge because of its large (17 Gbp) genome, consisting of between 80% and 90% repetitive sequence (Wanjugi et al., 2009, Šafář et al., 2010a). Bread wheat is also hexaploid, being derived from a combination of three diploid donor species which are proposed to have diverged from an ancestral diploid species between 2.5 and 6 MYA (Huang et al., 2002a, Chantret et al., 2005). There have been several efforts to sequence the genome of hexaploid bread wheat. The *de novo* assembly of sequence data from flow-sorted chromosome arms was initially performed for 7DS, demonstrating that it was possible to assemble all known 7DS genes (Berkman et al., 2011a). The same approach delimited a translocation between chromosome arms 7BS and 4AL (Berkman et al., 2012b), with a subsequent comparison of all group 7 chromosomes, highlighting genomic changes during the early evolution and domestication of this important crop (Berkman et al., 2013b). The application of a similar approach towards all chromosome arms with the exception of 3B, (IWGSC, 2014) together with a whole genome assembly of Roche 454 sequence data (Brenchley et al., 2012) provided the first draft genome assemblies for wheat cultivar Chinese Spring. Two additional cultivars, OpataM85 and W7984 have undergone whole genome shotgun sequencing using Illumina data, and although gene presence comparisons were performed using cDNA mapping, these assemblies were not annotated (Chapman et al., 2015), limiting their use for pangenome analysis. With the exception of Chapman et al. (2015), each of these studies have focussed on the cultivar Chinese Spring.

Crop breeding increasingly benefits from the application of molecular tools such as marker assisted selection (MAS) and more recently, genomic selection (GS), and the increasing availability of genomic information supports these advanced breeding tools (Poland et al., 2012b, Simeão Resende et al., 2014, Sallam et al., 2015, Cros et al., 2015, Crossa et al., 2014). Modern molecular breeding tools apply single

nucleotide polymorphism (SNP) molecular genetic markers, and numerous studies have discovered and validated large numbers of SNP markers across the wheat genome (Winfield et al., 2015, Wang et al., 2014, Lai et al., 2015a, Lai et al., 2012c). SNPs have been used to find genes undergoing selective sweeps and population bottlenecks (Cavanagh et al., 2013), and have also been used to map low diversity regions which could have been targets of selection (Lai et al., 2015a).

The decreasing cost of DNA sequencing has accelerated genomics research in recent years (Visendi et al., 2013, Edwards et al., 2013b). Most sequencing projects focus on reference genome assembly and the discovery of SNPs, however, the importance of structural variants is becoming increasingly acknowledged (Saxena et al., 2014, Wendel et al., 2016, Jordan et al., 2015). Studies in several plant species reveal the existence of extensive structural variation (Gordon et al., 2014, Xu et al., 2012, Springer et al., 2009b, Zhang et al., 2014, Hardigan et al., 2016, Li et al., 2014d). One form of structural variation, the presence or absence of genes or genomic regions between individuals of the same species, is being increasingly acknowledged as an important form of variation in plants, and the sum of core and variable regions of the genome for a species is known as the pangenome.

Several approaches to pangenome assembly and analysis have been developed (Golicz et al., 2015a). The traditional approach, first applied in bacteria involves whole genome assembly of all genotypes, followed by individual annotation and comparison of the gene content (Tettelin et al., 2005, Schatz et al., 2014, Li et al., 2014d). An alternative is a read mapping and assembly approach, where sequence reads are first mapped to an existing reference, and the unmapped reads are then assembled (Golicz et al., 2015a, Yao et al., 2015, Golicz et al., 2016b).

The first step towards the production of a pangenome for a crop species is the production of a suitable reference assembly, followed by the expansion of this reference with additional sequences from other varieties which are not present in the reference. In this study we have reassembled a draft Chinese Spring wheat genome reference and used this as the basis for a pangenome study, identifying core and variable genes across 18 cultivars (Edwards et al., 2012). We have also identified 36.4 million SNPs between these 18 cultivars. The Chinese Spring reference is different in gene content than the 18 cultivars, suggesting that this pangenome and the associated SNPs may provide a better reference for wheat crop improvement than the current Chinese Spring references.

Results and Discussion

Wheat (cv. Chinese Spring) genome assembly

An assessment of the sequence duplication in the IWGSC draft Chinese Spring assembly (IWGSC, 2014) showed that 663 Mb (7%) of the assembly consisted of exact duplications greater than 1 Kb, with more than 40% of chromosome arms 4AS and 4AL being duplicated (Figure S1). Following reassembly, producing 10.7 Gb of new reference, these duplications were reduced to of 0.4 Mb (0.004%). The high frequency of duplicated regions in the IWGSC assembly (Figure S1) may be an artefact of using the parallelised *de bruijn* graph assembler ABySS (Simpson et al., 2009) as they were not observed in the previous assemblies of group 7 data (Berkman et al., 2011a, Berkman et al., 2012b, Berkman et al., 2013b) which used the non-parallel *de bruijn* graph assembler Velvet (Zerbino and Birney, 2008).

A reassembly of the IWGSC data in this study using Velvet produced a reference with larger assembly size and greatly reduced frequency of duplicated regions (Figure S1) compared to the published draft genome (IWGSC, 2014). CEGMA analysis (Parra et al., 2009) was performed to assess the completeness of the assembly and identified 245 (98.8%) of the 248 core eukaryotic genes compared to 243 genes identified in the IWGSC assembly.

Pangenome assembly

Whole genome sequence reads from 18 wheat cultivars were mapped to the new Chinese Spring assembly, and unmapped reads assembled. The average sequencing depth ranged from 8.4x to 19.9x, except for Chinese Spring which had a coverage that ranged from 60X to 200X for each of the chromosome arms. (Table S5). After removal of contaminant sequences, the newly assembled sequence contained 221,991 scaffolds with a total length of 350 Mb (Table S1) and a total of 21,653 predicted genes. Mapping of Chinese Spring sequence reads to this pangenome demonstrated that this sequence was not present in the Chinese Spring reference and represents a 3.3% increase in the size of the wheat reference genome. A similar approach was used by Yao et al. (2015) with 1,483 rice accessions from the *japonica* and *indica* groups, where they assembled 15.8 Mb and

24.6 Mb of additional sequence for each subspecies respectively, representing an increase of 4% and 6% in genome size. Similarly, local reassembly in *Brachypodium distachyon* identified 19.2 Mb of additional sequence in 7 highly diverse inbred lines, a 5% increase in the size of the reference genome. Golizc et al (2016) characterised the pangenome of *Brassica oleracea* using 9 diverse morphotypes and assembled an additional 99 Mbp of sequence. The relatively small increase in pangenome assembly size we observe reflects the high degree of relatedness of the cultivars sequenced (Lai et al., 2015a). The additional sequence identified in this study is likely to be an underestimate of the total sequence content present in the cultivars as sequences present in only one or two of the cultivars are unlikely to have sufficient coverage to assemble, as IDBA-UD has 81% assembly efficiency for samples with a sequencing depth of 10X (Peng et al., 2012b).

Gene presence/absence discovery

The presence or absence of each gene was predicted for each cultivar based on the mapping of reads from each cultivar to the new pangenome assembly (Table S2). The approach followed the method of Golicz et al. (2016b) which demonstrates a 0.05% error rate using 10x read coverage. Based on Chinese Spring read mapping to the pangenome, none of the additional genes identified in the 18 cultivars were identified as present in Chinese Spring. On average, each cultivar contains 128,656 genes, with 89,795 (64.3%) shared by all 19 cultivars, while 49,952 genes represent the variable genome across these cultivars. Based on gene presence and absence in each of the 18 cultivars we estimate that the pangenome of modern wheat cultivars contains 140,500 +/- 102 genes (Figure 1). This is likely to be an underestimate of the broader wheat pangenome as it is predicted from a relatively narrow set of cultivars, and extending the study to more diverse landraces and wild relatives will provide a more comprehensive measure of the gene content of this important crop species.

Characterisation of Chinese Spring gene content identified 245 genes in Chinese Spring which are absent from the 18 cultivars, while a further 12,150 genes were identified in all 18 cultivars but are not found in Chinese Spring (Table S2). A dendrogram reconstructed using gene presence/absence variation places Chinese Spring in a separate cluster at the base of the tree (Figure 2). This is similar to a previous study using SSR markers where Chinese Spring was placed in the basal

node away from most modern wheat cultivars (Plaschke et al., 1995). Our results can also be explained by the history of Chinese Spring, which despite being a major source of cytogenetic stocks, used in the discovery of the seven homoeologous chromosome groups and in early gene mapping efforts (Sears, 1966, Sharp et al., 1988), and more recently in genome sequencing (IWGSC, 2014), it is not widely used in breeding programs due to its susceptibility to biotic and abiotic stress (Sears and Miller, 1985).

Variable genes were annotated, and functional enrichment analysis suggests that the variable genome is enriched with genes involved in response to environmental stress and defence response (Figure 3; Table S3). Similarly, Yao et al. (2015) found that the variable genome of rice was enriched with genes related to biotic stress defence including NBS LRR genes and genes coding for protein kinases and abiotic stress tolerance (Yao et al., 2015). Analysis of the Brassica oleracea pangenome by (2016b) also found that variable genes were enriched for annotated related to major agronomic traits, including disease resistance.

SNP discovery

Capturing and characterising diversity is essential in the design and execution of breeding programs. We have previously identified more than 4 million SNPs on the group 7 Chinese Spring chromosomes with a validation rate of 95% (Lai et al., 2015a). Using the same method, whole genome shotgun reads from the 18 wheat cultivars were mapped to the pangenome assembly and SNPs were identified using SGSautoSNP (Lorenc et al., 2012), leading to the identification of 36.4 million SNPs. Of these, 2.87 million were identified in scaffolds not present in the Chinese Spring assembly. The SGSautoSNP calls were compared with SNPs from a published Infinium array (Wang et al., 2014). A total of 13,541 Infinium SNPs were identified as being at the same location as the SGSautoSNP calls. Out of these 96.3% were identified as polymorphic. This is similar to the validation rate observed by Lai et al. (2015a) using the same approach. The majority of SNPs were found in intergenic regions, with only 392,142 (1%) SNPs located in coding regions. Of these 225,064 (57.4%) are predicted to be non-synonymous resulting in a potentially different functional protein. These results are comparable to those obtained by Jordan et al (2015) who found that 52.3% of the SNPs were non-synonimous (Jordan et al.,

2015). The dataset represents the most comprehensive SNP resource available for the improvement of elite bread wheat cultivars.

Conclusion

In this study, we constructed and analysed a draft wheat pangenome using a single reference and whole genome sequencing data from 18 cultivars. The pangenome contains 128,656 predicted genes of which 64.3% are identified as core, that is present in all cultivars, while the remainder are variable and display presence/absence variation. Additionally, 12,150 genes are absent in the Chinese Spring reference sequence but present in all the other cultivars analysed. The pangenome sequence is a valuable resource for the wheat genomics and breeding communities as understanding the diversity of genes is essential for their association with agronomic traits. The pangenome can be easily expanded to include additional genes from other diverse wheat cultivars and, along with the SNP dataset derived from it, provide markers that can be used to integrate this resource into current GWAS pipelines.

Experimental procedures

Genome assembly and annotation

Sequence data was downloaded from various repositories as described in (Table S4). Clonal reads were removed using an in-house script (remove_clones.pl). Quality trimming and adapter clipping was performed using TRIMMOMATIC v 0.33 (Bolger et al., 2014), and sequences shorter than 73bp were removed. VELVET v 1.2.10 (Zerbino and Birney, 2008, Seemann, 2012) was used for assembly using a kmer size of 71. RNA-seq reads were aligned to the reference genome using TOPHAT2 v2.1.0.1 (Kim et al., 2013b). Accepted alignments were transformed into hints files with the script bam2hints from the AUGUSTUS package.

REPEATMASKER (Smit et al., 2015) was used to mask repeated regions using RepBase version 20150807 (Jurka et al., 2005) and viridiplantae as species. AUGUSTUS v 2.1.0 (Keller et al., 2011) predicted gene models using the hints produced from the RNA-seq alignments. Gene models were first filtered for size (>=300bp). BEDOPS v 2.4.15 (Neph et al., 2012) was used to identify and remove gene models that were not supported by TOPHAT2 annotation or overlapped repeat-

masked regions. Finally, the protein sequence of the selected models were aligned to TE-related proteins with BLASTP and those with significant alignments (E ≤ 1e-5) were removed from the annotation. The protein sequences of the final gene set were aligned to the proteome of *Triticum uratrtu* to identify and merge split genes.

CEGMA (Parra et al., 2009) was used to assess the completeness of the reference genome prior to annotation with default parameters.

Pangenome assembly and annotation

Reads from the 16 wheat cultivars were mapped to the new Chinese Spring assembly using Bowtie2 v2.2.5, and unmapped reads pooled. The sequencing depth per cultivar is shown in Table S5. TRIMMOMATIC v 0.33 removed adapter and low quality sequence and the reads were assembled using IDBA_UD (Peng et al., 2012b) using standard parameters. The resulting scaffolds were compared with the NCBI non-redundant nucleotide database using BLASTN (E ≤ 1e-5) and the scaffolds with hits outside the seed plants taxonomy group were removed. REPEATMASKER v 4.0.6 masked repetitive elements using ‘viridiplantae’ as the species. Then, TBLASTX (Camacho et al., 2009) was used to align the green plant ESTs from genbank, and genes were predicted using AUGUSTUS v2.1.0., supported by the EST alignments. The reads from W7984, OataM85 and 90 doubled haploid offspring were mapped to the full pangenome assembly and unmapped reads were processed and assembled as described above. Libraries with mapping efficiency below 80% were not included for further analysis.

Gene presence/absence and pangenome prediction

BOWTIE2 v 2.2.5 was used to align the reads with standard parameters and an insert size between 0 and 1000 bp. Gene presence/absence was called as described by Golicz et al (Golicz et al., 2015b). SAMTOOLS was used to calculate the coverage of the annotated genes, and an in house script (pileup2cov.pl) predicted the presence/absence status of each gene based on the following requirements: coverage >2X and exon fraction covered >0.05. PVCLUST (Suzuki and Shimodaira, 2006) was used with the presence/absence binary matrix to estimate the relationship between the cultivars. One thousand resamplings were used for bootstrap calculations.

The program PANGP (Zhao et al., 2014) was used to count the core and total genes present in all possible combinations of the 19 cultivars. The average results gene count from each iteration was plotted and used to model the wheat pangenome expansion using a power law model ($f(x) = Ax^B+C$) (Tettelin et al., 2005) by means of the R nls function. Assuming a closed pangenome, the C parameter was used as an estimator of the total gene content in the pangenome. The same approach was used to estimate the core genome, using the average gene count to fit the model $f(x) = Ae^{Bx}+C$.

SNP discovery

Reads were mapped to the pangenome using BOWTIE2 v2.2.5 (--no-mixed --no-unal -I 0 -X 1000) (Langmead and Salzberg, 2012). Reads with MAPQ < 20 and with low base qualities were removed from the alignments along with their mates. SAM files were further processed and duplicated reads removed with samtools v 1.3.1 (Li et al., 2009b). SGSautoSNP (Lorenc et al., 2012) was used to identify SNPs. SNPs were validated as described in Lai et al (2015). SNPEFF v4.2 (Cingolani et al., 2012) was used to predict the effect of the SNPs on the gene annotations.

SNP validation

The sequence tags from the 90K SNP Infinium array (Wang et al., 2014) were aligned to the reference wheat pangenome using NCBI Blast plus (Camacho et al., 2009). High quality alignments (E-threshold < 1e-10 and >= 99% sequence identity) where used to count the number of common polymorphic SNPs as described in Lai et al (2015).

Acknowledgements

This work was supported by the Australian Research Council (Projects LP140100537 and LP130100925). Support is also acknowledged from the Australian Genome Research Facility (AGRF) and the Queensland Cyber Infrastructure Foundation (QCIF), the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia and resources from the National Computational Infrastructure (NCI), which is supported by the Australian Government.

Figures and Tables

Supplementary Information:

References

1995. Variability in wheat based on low-copy DNA sequence comparisons. *Genome,* 38**,** 951-957.

2010. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature,* 463**,** 763-768.

2011. Genome sequence and analysis of the tuber crop potato. *Nature,* 475**,** 189-195.

2014. The 3000 rice genomes project. *GigaScience,* 3.

ABBERTON, M., BATLEY, J., BENTLEY, A., BRYANT, J., CAI, H., COCKRAM, J., COSTA DE OLIVEIRA, A., CSEKE, L. J., DEMPEWOLF, H., DE PACE, C., EDWARDS, D., GEPTS, P., GREENLAND, A., HALL, A. E., HENRY, R., HORI, K., HOWE, G. T., HUGHES, S., HUMPHREYS, M., LIGHTFOOT, D., MARSHALL, A., MAYES, S., NGUYEN, H. T., OGBONNAYA, F. C., ORTIZ, R., PATERSON, A. H., TUBEROSA, R., VALLIYODAN, B., VARSHNEY, R. K. & YANO, M. 2015. Global agricultural intensification during climate change: a role for genomics. *Plant Biotechnology Journal,* 14**,** 1095-1098.

ABDEL-GHANY, S. E., HAMILTON, M., JACOBI, J. L., NGAM, P., DEVITT, N., SCHILKEY, F., BEN-HUR, A. & REDDY, A. S. N. 2016. A survey of the sorghum transcriptome using single-molecule long reads. 7**,** 11706.

ADAMS, K. L. & WENDEL, J. F. 2005a. Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology,* 8**,** 135-141.

ADAMS, K. L. & WENDEL, J. F. 2005b. Polyploidy and genome evolution in plants. *Curr Opin Plant Biol,* 8.

ADRIAN ALEXA, J. R. 2006. topGO: Enrichment Analysis for Gene Ontology.

AIR, G., SANGER, F., BARRELL, B., BROWN, N., COULSON, A., FIDDES, J., HUTCHISON, C., SLOCOMBE, P. & SMITH, M. NUCLEOTIDE-SEQUENCE OF DNA OF BACTERIOPHAGE-PHIX174. PROCEEDINGS OF THE AUSTRALIAN BIOCHEMICAL SOCIETY, 1977. PROC AUST BIOCHEMICAL SOC MONASH UNIV DEPT BIOCHEMISTRY, CLAYTON VICTORIA 3168, AUSTRALIA, 60-60.

AKBARI, M., WENZL, P., CAIG, V., CARLING, J., XIA, L. & YANG, S. 2006. Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. *Theor Appl Genet,* 113.

AKHUNOV, E. D., AKHUNOVA, A. R., ANDERSON, O. D., ANDERSON, J. A., BLAKE, N. & CLEGG, M. T. 2010. Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. *BMC Genomics,* 11.

AKHUNOV, E. D., SEHGAL, S., LIANG, H., WANG, S., AKHUNOVA, A. R. & KAUR, G. 2013. Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. *Plant Physiol,* 161.

AL-OKAILY, A. A. 2016. HGA: de novo genome assembly method for bacterial genomes using high coverage short sequencing reads. *BMC Genomics,* 17**,** 193.

ALLEN, A. M., BARKER, G. L. A., BERRY, S. T., COGHILL, J. A., GWILLIAM, R., KIRBY, S., ROBINSON, P., BRENCHLEY, R. C., D'AMORE, R., MCKENZIE, N., WAITE, D., HALL, A., BEVAN, M., HALL, N. & EDWARDS, K. J. 2011. Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (Triticum aestivum L.). *Plant Biotechnology Journal,* 9**,** 1086-1099.

ALLOUIS, S., MOORE, G., BELLEC, A., SHARP, R., RAMPANT, P. F., MORTIMER, K., PATEYRON, S., FOOTE, T. N., GRIFFITHS, S., CABOCHE, M. & CHALHOUB, B. 2003. Construction and characterisation of a hexaploid wheat (Triticum aestivum L.) BAC library from the reference germplasm 'Chinese Spring'. *Cereal Research Communications,* 31**,** 331-338.

ALONSO-BLANCO, C., ANDRADE, J., BECKER, C., BEMM, F., BERGELSON, J., BORGWARDT, K. M., CAO, J., CHAE, E., DEZWAAN, T. M., DING, W., ECKER, J. R., EXPOSITO-ALONSO, M., FARLOW, A., FITZ, J., GAN, X., GRIMM, D. G., HANCOCK, A. M., HENZ, S. R., HOLM, S., HORTON, M., JARSULIC, M., KERSTETTER, R. A., KORTE, A., KORTE, P., LANZ, C., LEE, C.-R., MENG, D., MICHAEL, T. P., MOTT, R., MULIYATI, N. W., NÄGELE, T., NAGLER, M., NIZHYNSKA, V., NORDBORG, M., NOVIKOVA, P. Y., PICÓ, F. X., PLATZER, A., RABANAL, F. A., RODRIGUEZ, A., ROWAN, B. A., SALOMÉ, P. A., SCHMID, K. J., SCHMITZ, R. J., SEREN, Ü., SPERONE, F. G., SUDKAMP, M., SVARDAL, H., TANZER, M. M., TODD, D.,

VOLCHENBOUM, S. L., WANG, C., WANG, G., WANG, X., WECKWERTH, W., WEIGEL, D. & ZHOU, X. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell,* 166**,** 481-491.

ANDERSON, J. A., CHURCHILL, G. A., AUTRIQUE, J. E., TANKSLEY, S. D. & SORRELLS, M. E. 1993. Optimizing parental selection for genetic linkage maps. *Genome,* 36**,** 181-186.

ANDREWS, S. *FastQC A Quality Control tool for High Throughput Sequence Data* [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [Accessed].

ARAI-KICHISE, Y., SHIWA, Y., NAGASAKI, H., EBANA, K., YOSHIKAWA, H., YANO, M. & WAKASA, K. 2011. Discovery of Genome-Wide DNA Polymorphisms in a Landrace Cultivar of Japonica Rice by Whole-Genome Sequencing. *Plant and Cell Physiology,* 52**,** 274-282.

ARCHER, J., BAILLIE, G., WATSON, S. J., KELLAM, P., RAMBAUT, A. & ROBERTSON, D. L. 2012. Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator II. *BMC Bioinformatics,* 13**,** 47.

ASHBY, M., EKHOLM, J. & HICKEY, L. 2017. Sequencing for Structural Variation and Isoform Discovery. *Genetic Engineering & Biotechnology News,* 37**,** 16-17.

ASSENG, S., EWERT, F., MARTRE, P., ROTTER, R. P., LOBELL, D. B., CAMMARANO, D., KIMBALL, B. A., OTTMAN, M. J., WALL, G. W., WHITE, J. W., REYNOLDS, M. P., ALDERMAN, P. D., PRASAD, P. V. V., AGGARWAL, P. K., ANOTHAI, J., BASSO, B., BIERNATH, C., CHALLINOR, A. J., DE SANCTIS, G., DOLTRA, J., FERERES, E., GARCIA-VILA, M., GAYLER, S., HOOGENBOOM, G., HUNT, L. A., IZAURRALDE, R. C., JABLOUN, M., JONES, C. D., KERSEBAUM, K. C., KOEHLER, A. K., MULLER, C., NARESH KUMAR, S., NENDEL, C., O'LEARY, G., OLESEN, J. E., PALOSUO, T., PRIESACK, E., EYSHI REZAEI, E., RUANE, A. C., SEMENOV, M. A., SHCHERBAK, I., STOCKLE, C., STRATONOVITCH, P., STRECK, T., SUPIT, I., TAO, F., THORBURN, P. J., WAHA, K., WANG, E., WALLACH, D., WOLF, J., ZHAO, Z. & ZHU, Y. 2015. Rising temperatures reduce global wheat production. *Nature Clim. Change,* 5**,** 143-147.

BACHLAVA, E., TAYLOR, C. A., TANG, S., BOWERS, J. E., MANDEL, J. R., BURKE, J. M. & KNAPP, S. J. 2012. SNP discovery and development of a high-density genotyping array for sunflower. *PLoS One,* 7.

BADAEVA, E. D., DEDKOVA, O. S., GAY, G., PUKHALSKYI, V. A., ZELENIN, A. V., BERNARD, S. & BERNARD, M. 2007. Chromosomal rearrangements in wheat: their types and distribution. *Genome,* 50**,** 907-926.

BAETS, G., DURME, J., REUMERS, J., MAURER-STROH, S., VANHEE, P. & DOPAZO, J. 2012. SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res,* 40.

BAIER, U., BELLER, T. & OHLEBUSCH, E. 2016. Graphical pan-genome analysis with compressed suffix trees and the Burrows–Wheeler transform. *Bioinformatics,* 32**,** 497-504.

BANKEVICH, A., NURK, S., ANTIPOV, D., GUREVICH, A. A., DVORKIN, M., KULIKOV, A. S., LESIN, V. M., NIKOLENKO, S. I., PHAM, S., PRJIBELSKI, A. D., PYSHKIN, A. V., SIROTKIN, A. V., VYAHHI, N., TESLER, G., ALEKSEYEV, M. A. & PEVZNER, P. A. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology,* 19**,** 455-477.

BARBARA, T., PALMA-SILVA, C., PAGGI, G. M., BERED, F., FAY, M. F. & LEXER, C. 2007. Cross-species transfer of nuclear microsatellite markers: potential and limitations. *Molecular Ecology,* 16**,** 3759-3767.

BARTOS, J., PAUX, E., KOFLER, R., HAVRANKOVA, M., KOPECKY, D., SUCHANKOVA, P., SAFAR, J., SIMKOVA, H., TOWN, C. D., LELLEY, T., FEUILLET, C. & DOLEZEL, J. 2008. A first survey of the rye (Secale cereale) genome composition through BAC end sequencing of the short arm of chromosome 1R. *BMC Plant Biol,* 8**,** 95.

BATLEY, J. & EDWARDS, D. 2007. SNP Applications in Plants. *In:* ORAGUZIE, N., RIKKERINK, E. A., GARDINER, S. & DE SILVA, H. N. (eds.) *Association Mapping in Plants.* Springer New York.

BATLEY, J. & EDWARDS, D. 2016. The application of genomics and bioinformatics to accelerate crop improvement in a changing climate. *Current Opinion in Plant Biology,* 30**,** 78-81.

BEKELE, W. A., WIECKHORST, S., FRIEDT, W. & SNOWDON, R. J. 2013. High-throughput genomics in sorghum: from whole-genome resequencing to a SNP screening array. *Plant Biotechnology Journal,* 11**,** 1112-1125.

BELLER, T. & OHLEBUSCH, E. 2016. A representation of a compressed de Bruijn graph for pan-genome analysis that enables search. *Algorithms for Molecular Biology,* 11**,** 20.

BELÓ, A., BEATTY, M. K., HONDRED, D., FENGLER, K. A., LI, B. & RAFALSKI, A. 2009. Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theoretical and Applied Genetics,* 120**,** 355.

BENNETT, M. D. 1972. Nuclear DNA Content and Minimum Generation Time in Herbaceous Plants. *Proceedings of the Royal Society of London. Series B. Biological Sciences,* 181**,** 109-135.

BENTLEY, S. 2009. Sequencing the species pan-genome. *Nat Rev Micro,* 7**,** 258-259.

BERKMAN, P., SKARSHEWSKI, A., MANOLI, S., LORENC, M., STILLER, J., SMITS, L., LAI, K., CAMPBELL, E., KUBALÁKOVÁ, M., ŠIMKOVÁ, H., BATLEY, J., DOLEŽEL, J., HERNANDEZ, P. & EDWARDS, D. 2012a. Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theoretical and Applied Genetics,* 124**,** 423-432.

BERKMAN, P. J., MANOLI, S., MCKENZIE, M., KUBALÁKOVÁ, M., ŠIMKOVÁ, H., BATLEY, J., FLEURY, D., DOLEŽEL, J., EDWARDS, D., SKARSHEWSKI, A., LORENC, M. T., LAI, K., DURAN, C., LING, E. Y. S., STILLER, J., SMITS, L. & IMELFORT, M. 2011a. Sequencing and assembly of low copy and genic regions of isolated Triticum aestivum chromosome arm 7DS. *Plant Biotechnology Journal,* 9**,** 768-775.

BERKMAN, P. J., SKARSHEWSKI, A., LORENC, M. T., LAI, K., DURAN, C., LING, E. Y., STILLER, J., SMITS, L., IMELFORT, M., MANOLI, S., MCKENZIE, M., KUBALAKOVA, M., SIMKOVA, H., BATLEY, J., FLEURY, D., DOLEZEL, J. & EDWARDS, D. 2011b. Sequencing and assembly of low copy and genic regions of isolated Triticum aestivum chromosome arm 7DS. *Plant Biotechnol J,* 9**,** 768-75.

BERKMAN, P. J., SKARSHEWSKI, A., MANOLI, S., LORENC, M. T., STILLER, J., LARS, SMITS, L., LAI, K., CAMPBELL, E., KUBALAKOVA, M., SIMKOVA, H., BATLEY, J., DOLEZEL, J., HERNANDEZ, P. & EDWARDS, D. 2012b. Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theoretical and Applied Genetics,* 124**,** 423-432.

BERKMAN, P. J., VISENDI, P., LEE, H. C., STILLER, J., MANOLI, S., LORENC, M. T., LAI, K., BATLEY, J., FLEURY, D., SIMKOVA, H., KUBALAKOVA, M., WEINING, S., DOLEZEL, J. & EDWARDS, D. 2013a. Dispersion and domestication shaped the genome of bread wheat. *Plant Biotechnol J,* 11**,** 564-71.

BERKMAN, P. J., VISENDI, P., LEE, H. C., STILLER, J., MANOLI, S., LORENC, M. T., LAI, K., BATLEY, J., FLEURY, D., SIMKOVA, H., KUBALAKOVA, M., WEINING, S., DOLEZEL, J. & EDWARDS, D. 2013b. Dispersion and domestication shaped the genome of bread wheat. *Plant Biotechnology Journal,* 11**,** 564-571.

BEUZEN, N. D., STEAR, M. J. & CHANG, K. C. 2000. Molecular markers and their use in animal breeding. *The Veterinary Journal,* 160**,** 42-52.

BIOINFOLOGICS. 2016. *The w2rap-contigger* [Online]. Available: http://bioinfologics.github.io/the-w2rap-contigger/ [Accessed 2016].

BOETZER, M., HENKEL, C. V., JANSEN, H. J., BUTLER, D. & PIROVANO, W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics,* 27.

BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-2120.

BOŽA, V., BREJOVÁ, B. & VINAŘ, T. 2014. GAML: Genome Assembly by Maximum Likelihood. *In:* BROWN, D. & MORGENSTERN, B. (eds.) *Algorithms in Bioinformatics: 14th International Workshop, WABI 2014, Wroclaw, Poland, September 8-10, 2014. Proceedings.* Berlin, Heidelberg: Springer Berlin Heidelberg.

BOŽA, V., BREJOVÁ, B. & VINAŘ, T. 2017. DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. *PLOS ONE,* 12**,** e0178751.

BRADNAM, K. R., FASS, J. N., ALEXANDROV, A., BARANAY, P., BECHNER, M., BIROL, I., BOISVERT, S., CHAPMAN, J. A., CHAPUIS, G., CHIKHI, R.,

CHITSAZ, H., CHOU, W.-C., CORBEIL, J., DEL FABBRO, C., DOCKING, T. R., DURBIN, R., EARL, D., EMRICH, S., FEDOTOV, P., FONSECA, N. A., GANAPATHY, G., GIBBS, R. A., GNERRE, S., GODZARIDIS, É., GOLDSTEIN, S., HAIMEL, M., HALL, G., HAUSSLER, D., HIATT, J. B., HO, I. Y., HOWARD, J., HUNT, M., JACKMAN, S. D., JAFFE, D. B., JARVIS, E. D., JIANG, H., KAZAKOV, S., KERSEY, P. J., KITZMAN, J. O., KNIGHT, J. R., KOREN, S., LAM, T.-W., LAVENIER, D., LAVIOLETTE, F., LI, Y., LI, Z., LIU, B., LIU, Y., LUO, R., MACCALLUM, I., MACMANES, M. D., MAILLET, N., MELNIKOV, S., NAQUIN, D., NING, Z., OTTO, T. D., PATEN, B., PAULO, O. S., PHILLIPPY, A. M., PINA-MARTINS, F., PLACE, M., PRZYBYLSKI, D., QIN, X., QU, C., RIBEIRO, F. J., RICHARDS, S., ROKHSAR, D. S., RUBY, J. G., SCALABRIN, S., SCHATZ, M. C., SCHWARTZ, D. C., SERGUSHICHEV, A., SHARPE, T., SHAW, T. I., SHENDURE, J., SHI, Y., SIMPSON, J. T., SONG, H., TSAREV, F., VEZZI, F., VICEDOMINI, R., VIEIRA, B. M., WANG, J., WORLEY, K. C., YIN, S., YIU, S.-M., YUAN, J., ZHANG, G., ZHANG, H., ZHOU, S. & KORF, I. F. 2013a. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience,* 2**,** 1-31.

BRADNAM, K. R., FASS, J. N. & KORF, I. F. 2013b. CEGMA gene predictions for Assemblathon 2 entries. *GigaScience Database.*

BRENCHLEY, R., SPANNAGL, M., PFEIFER, M., BARKER, G. L. A., D/'AMORE, R., ALLEN, A. M., MCKENZIE, N., KRAMER, M., KERHORNOU, A., BOLSER, D., KAY, S., WAITE, D., TRICK, M., BANCROFT, I., GU, Y., HUO, N., LUO, M.-C., SEHGAL, S., GILL, B., KIANIAN, S., ANDERSON, O., KERSEY, P., DVORAK, J., MCCOMBIE, W. R., HALL, A., MAYER, K. F. X., EDWARDS, K. J., BEVAN, M. W. & HALL, N. 2012b. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature,* 491**,** 705-710.

BRYAN, G. J., COLLINS, A. J., STEPHENSON, P., ORRY, A., SMITH, J. B. & GALE, M. D. 1997. Isolation and characterisation of microsatellites from hexaploid bread wheat. *Theoretical and Applied Genetics,* 94**,** 557-563.

BUCKLER, T. E. S. & HOLTSFORD, T. P. 1996. Zea ribosomal repeat evolution and substitution patterns. *Molecular Biology and Evolution,* 13**,** 623-632.

BUSH, S. J., CASTILLO-MORALES, A., TOVAR-CORONA, J. M., CHEN, L., KOVER, P. X. & URRUTIA, A. O. 2014. Presence–Absence Variation in A.

thaliana Is Primarily Associated with Genomic Signatures Consistent with Relaxed Selective Constraints. *Molecular Biology and Evolution,* 31**,** 59-69.

CALDWELL, K. S., DVORAK, J., LAGUDAH, E. S., AKHUNOV, E., LUO, M.-C., WOLTERS, P. & POWELL, W. 2004. Sequence Polymorphism in Polyploid Wheat and Their D-Genome Diploid Ancestor. *Genetics,* 167**,** 941-947.

CAMACHO, C., COULOURIS, G., AVAGYAN, V., MA, N., PAPADOPOULOS, J. & BEALER, K. 2009. BLAST+: architecture and applications. *BMC Bioinformatics,* 10.

CAMPBELL, M. S., LAW, M., HOLT, C., STEIN, J. C., MOGHE, G. D., HUFNAGEL, D. E., LEI, J., ACHAWANANTAKUN, R., JIAO, D., LAWRENCE, C. J., WARE, D., SHIU, S. H., CHILDS, K. L., SUN, Y., JIANG, N. & YANDELL, M. 2014. MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol,* 164.

CANTAREL, B. L., KORF, I., ROBB, S. M., PARRA, G., ROSS, E., MOORE, B., HOLT, C., SANCHEZ ALVARADO, A. & YANDELL, M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res,* 18.

CAO, J., SCHNEEBERGER, K., OSSOWSKI, S., GUNTHER, T., BENDER, S., FITZ, J., KOENIG, D., LANZ, C., STEGLE, O., LIPPERT, C., WANG, X., OTT, F., MULLER, J., ALONSO-BLANCO, C., BORGWARDT, K., SCHMID, K. J. & WEIGEL, D. 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat Genet,* 43**,** 956-963.

CARTOLANO, M., HUETTEL, B., HARTWIG, B., REINHARDT, R. & SCHNEEBERGER, K. 2016. cDNA Library Enrichment of Full Length Transcripts for SMRT Long Read Sequencing. *PLOS ONE,* 11**,** e0157779.

CAVANAGH, C. R., CHAO, S., WANG, S., HUANG, B. E., STEPHEN, S., KIANI, S., FORREST, K., SAINTENAC, C., BROWN-GUEDIRA, G. L., AKHUNOVA, A., SEE, D., BAI, G., PUMPHREY, M., TOMAR, L., WONG, D., KONG, S., REYNOLDS, M., DA SILVA, M. L., BOCKELMAN, H., TALBERT, L., ANDERSON, J. A., DREISIGACKER, S., BAENZIGER, S., CARTER, A., KORZUN, V., MORRELL, P. L., DUBCOVSKY, J., MORELL, M. K., SORRELLS, M. E., HAYDEN, M. J. & AKHUNOV, E. 2013. Genome-wide comparative diversity uncovers multiple targets of selection for improvement

in hexaploid wheat landraces and cultivars. *Proceedings of the National Academy of Sciences of the United States of America,* 110**,** 8057-8062.

CAZAUX, B., SACOMOTO, G. & RIVALS, E. 2016. Superstring Graph: A New Approach for Genome Assembly. *In:* DONDI, R., FERTIN, G. & MAURI, G. (eds.) *Algorithmic Aspects in Information and Management: 11th International Conference, AAIM 2016, Bergamo, Italy, July 18-20, 2016, Proceedings.* Cham: Springer International Publishing.

CHANTRET, N., SALSE, J., SABOT, F., RAHMAN, S., BELLEC, A., LAUBIN, B., DUBOIS, I., DOSSAT, C., SOURDILLE, P., JOUDRIER, P., GAUTIER, M. F., CATTOLICO, L., BECKERT, M., AUBOURG, S., WEISSENBACH, J., CABOCHE, M., BERNARD, M., LEROY, P. & CHALHOUB, B. 2005. Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *Plant Cell,* 17**,** 1033-1045.

CHAO, S., ZHANG, W., AKHUNOV, E., SHERMAN, J., MA, Y., LUO, M.-C. & DUBCOVSKY, J. 2009. Analysis of gene-derived SNP marker polymorphism in US wheat (Triticum aestivum L.) cultivars. *Molecular Breeding,* 23**,** 23-33.

CHAPMAN, J. A., MASCHER, M., BULUÇ, A., BARRY, K., GEORGANAS, E., SESSION, A., STRNADOVA, V., JENKINS, J., SEHGAL, S., OLIKER, L., SCHMUTZ, J., YELICK, K. A., SCHOLZ, U., WAUGH, R., POLAND, J. A., MUEHLBAUER, G. J., STEIN, N. & ROKHSAR, D. S. 2015a. A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biology,* 16**,** 1-17.

CHAUDHARI, N. M., GUPTA, V. K. & DUTTA, C. 2016. BPGA- an ultra-fast pan-genome analysis pipeline. 6**,** 24373.

CHEN, X., MIN, D., YASIR, T. A. & HU, Y.-G. 2012. Genetic Diversity, Population Structure and Linkage Disequilibrium in Elite Chinese Winter Wheat Investigated with SSR Markers. *PLOS ONE,* 7**,** e44510.

CHEN, Z. J. & NI, Z. 2006. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *Bioessays.,* 28.

CHOULET, F., WICKER, T., RUSTENHOLZ, C., PAUX, E., SALSE, J., LEROY, P., SCHLUB, S., LE PASLIER, M.-C., MAGDELENAT, G., GONTHIER, C., COULOUX, A., BUDAK, H., BREEN, J., PUMPHREY, M., LIU, S., KONG, X., JIA, J., GUT, M., BRUNEL, D., ANDERSON, J. A., GILL, B. S., APPELS, R.,

KELLER, B. & FEUILLET, C. 2010. Megabase Level Sequencing Reveals Contrasted Organization and Evolution Patterns of the Wheat Gene and Transposable Element Spaces. *The Plant Cell,* 22**,** 1686-1701.

CINGOLANI, P., PLATTS, A., WANG, L. L., COON, M., TUNG, N., WANG, L., LAND, S. J., LU, X. & RUDEN, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. *Fly,* 6**,** 80-92.

CLARKSON, J. J., LIM, K. Y., KOVARIK, A., CHASE, M. W., KNAPP, S. & LEITCH, A. R. 2005. Long-term genome diploidization in allopolyploid Nicotiana section Repandae (Solanaceae). *New Phytologist,* 168**,** 241-252.

CLAVIJO, B. J., VENTURINI, L., SCHUDOMA, C., ACCINELLI, G. G., KAITHAKOTTIL, G., WRIGHT, J., BORRILL, P., KETTLEBOROUGH, G., HEAVENS, D., CHAPMAN, H., LIPSCOMBE, J., BARKER, T., LU, F.-H., MCKENZIE, N., RAATS, D., RAMIREZ-GONZALEZ, R. H., COINCE, A., PEEL, N., PERCIVAL-ALWYN, L., DUNCAN, O., TRÖSCH, J., YU, G., BOLSER, D. M., NAMAATI, G., KERHORNOU, A., SPANNAGL, M., GUNDLACH, H., HABERER, G., DAVEY, R. P., FOSKER, C., PALMA, F. D., PHILLIPS, A. L., MILLAR, A. H., KERSEY, P. J., UAUY, C., KRASILEVA, K. V., SWARBRECK, D., BEVAN, M. W. & CLARK, M. D. 2017. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Research,* 27**,** 885-896.

CLAVIJO, B. J., VENTURINI, L., SCHUDOMA, C., GARCIA ACCINELLI, G., KAITHAKOTTIL, G., WRIGHT, J., BORRILL, P., KETTLEBOROUGH, G., HEAVENS, D., CHAPMAN, H., LIPSCOMBE, J., BARKER, T., LU, F.-H., MCKENZIE, N., RAATS, D., RAMIREZ-GONZALEZ, R. H., COINCE, A., PEEL, N., PERCIVAL-ALWYN, L., DUNCAN, O., TRÖSCH, J., YU, G., BOLSER, D., NAAMATI, G., KERHORNOU, A., SPANNAGL, M., GUNDLACH, H., HABERER, G., DAVEY, R. P., FOSKER, C., DI PALMA, F., PHILLIPS, A., MILLAR, A. H., KERSEY, P. J., UAUY, C., KRASILEVA, K. V., SWARBRECK, D., BEVAN, M. W. & CLARK, M. D. 2016. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *bioRxiv*.

COGEPEDIA. 2017. *Sequenced Plant genomes* [Online]. online: onlie. Available: https://genomevolution.org/wiki/index.php/Sequenced_plant_genomes [Accessed 2017].

COKUS, S. J., FENG, S., ZHANG, X., CHEN, Z., MERRIMAN, B. & HAUDENSCHILD, C. D. 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature.,* 452.

COLLARD, B. C. Y. & MACKILL, D. J. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 363**,** 557-572.

COLLINS, R. E. & HIGGS, P. G. 2012. Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Mol Biol Evol,* 29**,** 3413-25.

CONANT, G. C., BIRCHLER, J. A. & PIRES, J. C. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr Opin Plant Biol.,* 19.

CONESA, A. & GÖTZ, S. 2008. Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. *International Journal of Plant Genomics,* 2008**,** 619832.

CONESA, A., GÖTZ, S., GARCÍA-GÓMEZ, J. M., TEROL, J., TALÓN, M. & ROBLES, M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics,* 21**,** 3674-3676.

CONSORTIUM, I. W. G. S. 2016a. IWGSC whole genome shotgun sequencing of Chinese Spring: Towards a Reference Sequence of Wheat. *Plan and Animal Genoem Conference XXIV.* San Diego.

CONSORTIUM, I. W. G. S. 2016b. Wheat Sequencing Consortium Releases Key Resource to the Scientific Community. wheatgenome.org.

CONSORTIUM, T. I. W. G. S. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science,* 345**,** 1251788.

COULONDRE, C., MILLER, J. H., FARABAUGH, P. J. & GILBERT, W. 1978. Molecular basis of base substitution hotspots in Escherichia coli. *Nature,* 274**,** 775-780.

CRONN, R., LISTON, A., PARKS, M., GERNANDT, D. S., SHEN, R. & MOCKLER, T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research,* 36**,** e122-e122.

CROS, D., DENIS, M., SANCHEZ, L., COCHARD, B., FLORI, A., DURAND-GASSELIN, T., NOUY, B., OMORE, A., POMIES, V., RIOU, V., SURYANA, E. & BOUVET, J.-M. 2015. Genomic selection prediction accuracy in a perennial crop: case study of oil palm (Elaeis guineensis Jacq.). *Theoretical and Applied Genetics,* 128**,** 397-410.

CROSSA, J., PEREZ, P., HICKEY, J., BURGUENO, J., ORNELLA, L., CERON-ROJAS, J., ZHANG, X., DREISIGACKER, S., BABU, R., LI, Y., BONNETT, D. & MATHEWS, K. 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity,* 112**,** 48-60.

CUI, P., LIU, H., LIN, Q., DING, F., ZHUO, G., HU, S., LIU, D., YANG, W., ZHAN, K., ZHANG, A. & YU, J. 2009. A complete mitochondrial genome of wheat (Triticum aestivum cv. Chinese Yumai), and fast evolving mitochondrial genes in higher plants. *J Genet,* 88**,** 299-307.

DAMANIA, A. B. 1998. Diversity of major cultivated plants domesticated in the Near East.

DANECEK, P., AUTON, A., ABECASIS, G., ALBERS, C. A., BANKS, E., DEPRISTO, M. A., HANDSAKER, R. E., LUNTER, G., MARTH, G. T., SHERRY, S. T., MCVEAN, G., DURBIN, R. & GROUP, G. P. A. 2011. The variant call format and VCFtools. *Bioinformatics,* 27**,** 2156-2158.

DAVENPORT, C. F. & TÜMMLER, B. 2013. Advances in computational analysis of metagenome sequences. *Environmental Microbiology,* 15**,** 1-5.

DAVID, M., DURSI, L. J., YAO, D., BOUTROS, P. C. & SIMPSON, J. T. 2017. Nanocall: an open source basecaller for Oxford Nanopore sequencing data. *Bioinformatics,* 33**,** 49-55.

DE LA BASTIDE, M. & MCCOMBIE, W. R. 2007. Assembling genomic DNA sequences with PHRAP. *Curr Protoc Bioinformatics,* Chapter 11**,** Unit11.4.

DEL BLANCO, I. A., RAJARAM, S. & KRONSTAD, W. E. 2001. Agronomic Potential of Synthetic Hexaploid Wheat-Derived Populations I.A. del Blanco present address: Dep. of Plant Sciences, North Dakota State Univ., Fargo, ND 58105. Technical Paper no. 11547 of the Oregon State Univ. Agric. Expt. Stn. *Crop Science,* 41**,** 670-676.

DELCHER, A. L., PHILLIPPY, A., CARLTON, J. & SALZBERG, S. L. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res,* 30**,** 2478-83.

DÍAZ, A., ZIKHALI, M., TURNER, A. S., ISAAC, P. & LAURIE, D. A. 2012. Copy Number Variation Affecting the Photoperiod-B1 and Vernalization-A1 Genes Is Associated with Altered Flowering Time in Wheat (Triticum aestivum). *PLOS ONE,* 7**,** e33234.

DING, J., ARAKI, H., WANG, Q., ZHANG, P., YANG, S., CHEN, J.-Q. & TIAN, D. 2007. Highly asymmetric rice genomes. *BMC Genomics,* 8**,** 154.

DUBCOVSKY, J. & DVORAK, J. 2007b. Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* 316**,** 1862-1866.

DUITAMA, J., SILVA, A., SANABRIA, Y., CRUZ, D. F., QUINTERO, C., BALLEN, C., LORIEUX, M., SCHEFFLER, B., FARMER, A., TORRES, E., OARD, J. & TOHME, J. 2015. Whole Genome Sequencing of Elite Rice Cultivars as a Comprehensive Information Resource for Marker Assisted Selection. *PLOS ONE,* 10**,** e0124617.

DVORAK, J., AKHUNOV, E. D., AKHUNOV, A. R., DEAL, K. R. & LUO, M.-C. 2006. Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat. *Mol Biol Evol,* 23.

DVORAK, J., LUO, M. C., YANG, Z. L. & ZHANG, H. B. 1998. The structure of the Aegilops tauschii genepool and the evolution of hexaploid wheat. *TAG Theor Appl Genet,* 97.

DVORAK, J., MCGUIRE, P. & CASSIDY, B. 1988. Apparent sources of the A genomes of wheats inferred from the polymorphism in abundance and restriction fragment length of repeated nucleotide sequences. *Genome,* 30.

DVORÁK, J., TERLIZZI, P., ZHANG, H. B. & RESTA, P. 1993. The evolution of polyploid wheats: identification of the A genome donor species. *Genome,* 36.

DVOŘÁK, J., TERLIZZI, P. D., ZHANG, H.-B. & RESTA, P. 1993. The evolution of polyploid wheats: identification of the A genome donor species. *Genome,* 36**,** 21-31.

EDAE, E. A., BOWDEN, R. L. & POLAND, J. 2015. Application of Population Sequencing (POPSEQ) for Ordering and Imputing Genotyping-by-Sequencing Markers in Hexaploid Wheat. *G3: Genes|Genomes|Genetics*.

EDWARDS, D. & BATLEY, J. 2004. Plant bioinformatics: from genome to phenome. *Trends Biotechnol,* 22**,** 232-7.

EDWARDS, D., BATLEY, J. & SNOWDON, R. 2013a. Accessing complex crop genomes with next-generation sequencing. *Theoretical and Applied Genetics,* 126**,** 1-11.

EDWARDS, D., BATLEY, J. & SNOWDON, R. J. 2013b. Accessing complex crop genomes with next-generation sequencing. *Theoretical and Applied Genetics,* 126**,** 1-11.

EDWARDS, D., WILCOX, S., BARRERO, R. A., FLEURY, D., CAVANAGH, C. R., FORREST, K. L., HAYDEN, M. J., MOOLHUIJZEN, P., KEEBLE-GAGNÈRE, G., BELLGARD, M. I., LORENC, M. T., SHANG, C. A., BAUMANN, U., TAYLOR, J. M., MORELL, M. K., LANGRIDGE, P., APPELS, R. & FITZGERALD, A. 2012. Bread matters: a national initiative to profile the genetic diversity of Australian wheat. *Plant Biotechnology Journal,* 10**,** 703-708.

EICHTEN, S. R., FOERSTER, J. M., DE LEON, N., KAI, Y., YEH, C.-T., LIU, S., JEDDELOH, J. A., SCHNABLE, P. S., KAEPPLER, S. M. & SPRINGER, N. M. 2011. B73-Mo17 Near-Isogenic Lines Demonstrate Dispersed Structural Variation in Maize. *Plant Physiology,* 156**,** 1679-1690.

EID, J., FEHR, A., GRAY, J., LUONG, K., LYLE, J., OTTO, G., PELUSO, P., RANK, D., BAYBAYAN, P., BETTMAN, B., BIBILLO, A., BJORNSON, K., CHAUDHURI, B., CHRISTIANS, F., CICERO, R., CLARK, S., DALAL, R., DEWINTER, A., DIXON, J., FOQUET, M., GAERTNER, A., HARDENBOL, P., HEINER, C., HESTER, K., HOLDEN, D., KEARNS, G., KONG, X., KUSE, R., LACROIX, Y., LIN, S., LUNDQUIST, P., MA, C., MARKS, P., MAXHAM, M., MURPHY, D., PARK, I., PHAM, T., PHILLIPS, M., ROY, J., SEBRA, R., SHEN, G., SORENSON, J., TOMANEY, A., TRAVERS, K., TRULSON, M., VIECELI, J., WEGENER, J., WU, D., YANG, A., ZACCARIN, D., ZHAO, P., ZHONG, F., KORLACH, J. & TURNER, S. 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science,* 323**,** 133-138.

EISENSTEIN, M. 2012. Oxford Nanopore announcement sets sequencing sector abuzz. *Nat Biotech,* 30**,** 295-296.

ELSIK, C. G., WORLEY, K. C., ZHANG, L., MILSHINA, N. V., JIANG, H., REESE, J. T., CHILDS, K. L., VENKATRAMAN, A., DICKENS, C. M. & WEINSTOCK, G. M. 2006. Community annotation: procedures, protocols and supporting tools. *Genome Res,* 16.

ENSEMBL 2017. Plant genomes.

EVENSON, R. E. & GOLLIN, D. 2003. Assessing the Impact of the Green Revolution, 1960 to 2000. *Science,* 300**,** 758-762.

FAN, L., ZHANG, M.-Y., LIU, Q.-Z., LI, L.-T., SONG, Y., WANG, L.-F., ZHANG, S.-L. & WU, J. 2013. Transferability of Newly Developed Pear SSR Markers to Other Rosaceae Species. *Plant Molecular Biology Reporter,* 31**,** 1271-1282.

FAO 2016. *Save and grow in practice. A guide to sustainable cereal production,* Rome, FAO.

FELDMAN, M. & KISLEV, M. E. 2007. Domestication of emmer wheat and evolution of free-threshing tetraploid wheat. *Israel Journal of Plant Sciences,* 55**,** 207-221.

FERREIRA, A., SILVA, M. F. D., SILVA, L. D. C. E. & CRUZ, C. D. 2006. Estimating the effects of population size and type on the accuracy of genetic maps. *Genetics and Molecular Biology,* 29**,** 187-192.

FEUILLET, C., LANGRIDGE, P. & WAUGH, R. 2008. Cereal breeding takes a walk on the wild side. *Trends Genet,* 24.

FEUK, L., CARSON, A. R. & SCHERER, S. W. 2006. Structural variation in the human genome. *Nat Rev Genet,* 7**,** 85-97.

FLAVELL, R., BENNETT, M., SMITH, J. & SMITH, D. 1974. Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochem Genet,* 12.

FORREST, K. L., PUJOL, V., BULLI, P., PUMPHREY, M., WELLINGS, C. & HERRERA-FOESSEL, S. 2014. Development of a SNP marker assay for the Lr67 gene of wheat using a genotyping by sequencing approach. *Mol Breed,* 34.

FREELING, M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu Rev Plant Biol,* 60**,** 433-53.

FRIEBE, B. & GILL, B. S. 1994. C-band polymorphism and structural rearrangements detected in common wheat (Triticum aestivum). *Euphytica,* 78**,** 1-5.

FROMMER, M., MCDONALD, L. E., MILLAR, D. S., COLLIS, C. M., WATT, F., GRIGG, G. W., MOLLOY, P. L. & PAUL, C. L. 1992. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in

individual DNA strands. *Proceedings of the National Academy of Sciences,* 89**,** 1827-1831.

FU, Y., LUO, G. Z., CHEN, K., DENG, X., YU, M. & HAN, D. 2015. N6-methyldeoxyadenosine marks active transcription start sites in chlamydomonas. *Cell,* 161.

GABALDON, T. & KOONIN, E. V. 2013. Functional and evolutionary implications of gene orthology. *Nat Rev Genet,* 14**,** 360-366.

GAN, X., STEGLE, O., BEHR, J., STEFFEN, J. G., DREWE, P., HILDEBRAND, K. L., LYNGSOE, R., SCHULTHEISS, S. J., OSBORNE, E. J., SREEDHARAN, V. T., KAHLES, A., BOHNERT, R., JEAN, G., DERWENT, P., KERSEY, P., BELFIELD, E. J., HARBERD, N. P., KEMEN, E., TOOMAJIAN, C., KOVER, P. X., CLARK, R. M., RATSCH, G. & MOTT, R. 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature,* 477**,** 419-423.

GARDINER, L.-J., QUINTON-TULLOCH, M., OLOHAN, L., PRICE, J., HALL, N. & HALL, A. 2015. A genome-wide survey of DNA methylation in hexaploid wheat. *Genome Biology,* 16**,** 273.

GILES, R. J. & BROWN, T. A. 2006. GluDy allele variations in Aegilops tauschii and Triticum aestivum: implications for the origins of hexaploid wheats. *Theor Appl Genet,* 112**,** 1563-72.

GILL, B. S., APPELS, R., BOTHA-OBERHOLSTER, A.-M., BUELL, C. R., BENNETZEN, J. L., CHALHOUB, B., CHUMLEY, F., DVOŘÁK, J., IWANAGA, M., KELLER, B., LI, W., MCCOMBIE, W. R., OGIHARA, Y., QUETIER, F. & SASAKI, T. 2004. A Workshop Report on Wheat Genome Sequencing: International Genome Research on Wheat Consortium. *Genetics,* 168**,** 1087-1096.

GILL, K. S., GILL, B. S., ENDO, T. R. & TAYLOR, T. 1996. Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. *Genetics,* 144.

GIOVANNI PARMIGIANI, E. S. G., RAFAEL A. IRIZARRY, SCOTT L. ZEGER 2003. *The Analysis of Gene Expression Data*, Springer New York.

GOFF, S. A., RICKE, D., LAN, T.-H., PRESTING, G., WANG, R., DUNN, M., GLAZEBROOK, J., SESSIONS, A., OELLER, P., VARMA, H., HADLEY, D., HUTCHISON, D., MARTIN, C., KATAGIRI, F., LANGE, B. M., MOUGHAMER, T., XIA, Y., BUDWORTH, P., ZHONG, J., MIGUEL, T., PASZKOWSKI, U.,

ZHANG, S., COLBERT, M., SUN, W.-L., CHEN, L., COOPER, B., PARK, S., WOOD, T. C., MAO, L., QUAIL, P., WING, R., DEAN, R., YU, Y., ZHARKIKH, A., SHEN, R., SAHASRABUDHE, S., THOMAS, A., CANNINGS, R., GUTIN, A., PRUSS, D., REID, J., TAVTIGIAN, S., MITCHELL, J., ELDREDGE, G., SCHOLL, T., MILLER, R. M., BHATNAGAR, S., ADEY, N., RUBANO, T., TUSNEEM, N., ROBINSON, R., FELDHAUS, J., MACALMA, T., OLIPHANT, A. & BRIGGS, S. 2002. A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. japonica). *Science,* 296**,** 92-100.

GOLICZ, A. A. 2016. *Construction and analysis of the Brassica oleracea pangenome.* Phylosophical doctor PhD, The University of Queensland.

GOLICZ, A. A., BATLEY, J. & EDWARDS, D. 2015a. Towards plant pangenomics. *Plant Biotechnology Journal,* 14**,** 1099-1105.

GOLICZ, A. A., BATLEY, J. & EDWARDS, D. 2016a. Towards plant pangenomics. *Plant Biotechnol J,* 14**,** 1099-105.

GOLICZ, A. A., BAYER, P. E., BARKER, G. C., EDGER, P. P., KIM, H., MARTINEZ, P. A., CHAN, C. K. K., SEVERN-ELLIS, A., MCCOMBIE, W. R., PARKIN, I. A. P., PATERSON, A. H., PIRES, J. C., SHARPE, A. G., TANG, H., TEAKLE, G. R., TOWN, C. D., BATLEY, J. & EDWARDS, D. 2016b. The pangenome of an agronomically important crop plant Brassica oleracea. *Nature Communications,* 7**,** 13390.

GOLICZ, A. A., MARTINEZ, P. A., ZANDER, M., PATEL, D. A., VAN DE WOUW, A. P., VISENDI, P., FITZGERALD, T. L., EDWARDS, D. & BATLEY, J. 2015b. Gene loss in the fungal canola pathogen Leptosphaeria maculans. *Functional & Integrative Genomics,* 15**,** 189-196.

GOLICZ, A. A., SCHLIEP, M., LEE, H. T., LARKUM, A. W. D., DOLFERUS, R., BATLEY, J., CHAN, C.-K. K., SABLOK, G., RALPH, P. J. & EDWARDS, D. 2015c. Genome-wide survey of the seagrass Zostera muelleri suggests modification of the ethylene signalling network. *Journal of Experimental Botany*.

GOLIKZ, A. A. 2016. *Construction and analysis of the Brassica aleracea pangenome.* Doctor in Philosophy, The university of Queensland.

GONZALEZ-GARAY, M. L. 2016. Introduction to Isoform Sequencing Using Pacific Biosciences Technology (Iso-Seq). *In:* WU, J. (ed.) *Transcriptomics and Gene Regulation.* Dordrecht: Springer Netherlands.

GOODWIN, S., GURTOWSKI, J., ETHE-SAYERS, S., DESHPANDE, P., SCHATZ, M. C. & MCCOMBIE, W. R. 2015. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research,* 25**,** 1750-1756.

GOODWIN, S., MCPHERSON, J. D. & MCCOMBIE, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet,* 17**,** 333-351.

GORDON, S. P., PRIEST, H., DES MARAIS, D. L., SCHACKWITZ, W., FIGUEROA, M., MARTIN, J., BRAGG, J. N., TYLER, L., LEE, C.-R., BRYANT, D., WANG, W., MESSING, J., MANZANEDA, A. J., BARRY, K., GARVIN, D. F., BUDAK, H., TUNA, M., MITCHELL-OLDS, T., PFENDER, W. F., JUENGER, T. E., MOCKLER, T. C. & VOGEL, J. P. 2014a. Genome diversity in Brachypodium distachyon: deep sequencing of highly diverse inbred lines. *The Plant Journal,* 79**,** 361-374.

GORE, M. A., CHIA, J.-M., ELSHIRE, R. J., SUN, Q., ERSOZ, E. S. & HURWITZ, B. L. 2009. A first-generation haplotype map of maize. *Science,* 326.

GREER, E. L., BLANCO, M. A., GU, L., SENDINC, E., LIU, J. & ARISTIZÁBAL-CORRALES, D. 2015. DNA methylation on N6-adenine in C. elegans. *Cell,* 161.

GREGORY R. WARNES, B. B., LODEWIJK BONEBAKKER, ROBERT GENTLEMAN, WOLFGANG HUBER ANDY LIAW, THOMAS LUMLEY, MARTIN MAECHLER, ARNI MAGNUSSON, STEFFEN MOELLER, MARC SCHWARTZ AND BILL VENABLES 2015. gplots: Various R Programming Tools for Plotting Data. R package version 2.17.0 ed.

GRIFFITHS, S., SHARP, R., FOOTE, T. N., BERTIN, I., WANOUS, M., READER, S., COLAS, I. & MOORE, G. 2006. Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. *Nature,* 439**,** 749-752.

GROVER, C., GALLAGHER, J., SZADKOWSKI, E., YOO, M., FLAGEL, L. & WENDEL, J. 2012. Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytol.,* 196.

GUO, Y., CHEN, S., LI, Z. & COWLING, W. A. 2014. Center of Origin and Centers of Diversity in an Ancient Crop, Brassica rapa (Turnip Rape). *Journal of Heredity,* 105**,** 555-565.

GUPTA, P. K., KULWAL, P. L. & RUSTGI, S. 2005. Wheat cytogenetics in the genomics era and its relevance to breeding. *Cytogenetic and Genome Research,* 109**,** 315-327.

GUPTA, P. K., MIR, R. R., MOHAN, A. & KUMAR, J. 2008. Wheat Genomics: Present Status and Future Prospects. *International Journal of Plant Genomics,* 2008**,** 896451.

HARCOURT, R. L. & GALE, M. D. 1991. A chromosome-specific DNA sequence which reveals a high level of RFLP in wheat. *Theoretical and Applied Genetics,* 81**,** 397-400.

HARDIGAN, M. A., CRISOVAN, E., HAMILTON, J. P., KIM, J., LAIMBEER, P., LEISNER, C. P., MANRIQUE-CARPINTERO, N. C., NEWTON, L., PHAM, G. M., VAILLANCOURT, B., YANG, X., ZENG, Z., DOUCHES, D. S., JIANG, J., VEILLEUX, R. E. & BUELL, C. R. 2016. Genome Reduction Uncovers a Large Dispensable Genome and Adaptive Role for Copy Number Variation in Asexually Propagated Solanum tuberosum. *The Plant Cell,* 28**,** 388-405.

HAUDRY, A., CENCI, A., RAVEL, C., BATAILLON, T., BRUNEL, D., PONCET, C., HOCHU, I., POIRIER, S., SANTONI, S., GLÉMIN, S. & DAVID, J. 2007. Grinding up Wheat: A Massive Loss of Nucleotide Diversity Since Domestication. *Molecular Biology and Evolution,* 24**,** 1506-1517.

HAYATSU, H., WATAYA, Y. & KAI, K. 1970. Addition of sodium bisulfite to uracil and to cytosine. *Journal of the American Chemical Society,* 92**,** 724-726.

HAYWARD, A. M., ANNALIESE S.; DALTON-MORGAN, JESSICA; ZANDER, MANUEL; EDWARDS, DAVID; BATLEY, JACQUELINE 2012. SNP discovery and applications in Brassica napus. *Journal of Plant Biotechnology,* 39**,** 49-61.

HEDDEN, P. 2003. The genes of the Green Revolution. *Trends in Genetics,* 19**,** 5-9.

HEFFNER, E. L., LORENZ, A. J., JANNINK, J.-L. & SORRELLS, M. E. 2010. Plant Breeding with Genomic Selection: Gain per Unit Time and Cost All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher. *Crop Science,* 50**,** 1681-1690.

HESLOT, N., YANG, H.-P., SORRELLS, M. E. & JANNINK, J.-L. 2012. Genomic Selection in Plant Breeding: A Comparison of Models. *Crop Science,* 52**,** 146-160.

HEUN, M., SCHÄFER-PREGL, R., KLAWAN, D., CASTAGNA, R., ACCERBI, M., BORGHI, B. & SALAMINI, F. 1997. Site of Einkorn Wheat Domestication Identified by DNA Fingerprinting. *Science,* 278**,** 1312-1314.

HIRSCH, C. N., FOERSTER, J. M., JOHNSON, J. M., SEKHON, R. S., MUTTONI, G. & VAILLANCOURT, B. 2014. Insights into the maize pan-genome and pan-transcriptome. *Plant Cell.,* 26.

HOLT, C. & YANDELL, M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics,* 12**,** 491.

HUANG, S., SIRIKHACHORNKIT, A., SU, X., FARIS, J., GILL, B., HASELKORN, R. & GORNICKI, P. 2002a. Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 8133-8138.

HUANG, X., BÖRNER, A., RÖDER, M. & GANAL, M. 2002b. Assessing genetic diversity of wheat (Triticum aestivum L.) germplasm using microsatellite markers. *Theoretical and Applied Genetics,* 105**,** 699-707.

HUANG, X., WANG, J., ALURU, S., YANG, S.-P. & HILLIER, L. 2003. PCAP: A Whole-Genome Assembly Program. *Genome Research,* 13**,** 2164-2170.

HUNT, M., KIKUCHI, T., SANDERS, M., NEWBOLD, C., BERRIMAN, M. & OTTO, T. D. 2013. REAPR: a universal tool for genome assembly evaluation. *Genome Biology,* 14**,** R47.

HUO, N., VOGEL, J., LAZO, G., YOU, F., MA, Y., MCMAHON, S., DVORAK, J., ANDERSON, O., LUO, M.-C. & GU, Y. 2009. Structural characterization of Brachypodium genome and its syntenic relationship with rice and wheat. *Plant Molecular Biology,* 70**,** 47-61.

HUSE, S. M., HUBER, J. A., MORRISON, H. G., SOGIN, M. L. & WELCH, D. M. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol,* 8**,** R143.

IAFRATE, A. J., FEUK, L., RIVERA, M. N., LISTEWNIK, M. L., DONAHOE, P. K., QI, Y., SCHERER, S. W. & LEE, C. 2004. Detection of large-scale variation in the human genome. *Nat Genet,* 36**,** 949-951.

IAKOUBOV, L., MOSSAKOWSKA, M., SZWED, M., DUAN, Z., SESTI, F. & PUZIANOWSKA-KUZNICKA, M. 2013. A Common Copy Number Variation (CNV) Polymorphism in the CNTNAP4 Gene: Association with Aging in Females. *PLOS ONE,* 8**,** e79790.

IEHISA, J. C. M., OHNO, R., KIMURA, T., ENOKI, H., NISHIMURA, S., OKAMOTO, Y., NASUDA, S. & TAKUMI, S. 2014. A High-Density Genetic Map with Array-Based Markers Facilitates Structural and Quantitative Trait Locus Analyses of the Common Wheat Genome. *DNA Research,* 21**,** 555-567.

INTERNATIONAL BRACHYPODIUM, I. 2010. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature,* 463**,** 763-768.

IOVENE, M., ZHANG, T., LOU, Q., BUELL, C. R. & JIANG, J. 2013. Copy number variation in potato – an asexually propagated autotetraploid species. *The Plant Journal,* 75**,** 80-89.

IQBAL, Z., CACCAMO, M., TURNER, I., FLICEK, P. & MCVEAN, G. 2012. De novo assembly and genotyping of variants using colored de bruijn graphs. *Nat Genet,* 44.

IRISH, V. F. & LITT, A. 2005. Flower development and evolution: gene duplication, diversification and redeployment. *Current Opinion in Genetics & Development,* 15**,** 454-460.

ISHII, T., MORI, N. & OGIHARA, Y. 2001. Evaluation of allelic diversity at chloroplast microsatellite loci among common wheat and its ancestral species. *Theoretical and Applied Genetics,* 103**,** 896-904.

IWAKI, K., HARUNA, S., NIWA, T. & KATO, K. 2001. Adaptation and ecological differentiation in wheat with special reference to geographical variation of growth habit and Vrn genotype. *Plant Breeding,* 120**,** 107-114.

IWGSC 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science,* 345.

JACOBSEN, A., HENDRIKSEN, R. S., AARESTURP, F. M., USSERY, D. W. & FRIIS, C. 2011. The Salmonella enterica pan-genome. *Microb Ecol,* 62**,** 487-504.

JANNINK, J.-L., LORENZ, A. J. & IWATA, H. 2010. Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics,* 9**,** 166-177.

JIA, J., ZHAO, S., KONG, X., LI, Y., ZHAO, G., HE, W., APPELS, R., PFEIFER, M., TAO, Y., ZHANG, X., JING, R., ZHANG, C., MA, Y., GAO, L., GAO, C., SPANNAGL, M., MAYER, K. F. X., LI, D., PAN, S., ZHENG, F., HU, Q., XIA, X., LI, J., LIANG, Q., CHEN, J., WICKER, T., GOU, C., KUANG, H., HE, G., LUO, Y., KELLER, B., XIA, Q., LU, P., WANG, J., ZOU, H., ZHANG, R., XU, J., GAO, J., MIDDLETON, C., QUAN, Z., LIU, G., WANG, J., YANG, H., LIU, X., HE, Z., MAO, L. & WANG, J. 2013. Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature,* 496**,** 91-95.

JIN, M., LIU, H., HE, C., FU, J., XIAO, Y., WANG, Y., XIE, W., WANG, G. & YAN, J. 2016. Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. *Scientific Reports,* 6**,** 18936.

JORDAN, K. W., WANG, S., LUN, Y., GARDINER, L.-J., MACLACHLAN, R., HUCL, P., WIEBE, K., WONG, D., FORREST, K. L., SHARPE, A. G., SIDEBOTTOM, C. H., HALL, N., TOOMAJIAN, C., CLOSE, T., DUBCOVSKY, J., AKHUNOVA, A., TALBERT, L., BANSAL, U. K., BARIANA, H. S., HAYDEN, M. J., POZNIAK, C., JEDDELOH, J. A., HALL, A. & AKHUNOV, E. 2015. A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biology,* 16**,** 48.

JURKA, J., KAPITONOV, V. V., PAVLICEK, A., KLONOWSKI, P., KOHANY, O. & WALICHIEWICZ, J. 2005a. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res,* 110**,** 462-7.

KAEPPLER, S. 2012. Heterosis: Many Genes, Many Mechanisms—End the Search for an Undiscovered Unifying Theory. *ISRN Botany,* 2012**,** 12.

KAM-MORGAN, L. N. W. 1988. DNA restriction fragment length polymorphisms as genetic markers in mapping the wheat genome. *Dissertation Abstracts International, B (Sciences and Engineering),* 48**,** 2201B-2202B.

KASHKUSH, K., FELDMAN, M. & LEVY, A. A. 2002. Gene Loss, Silencing and Activation in a Newly Synthesized Wheat Allotetraploid. *Genetics,* 160**,** 1651-1659.

KEANE, T. M., GOODSTADT, L., DANECEK, P., WHITE, M. A., WONG, K., YALCIN, B., HEGER, A., AGAM, A., SLATER, G., GOODSON, M.,

FURLOTTE, N. A., ESKIN, E., NELLAKER, C., WHITLEY, H., CLEAK, J., JANOWITZ, D., HERNANDEZ-PLIEGO, P., EDWARDS, A., BELGARD, T. G., OLIVER, P. L., MCINTYRE, R. E., BHOMRA, A., NICOD, J., GAN, X., YUAN, W., VAN DER WEYDEN, L., STEWARD, C. A., BALA, S., STALKER, J., MOTT, R., DURBIN, R., JACKSON, I. J., CZECHANSKI, A., GUERRA-ASSUNCAO, J. A., DONAHUE, L. R., REINHOLDT, L. G., PAYSEUR, B. A., PONTING, C. P., BIRNEY, E., FLINT, J. & ADAMS, D. J. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature,* 477**,** 289-294.

KELLER, O., KOLLMAR, M., STANKE, M. & WAACK, S. 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics,* 27**,** 757-763.

KENAN-EICHLER, M., LESHKOWITZ, D., TAL, L., NOOR, E., MELAMED-BESSUDO, C. & FELDMAN, M. 2011. Wheat hybridization and polyploidization results in deregulation of small RNAs. *Genetics.,* 188.

KERSEY, P. J., ALLEN, J. E., ARMEAN, I., BODDU, S., BOLT, B. J., CARVALHO-SILVA, D., CHRISTENSEN, M., DAVIS, P., FALIN, L. J., GRABMUELLER, C., HUMPHREY, J., KERHORNOU, A., KHOBOVA, J., ARANGANATHAN, N. K., LANGRIDGE, N., LOWY, E., MCDOWALL, M. D., MAHESWARI, U., NUHN, M., ONG, C. K., OVERDUIN, B., PAULINI, M., PEDRO, H., PERRY, E., SPUDICH, G., TAPANARI, E., WALTS, B., WILLIAMS, G., TELLO–RUIZ, M., STEIN, J., WEI, S., WARE, D., BOLSER, D. M., HOWE, K. L., KULESHA, E., LAWSON, D., MASLEN, G. & STAINES, D. M. 2016b. Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Research,* 44**,** D574-D580.

KIHARA, H. 1919. Über cytologische Studien nei einige Getreidearten. I. Species-Bastarde des Weizens und Weizenroggen-Bastarde. *Botanical Magazine,* 32**,** 17-38.

KIHARA, H. 1966. Factors affecting the evolution of common wheat. *Indian J. Genet.,* 26A**,** 14-28.

KILIAN, B., OZKAN, H., WALTHER, A., KOHL, J., DAGAN, T., SALAMINI, F. & MARTIN, W. 2007. Molecular diversity at 18 loci in 321 wild and 92 domesticate lines reveal no reduction of nucleotide diversity during Triticum

monococcum (Einkorn) domestication: implications for the origin of agriculture. *Mol Biol Evol,* 24**,** 2657-68.

KIM, D., LANGMEAD, B. & SALZBERG, S. L. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Meth,* 12**,** 357-360.

KIM, D., PERTEA, G., TRAPNELL, C., PIMENTEL, H., KELLEY, R. & SALZBERG, S. L. 2013a. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology,* 14**,** 1-13.

KIM, D., PERTEA, G., TRAPNELL, C., PIMENTEL, H., KELLEY, R. & SALZBERG, S. L. 2013b. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology,* 14.

KOREN, S. & PHILLIPPY, A. M. 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology,* 23**,** 110-120.

KORLACH, J., BJORNSON, K. P., CHAUDHURI, B. P., CICERO, R. L., FLUSBERG, B. A., GRAY, J. J., HOLDEN, D., SAXENA, R., WEGENER, J. & TURNER, S. W. 2010. Real-Time DNA Sequencing from Single Polymerase Molecules. *Methods in Enzymology,* 472**,** 431-455.

KOZIOL, M. J., BRADSHAW, C. R., ALLEN, G. E., COSTA, A. S., FREZZA, C. & GURDON, J. B. 2015. Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. *Nat Struct Mol Biol*.

LAI, J., LI, R., XU, X., JIN, W., XU, M., ZHAO, H., XIANG, Z., SONG, W., YING, K., ZHANG, M., JIAO, Y., NI, P., ZHANG, J., LI, D., GUO, X., YE, K., JIAN, M., WANG, B., ZHENG, H., LIANG, H., ZHANG, X., WANG, S., CHEN, S., LI, J., FU, Y., SPRINGER, N. M., YANG, H., WANG, J., DAI, J., SCHNABLE, P. S. & WANG, J. 2010. Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet,* 42**,** 1027-1030.

LAI, K. 2015. *Genome Diversity in Triticum aestivum.* Philosophical doctor Research, University of Queensland.

LAI, K., BERKMAN, P. J., LORENC, M. T., DURAN, C., SMITS, L., MANOLI, S., STILLER, J. & EDWARDS, D. 2012a. WheatGenome.info: An Integrated Database and Portal for Wheat Genome Information. *Plant and Cell Physiology,* 53**,** e2.

LAI, K., DURAN, C., BERKMAN, P. J., LORENC, M. T., STILLER, J., MANOLI, S., HAYDEN, M. J., FORREST, K. L., FLEURY, D. & BAUMANN, U. 2012b.

Single nucleotide polymorphism discovery from wheat next-generation sequence data. *Plant Biotechnol J,* 10.

LAI, K., DURAN, C., BERKMAN, P. J., LORENC, M. T., STILLER, J., MANOLI, S., HAYDEN, M. J., FORREST, K. L., FLEURY, D., BAUMANN, U., ZANDER, M., MASON, A. S., BATLEY, J. & EDWARDS, D. 2012c. Single nucleotide polymorphism discovery from wheat next-generation sequence data. *Plant Biotechnology Journal,* 10**,** 743-749.

LAI, K., LORENC, M. T., LEE, H., BERKMAN, P. J., BAYER, P. E., VISENDI, P., RUPERAO, P., FITZGERALD, T. L., ZANDER, M., CHAN, C. K., MANOLI, S., STILLER, J., BATLEY, J. & EDWARDS, D. 2015a. Identification and characterisation of more than 4 million inter-varietal SNPs across the group 7 chromosomes of bread wheat. *Plant Biotechnology Journal* 13**,** 97-104.

LAI, K., LORENC, M. T., LEE, H. C., BERKMAN, P. J., BAYER, P. E., VISENDI, P., RUPERAO, P., FITZGERALD, T. L., ZANDER, M., CHAN, C.-K. K., MANOLI, S., STILLER, J., BATLEY, J. & EDWARDS, D. 2015b. Identification and characterization of more than 4 million intervarietal SNPs across the group 7 chromosomes of bread wheat. *Plant Biotechnology Journal,* 13**,** 97-104.

LAING, C., BUCHANAN, C., TABOADA, E. N., ZHANG, Y., KROPINSKI, A., VILLEGAS, A., THOMAS, J. E. & GANNON, V. P. 2010. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics,* 11**,** 461.

LAM, H.-M., XU, X., LIU, X., CHEN, W., YANG, G., WONG, F.-L., LI, M.-W., HE, W., QIN, N., WANG, B., LI, J., JIAN, M., WANG, J., SHAO, G., WANG, J., SUN, S. S.-M. & ZHANG, G. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet,* 42**,** 1053-1059.

LAMOUREUX, D., PETERSON, D. G., LI, W., FELLERS, J. P. & GILL, B. S. 2005. The efficacy of Cot-based gene enrichment in wheat (Triticum aestivum L.). *Genome,* 48**,** 1120-1126.

LANDER, E. S. & WATERMAN, M. S. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics,* 2**,** 231-9.

LANGMEAD, B. & SALZBERG, S. L. 2012a. Fast gapped-read alignment with Bowtie 2. *Nat Meth,* 9**,** 357-359.

LAPIERRE, P. & GOGARTEN, J. P. 2009. Estimating the size of the bacterial pan-genome. *Trends in Genetics,* 25**,** 107-110.

LAZO, G. R., CHAO, S., HUMMEL, D. D., EDWARDS, H., CROSSMAN, C. C., LUI, N., MATTHEWS, D. E., CAROLLO, V. L., HANE, D. L., YOU, F. M., BUTLER, G. E., MILLER, R. E., CLOSE, T. J., PENG, J. H., LAPITAN, N. L. V., GUSTAFSON, J. P., QI, L. L., ECHALIER, B., GILL, B. S., DILBIRLIGI, M., RANDHAWA, H. S., GILL, K. S., GREENE, R. A., SORRELLS, M. E., AKHUNOV, E. D., DVOŘÁK, J., LINKIEWICZ, A. M., DUBCOVSKY, J., HOSSAIN, K. G., KALAVACHARLA, V., KIANIAN, S. F., MAHMOUD, A. A., MIFTAHUDIN, MA, X.-F., CONLEY, E. J., ANDERSON, J. A., PATHAN, M. S., NGUYEN, H. T., MCGUIRE, P. E., QUALSET, C. O. & ANDERSON, O. D. 2004. Development of an Expressed Sequence Tag (EST) Resource for Wheat (Triticum aestivum L.). *EST Generation, Unigene Analysis, Probe Selection and Bioinformatics for a 16,000-Locus Bin-Delineated Map,* 168**,** 585-593.

LEE, C., IAFRATE, A. J. & BROTHMAN, A. R. 2007. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat Genet*.

LEGGETT, R. M., CLAVIJO, B. J., CLISSOLD, L., CLARK, M. D. & CACCAMO, M. 2014. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics,* 30**,** 566-568.

LELLEY, T., STACHEL, M., GRAUSGRUBER, H. & VOLLMANN, J. 2000. Analysis of relationships between Aegilops tauschii and the D genome of wheat utilizing microsatellites. *Genome,* 43**,** 661-668.

LEV-YADUN, S., GOPHER, A. & ABBO, S. 2000. The Cradle of Agriculture. *Science,* 288**,** 1602-1603.

LEVENE, M. J., KORLACH, J., TURNER, S. W., FOQUET, M., CRAIGHEAD, H. G. & WEBB, W. W. 2003. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. *Science,* 299**,** 682-686.

LEVY, S., SUTTON, G., NG, P. C., FEUK, L., HALPERN, A. L., WALENZ, B. P., AXELROD, N., HUANG, J., KIRKNESS, E. F., DENISOV, G., LIN, Y., MACDONALD, J. R., PANG, A. W., SHAGO, M., STOCKWELL, T. B., TSIAMOURI, A., BAFNA, V., BANSAL, V., KRAVITZ, S. A., BUSAM, D. A., BEESON, K. Y., MCINTOSH, T. C., REMINGTON, K. A., ABRIL, J. F., GILL, J., BORMAN, J., ROGERS, Y. H., FRAZIER, M. E., SCHERER, S. W. &

STRAUSBERG, R. L. 2007. The diploid genome sequence of an individual human. *PLoS Biol,* 5.

LI, A.-L., GENG, S.-F., ZHANG, L.-Q., LIU, D.-C. & MAO, L. 2015a. Making the Bread: Insights from Newly Synthesized Allohexaploid Wheat. *Molecular Plant,* 8**,** 847-859.

LI, C., BAI, G., CHAO, S. & WANG, Z. 2015b. A High-Density SNP and SSR Consensus Map Reveals Segregation Distortion Regions in Wheat. *BioMed Research International,* 2015**,** 830618.

LI, H. 2011. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics,* 27.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009a. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & GENOME PROJECT DATA PROCESSING, S. 2009b. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LI, H., VIKRAM, P., SINGH, R. P., KILIAN, A., CARLING, J., SONG, J., BURGUENO-FERREIRA, J. A., BHAVANI, S., HUERTA-ESPINO, J., PAYNE, T., SEHGAL, D., WENZL, P. & SINGH, S. 2015c. A high density GBS map of bread wheat and its application for dissecting complex disease resistance traits. *BMC Genomics,* 16**,** 216.

LI, J. Y., WANG, J. & ZEIGLER, R. S. 2014a. The 3,000 rice genomes project: new opportunities and challenges for future rice research. *Gigascience,* 3.

LI, J. Y., WANG, J. & ZEIGLER, R. S. 2014b. The 3000 Rice Genome Project: opportunities and challenges for future rice research. *GigaScience,* 3.

LI, L., STOECKERT, C. J., JR. & ROOS, D. S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res,* 13**,** 2178-89.

LI, R., YU, C., LI, Y., LAM, T.-W., YIU, S.-M., KRISTIANSEN, K. & WANG, J. 2009c. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics,* 25**,** 1966-1967.

LI, R., ZHU, H., RUAN, J., QIAN, W., FANG, X., SHI, Z., LI, Y., LI, S., SHAN, G., KRISTIANSEN, K., LI, S., YANG, H., WANG, J. & WANG, J. 2010. De novo

assembly of human genomes with massively parallel short read sequencing. *Genome Research,* 20**,** 265-272.

LI, S., JIA, J., WEI, X., ZHANG, X., LI, L., CHEN, H., FAN, Y., SUN, H., ZHAO, X., LEI, T., XU, Y., JIANG, F., WANG, H. & LI, L. 2007. A intervarietal genetic map and QTL analysis for yield traits in wheat. *Molecular Breeding,* 20**,** 167-178.

LI, W., ZHANG, P., FELLERS, J. P., FRIEBE, B. & GILL, B. S. 2004. Sequence composition, organization, and evolution of the core Triticeae genome. *The Plant Journal,* 40**,** 500-511.

LI, Y.-H., ZHOU, G., MA, J., JIANG, W., JIN, L.-G., ZHANG, Z., GUO, Y., ZHANG, J., SUI, Y., ZHENG, L., ZHANG, S.-S., ZUO, Q., SHI, X.-H., LI, Y.-F., ZHANG, W.-K., HU, Y., KONG, G., HONG, H.-L., TAN, B., SONG, J., LIU, Z.-X., WANG, Y., RUAN, H., YEUNG, C. K. L., LIU, J., WANG, H., ZHANG, L.-J., GUAN, R.-X., WANG, K.-J., LI, W.-B., CHEN, S.-Y., CHANG, R.-Z., JIANG, Z., JACKSON, S. A., LI, R. & QIU, L.-J. 2014c. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotech,* 32**,** 1045-1052.

LI, Y. H., ZHOU, G. Y., MA, J. X., JIANG, W. K., JIN, L. G., ZHANG, Z. H., GUO, Y., ZHANG, J. B., SUI, Y., ZHENG, L. T., ZHANG, S. S., ZUO, Q. Y., SHI, X. H., LI, Y. F., ZHANG, W. K., HU, Y. Y., KONG, G. Y., HONG, H. L., TAN, B., SONG, J., LIU, Z. X., WANG, Y. S., RUAN, H., YEUNG, C. K. L., LIU, J., WANG, H. L., ZHANG, L. J., GUAN, R. X., WANG, K. J., LI, W. B., CHEN, S. Y., CHANG, R. Z., JIANG, Z., JACKSON, S. A., LI, R. Q. & QIU, L. J. 2014d. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology,* 32**,** 1045-+.

LIAO, H.-M., CHAO, Y.-L., HUANG, A.-L., CHENG, M.-C., CHEN, Y.-J., LEE, K.-F., FANG, J.-S., HSU, C.-H. & CHEN, C.-H. 2012. Identification and characterization of three inherited genomic copy number variations associated with familial schizophrenia. *Schizophrenia Research,* 139**,** 229-236.

LIAO, Y.-C., LIN, S.-H. & LIN, H.-H. 2015. Completing bacterial genome assemblies: strategy and performance comparisons. 5**,** 8747.

LING, H.-Q., ZHAO, S., LIU, D., WANG, J., SUN, H., ZHANG, C., FAN, H., LI, D., DONG, L., TAO, Y., GAO, C., WU, H., LI, Y., CUI, Y., GUO, X., ZHENG, S., WANG, B., YU, K., LIANG, Q., YANG, W., LOU, X., CHEN, J., FENG, M.,

JIAN, J., ZHANG, X., LUO, G., JIANG, Y., LIU, J., WANG, Z., SHA, Y., ZHANG, B., WU, H., TANG, D., SHEN, Q., XUE, P., ZOU, S., WANG, X., LIU, X., WANG, F., YANG, Y., AN, X., DONG, Z., ZHANG, K., ZHANG, X., LUO, M.-C., DVORAK, J., TONG, Y., WANG, J., YANG, H., LI, Z., WANG, D., ZHANG, A. & WANG, J. 2013. Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature,* 496**,** 87-90.

LIU, M., STILLER, J., HOLUŠOVÁ, K., VRÁNA, J., LIU, D., DOLEŽEL, J. & LIU, C. 2016a. Chromosome-specific sequencing reveals an extensive dispensable genome component in wheat. *Scientific Reports,* 6**,** 36398.

LIU, M., STILLER, J., HOLUŠOVÁ, K., VRÁNA, J., LIU, D., DOLEŽEL, J. & LIU, C. 2016b. Chromosome-specific sequencing reveals an extensive dispensable genome component in wheat. 6**,** 36398.

LOBELL, D. B., SCHLENKER, W. & COSTA-ROBERTS, J. 2011. Climate Trends and Global Crop Production Since 1980. *Science,* 333**,** 616-620.

LOMAN, N. J., QUICK, J. & SIMPSON, J. T. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Meth,* 12**,** 733-735.

LONGIN, C. F. H. & REIF, J. C. 2014. Redesigning the exploitation of wheat genetic resources. *Trends in Plant Science,* 19**,** 631-636.

LOPES, M. S., EL-BASYONI, I., BAENZIGER, P. S., SINGH, S., ROYO, C., OZBEK, K., AKTAS, H., OZER, E., OZDEMIR, F., MANICKAVELU, A., BAN, T. & VIKRAM, P. 2015. Exploiting genetic diversity from landraces in wheat breeding for adaptation to climate change. *Journal of Experimental Botany,* 66**,** 3477-3486.

LORENC, M. T., HAYASHI, S., STILLER, J., LEE, H., MANOLI, S., RUPERAO, P., VISENDI, P., BERKMAN, P. J., LAI, K., BATLEY, J. & EDWARDS, D. 2012. Discovery of Single Nucleotide Polymorphisms in Complex Genomes Using SGSautoSNP. *Biology,* 1**,** 370-382.

LOVE, R. R., WEISENFELD, N. I., JAFFE, D. B., BESANSKY, N. J. & NEAFSEY, D. E. 2016. Evaluation of DISCOVAR de novo using a mosquito sample for cost-effective short-read genome assembly. *BMC Genomics,* 17**,** 187.

LU, F., ROMAY, M. C., GLAUBITZ, J. C., BRADBURY, P. J., ELSHIRE, R. J. & WANG, T. 2015. High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat Commun.,* 6.

LU, J., TANG, T., TANG, H., HUANG, J., SHI, S. & WU, C.-I. 2006. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends in Genetics,* 22**,** 126-131.

LU, P., HAN, X., QI, J., YANG, J., WIJERATNE, A. J., LI, T. & MA, H. 2012. Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. *Genome Research,* 22**,** 508-518.

LUKENS, L. N., PIRES, J. C., LEON, E., VOGELZANG, R., OSLACH, L. & OSBORN, T. 2006a. Patterns of Sequence Loss and Cytosine Methylation within a Population of Newly Resynthesized Brassica napus Allopolyploids. *Plant Physiology,* 140**,** 336-348.

LUKENS, L. N., PIRES, J. C., LEON, E. J., VOGELZANG, R., OSLACH, L. & OSBORN, T. C. 2006b. Patterns of sequence loss and cytosine methylation within a population of newly resynthesized Brassica napusallopolyploids. *Plant Physiology,* 140.

LUO, C., TSEMENTZI, D., KYRPIDES, N., READ, T. & KONSTANTINIDIS, K. T. 2012a. Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. *PLoS ONE,* 7**,** e30087.

LUO, M.-C., GU, Y. Q., YOU, F. M., DEAL, K. R., MA, Y. & HU, Y. 2013. A 4-gigabase physical map unlocks the structure and evolution of the complex genome of Aegilops tauschii, the wheat D-genome progenitor. *Proc Natl Acad Sci U S A,* 110.

LUO, M.-C., YANG, Z.-L., YOU, F. M., KAWAHARA, T., WAINES, J. G. & DVORAK, J. 2007. The structure of wild and domesticated emmer wheat populations, gene flow between them, and the site of emmer domestication. *Theoretical and Applied Genetics,* 114**,** 947-959.

LUO, R., LIU, B., XIE, Y., LI, Z., HUANG, W., YUAN, J., HE, G., CHEN, Y., PAN, Q., LIU, Y., TANG, J., WU, G., ZHANG, H., SHI, Y., LIU, Y., YU, C., WANG, B., LU, Y., HAN, C., CHEUNG, D., YIU, S.-M., PENG, S., XIAOQIAN, Z., LIU, G., LIAO, X., LI, Y., YANG, H., WANG, J., LAM, T.-W. & WANG, J. 2012b. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience,* 1.

MARCAIS, G. & KINGSFORD, C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics,* 27**,** 764-70.

MARÇAIS, G., YORKE, J. A. & ZIMIN, A. 2015. QuorUM: An Error Corrector for Illumina Reads. *PLOS ONE,* 10**,** e0130821.

MARCUS, S., LEE, H. & SCHATZ, M. C. 2014. SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips. *Bioinformatics,* 30**,** 3476-3483.

MARCUSSEN, T., SANDVE, S. R., HEIER, L., SPANNAGL, M., PFEIFER, M. & JAKOBSEN, K. S. 2014. Ancient hybridizations among the ancestral genomes of bread wheat. *Science,* 345.

MARIETTE, J., NOIROT, C. & KLOPP, C. 2011. Assessment of replicate bias in 454 pyrosequencing and a multi-purpose read-filtering tool. *BMC Research Notes,* 4**,** 149.

MARTIN, J. M., TALBERT, L. E., LANNING, S. P. & BLAKE, N. K. 1995. Hybrid Performance in Wheat as Related to Parental Diversity. *Crop Science,* 35**,** 104-108.

MARTINEZ-PEREZ, E., SHAW, P. & MOORE, G. 2001. The Ph1 locus is needed to ensure specific somatic and meiotic centromere association. *Nature,* 411**,** 204-207.

MASCHER, M., MUEHLBAUER, G. J., ROKHSAR, D. S., CHAPMAN, J., SCHMUTZ, J., BARRY, K., MUNOZ-AMATRIAIN, M., CLOSE, T. J., WISE, R. P., SCHULMAN, A. H., HIMMELBACH, A., MAYER, K. F., SCHOLZ, U., POLAND, J. A., STEIN, N. & WAUGH, R. 2013. Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *Plant J,* 76**,** 718-27.

MAYER, K. F., ROGERS, J., DOLEŽEL, J., POZNIAK, C., EVERSOLE, K. & FEUILLET, C. 2014a. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science.,* 345.

MAYER, K. F., WAUGH, R., BROWN, J. W., SCHULMAN, A. & LANGRIDGE, P. 2012. A physical, genetic and functional sequence assembly of the barley genome. *Nature,* 491.

MAYER, K. F. X., MARTIS, M., HEDLEY, P. E., ŠIMKOVÁ, H., LIU, H., MORRIS, J. A., STEUERNAGEL, B., TAUDIEN, S., ROESSNER, S., GUNDLACH, H., KUBALÁKOVÁ, M., SUCHÁNKOVÁ, P., MURAT, F., FELDER, M., NUSSBAUMER, T., GRANER, A., SALSE, J., ENDO, T., SAKAI, H., TANAKA, T., ITOH, T., SATO, K., PLATZER, M., MATSUMOTO, T.,

SCHOLZ, U., DOLEŽEL, J., WAUGH, R. & STEIN, N. 2011. Unlocking the Barley Genome by Chromosomal and Comparative Genomics. *The Plant Cell,* 23**,** 1249-1263.

MAYER, K. F. X., ROGERS, J., POZNIAK, C., EVERSOLE, K., FEUILLET, C. & GILL, B. 2014b. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science,* 345.

MAYER, K. F. X., TAUDIEN, S., MARTIS, M., SIMKOVÁ, H., SUCHÁNKOVÁ, P. & GUNDLACH, H. 2009. Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol,* 151.

MCCARROLL, S. A. & ALTSHULER, D. M. 2007. Copy-number variation and association studies of human disease. *Nat Genet*.

MCCARTHY, A. 2010. Third Generation DNA Sequencing: Pacific Biosciences' Single Molecule Real Time Technology. *Chemistry & Biology,* 17**,** 675-676.

MCFADDEN, E. S. & SEARS, E. R. 1946. The origin of Triticum spelta and its free-threshing hexaploid relatives. *J Hered,* 37.

MEDINI, D., DONATI, C., TETTELIN, H., MASIGNANI, V. & RAPPUOLI, R. 2005. The microbial pan-genome. *Current Opinion in Genetics & Development,* 15**,** 589-594.

MEISSNER, A., GNIRKE, A., BELL, G. W., RAMSAHOYE, B., LANDER, E. S. & JAENISCH, R. 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research,* 33**,** 5868-5877.

MENGISTU, D. K., KIDANE, Y. G., CATELLANI, M., FRASCAROLI, E., FADDA, C., PÈ, M. E. & DELL'ACQUA, M. 2016. High-density molecular characterization and association mapping in Ethiopian durum wheat landraces reveals high diversity and potential for wheat breeding. *Plant Biotechnology Journal,* 14**,** 1800-1812.

METTE, M. F., GILS, M., LONGIN, C. F. H. & REIF, J. C. 2015. Hybrid Breeding in Wheat. *In:* OGIHARA, Y., TAKUMI, S. & HANDA, H. (eds.) *Advances in Wheat Genetics: From Genome to Field: Proceedings of the 12th International Wheat Genetics Symposium.* Tokyo: Springer Japan.

MEYER, K. D. & JAFFREY, S. R. 2016. Expanding the diversity of DNA base modifications with N 6-methyldeoxyadenosine. *Genome Biology,* 17**,** 5.

MICHAEL, T. P. & JACKSON, S. 2013. The First 50 Plant Genomes. *The Plant Genome,* 6.

MIDDLETON, C. P., SENERCHIA, N., STEIN, N., AKHUNOV, E. D., KELLER, B., WICKER, T. & KILIAN, B. 2014. Sequencing of chloroplast genomes from wheat, barley, rye and their relatives provides a detailed insight into the evolution of the Triticeae tribe. *PLoS One,* 9**,** e85761.

MIKHEYEV, A. S. & TIN, M. M. Y. 2014. A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources,* 14**,** 1097-1102.

MOCHIDA, K., SAKURAI, T., SEKI, H., YOSHIDA, T., TAKAHAGI, K., SAWAI, S., UCHIYAMA, H., MURANAKA, T. & SAITO, K. 2017. Draft genome assembly and annotation of Glycyrrhiza uralensis, a medicinal legume. *The Plant Journal,* 89**,** 181-194.

MOORE, R. C. & PURUGGANAN, M. D. 2005. The evolutionary dynamics of plant duplicate genes. *Current Opinion in Plant Biology,* 8**,** 122-128.

MOVAHEDI, N. S., EMBREE, M., NAGARAJAN, H., ZENGLER, K. & CHITSAZ, H. 2016. Efficient Synergistic Single-Cell Genome Assembly. *Frontiers in Bioengineering and Biotechnology,* 4.

MUHINDIRA, P. V. 2016. A novel approach for the assembly of complex genomic DNA cloned into bacterial artificial chromosome vectors: assembly and analysis of Triticum aestivum chromosome arm 7DS. The University of Queensland, School of Agriculture and Food Sciences.

MUJEEB-KAZI, A., GUL, A., FAROOQ, M., RIZWAN, S. & AHMAD, I. 2008. Rebirth of synthetic hexaploids with global implications for wheat improvement. *Australian Journal of Agricultural Research,* 59**,** 391-398.

MURAT, F., XU, J.-H., TANNIER, E., ABROUK, M., GUILHOT, N., PONT, C., MESSING, J. & SALSE, J. 2010. Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Research,* 20**,** 1545-1557.

MYERS, E. W., SUTTON, G. G., DELCHER, A. L., DEW, I. M., FASULO, D. P. & FLANIGAN, M. J. 2000. A whole-genome assembly of Drosophila. *Science,* 287.

NEALE, D. B., WEGRZYN, J. L., STEVENS, K. A., ZIMIN, A. V., PUIU, D., CREPEAU, M. W., CARDENO, C., KORIABINE, M., HOLTZ-MORRIS, A. E., LIECHTY, J. D., MARTÍNEZ-GARCÍA, P. J., VASQUEZ-GROSS, H. A., LIN,

B. Y., ZIEVE, J. J., DOUGHERTY, W. M., FUENTES-SORIANO, S., WU, L.-S., GILBERT, D., MARÇAIS, G., ROBERTS, M., HOLT, C., YANDELL, M., DAVIS, J. M., SMITH, K. E., DEAN, J. F., LORENZ, W. W., WHETTEN, R. W., SEDEROFF, R., WHEELER, N., MCGUIRE, P. E., MAIN, D., LOOPSTRA, C. A., MOCKAITIS, K., DEJONG, P. J., YORKE, J. A., SALZBERG, S. L. & LANGLEY, C. H. 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology,* 15**,** R59.

NEPH, S., KUEHN, M. S., REYNOLDS, A. P., HAUGEN, E., THURMAN, R. E., JOHNSON, A. K., RYNES, E., MAURANO, M. T., VIERSTRA, J., THOMAS, S., SANDSTROM, R., HUMBERT, R. & STAMATOYANNOPOULOS, J. A. 2012. BEDOPS: high-performance genomic feature operations. *Bioinformatics,* 28**,** 1919-1920.

NEWELL, M. A. & JANNINK, J.-L. 2014. Genomic Selection in Plant Breeding. *In:* FLEURY, D. & WHITFORD, R. (eds.) *Crop Breeding: Methods and Protocols.* New York, NY: Springer New York.

NRGENE. 2016. *DeNovoMAGIC2.0* [Online]. Available: http://nrgene.com/products-technology/denovomagic/ [Accessed 11-11 2016].

NRGENE. 2017. *DeNovoMagic* [Online]. Available: http://nrgene.com/ [Accessed 2017 2017].

NURK, S. M., DMITRY; KOROBEYNIKOV, ANTON; PEVZNER, PAVEL 2016. metaSPAdes: a new versatile de novo metagenomics assembler. *bioarXiv*.

O'MARA, J. G. 1951. Cytogenetic Studies on Triticale II

The kinds of intergeneric chromosome addition. *CYTOLOGIA,* 16**,** 225-232.

O'MARA, J. G. 1953. The cytogenetics of Triticale. *The Botanical Review,* 19**,** 587-605.

OGIHARA, Y. & TSUNEWAKI, K. 1988a. Diversity and evolution of chloroplast DNA in Triticum and Aegilops as revealed by restriction fragment analysis. *Theoretical and Applied Genetics,* 76**,** 321-332.

OGIHARA, Y. & TSUNEWAKI, K. 1988b. Diversity and evolution of chloroplast DNA in Triticum and Aegilops as revealed by restriction fragment analysis. *Theor Appl Genet,* 76**,** 321-32.

OSSOWSKI, S., SCHNEEBERGER, K., CLARK, R. M., LANZ, C., WARTHMANN, N. & WEIGEL, D. 2008. Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res,* 18**,** 2024-33.

OZKAN, H., BRANDOLINI, A., SCHAFER-PREGL, R. & SALAMINI, F. 2002. AFLP analysis of a collection of tetraploid wheats indicates the origin of emmer and hard wheat domestication in southeast Turkey. *Mol Biol Evol,* 19**,** 1797-801.

PANKRATZ, N., DUMITRIU, A., HETRICK, K. N., SUN, M., LATOURELLE, J. C., WILK, J. B., HALTER, C., DOHENY, K. F., GUSELLA, J. F., NICHOLS, W. C., MYERS, R. H., FOROUD, T., DESTEFANO, A. L., THE, P. P., GENEPD INVESTIGATORS, C. & MOLECULAR GENETIC, L. 2011. Copy Number Variation in Familial Parkinson Disease. *PLOS ONE,* 6**,** e20988.

PARKIN, I. A., KOH, C., TANG, H., ROBINSON, S. J., KAGALE, S., CLARKE, W. E., TOWN, C. D., NIXON, J., KRISHNAKUMAR, V., BIDWELL, S. L., DENOEUD, F., BELCRAM, H., LINKS, M. G., JUST, J., CLARKE, C., BENDER, T., HUEBERT, T., MASON, A. S., PIRES, C. J., BARKER, G., MOORE, J., WALLEY, P. G., MANOLI, S., BATLEY, J., EDWARDS, D., NELSON, M. N., WANG, X., PATERSON, A. H., KING, G., BANCROFT, I., CHALHOUB, B. & SHARPE, A. G. 2014. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. *Genome Biol,* 15**,** R77.

PARRA, G., BRADNAM, K. & KORF, I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics,* 23**,** 1061-7.

PARRA, G., BRADNAM, K., NING, Z., KEANE, T. & KORF, I. 2009. Assessing the gene space in draft genomes. *Nucleic Acids Research,* 37**,** 289-297.

PATERSON, A. H., BOWERS, J. E., BRUGGMANN, R., DUBCHAK, I., GRIMWOOD, J., GUNDLACH, H., HABERER, G., HELLSTEN, U., MITROS, T., POLIAKOV, A., SCHMUTZ, J., SPANNAGL, M., TANG, H., WANG, X., WICKER, T., BHARTI, A. K., CHAPMAN, J., FELTUS, F. A., GOWIK, U., GRIGORIEV, I. V., LYONS, E., MAHER, C. A., MARTIS, M., NARECHANIA, A., OTILLAR, R. P., PENNING, B. W., SALAMOV, A. A., WANG, Y., ZHANG, L., CARPITA, N. C., FREELING, M., GINGLE, A. R., HASH, C. T., KELLER, B., KLEIN, P., KRESOVICH, S., MCCANN, M. C., MING, R., PETERSON, D. G., MEHBOOB UR, R., WARE, D., WESTHOFF, P., MAYER, K. F. X.,

MESSING, J. & ROKHSAR, D. S. 2009. The Sorghum bicolor genome and the diversification of grasses. *Nature,* 457**,** 551-556.

PATERSON AH, W. X., LI J, TANG H. 2012. Ancient and recent polyploidy in monocots. *In:* SOLTIS P, S. D. (ed.) *Polyploidy and genome evolution.* Berlin: Springer.

PAUX, E., ROGER, D., BADAEVA, E., GAY, G., BERNARD, M., SOURDILLE, P. & FEUILLET, C. 2006. Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. *Plant J,* 48**,** 463-74.

PAUX, E., SOURDILLE, P., SALSE, J., SAINTENAC, C., CHOULET, F., LEROY, P., KOROL, A., MICHALAK, M., KIANIAN, S., SPIELMEYER, W., LAGUDAH, E., SOMERS, D., KILIAN, A., ALAUX, M., VAUTRIN, S., BERGÈS, H., EVERSOLE, K., APPELS, R., SAFAR, J., SIMKOVA, H., DOLEZEL, J., BERNARD, M. & FEUILLET, C. 2008. A Physical Map of the 1-Gigabase Bread Wheat Chromosome 3B. *Science,* 322**,** 101-104.

PENG, J., SUN, D. & NEVO, E. 2011. Domestication evolution, genetics and genomics in wheat. *Molecular Breeding,* 28**,** 281-301.

PENG, J., ZADEH, H., LAZO, G., GUSTAFSON, J., CHAO, S., ANDERSON, O., QI, L., ECHALIER, B., GILL, B. & DILBIRLIGI, M. 2004. Chromosome bin map of expressed sequence tags in homoeologous group 1 of hexaploid wheat and homoeology with rice and Arabidopsis. *Genetics,* 168.

PENG, Y., LEUNG, H. C., YIU, S. M. & CHIN, F. Y. 2012a. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics,* 28**,** 1420-8.

PENG, Y., LEUNG, H. C. M., YIU, S. M. & CHIN, F. Y. L. 2010. IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler. *In:* BERGER, B. (ed.) *Research in Computational Molecular Biology: 14th Annual International Conference, RECOMB 2010, Lisbon, Portugal, April 25-28, 2010. Proceedings.* Berlin, Heidelberg: Springer Berlin Heidelberg.

PENG, Y., LEUNG, H. C. M., YIU, S. M. & CHIN, F. Y. L. 2012b. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics,* 28**,** 1420-1428.

PESTSOVA, E. G., BORNER, A. & RODER, M. S. 2001. Development of a set of Triticum aestivum-Aegilops tauschii introgression lines. *Hereditas,* 135**,** 139-43.

PETERSON, D. G., SCHULZE, S. R., SCIARA, E. B., LEE, S. A., BOWERS, J. E., NAGEL, A., JIANG, N., TIBBITTS, D. C., WESSLER, S. R. & PATERSON, A. H. 2002. Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Res,* 12**,** 795-807.

PEVZNER, P. A., TANG, H. & WATERMAN, M. S. 2001. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci USA,* 98.

PFEIFER, M., KUGLER, K. G., SANDVE, S. R., ZHAN, B., RUDI, H. & HVIDSTEN, T. R. 2014. Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science.,* 345.

PFEIFER, M., MARTIS, M., ASP, T., MAYER, K. F., LUBBERSTEDT, T., BYRNE, S., FREI, U. & STUDER, B. 2013. The perennial ryegrass GenomeZipper: targeted use of genome resources for comparative grass genomics. *Plant Physiol,* 161**,** 571-82.

PHANSTIEL, D. H., BOYLE, A. P., ARAYA, C. L. & SNYDER, M. P. 2014. Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics,* 30**,** 2808-10.

PHILLIPPY, A. M. 2017. New advances in sequence assembly. *Genome Research,* 27**,** xi-xiii.

PLASCHKE, J., GANAL, M. W. & RODER, M. S. 1995a. Detection of genetics diversity in closely-related bread wheat using microsatellite markers. *Theoretical and Applied Genetics,* 91**,** 1001-1007.

POLAND, J., ENDELMAN, J., DAWSON, J., RUTKOSKI, J., WU, S., MANES, Y., DREISIGACKER, S., CROSSA, J., SÁNCHEZ-VILLEDA, H., SORRELLS, M. & JANNINK, J.-L. 2012a. Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *The Plant Genome,* 5**,** 103-113.

POLAND, J., ENDELMAN, J., DAWSON, J., RUTKOSKI, J., WU, S., MANES, Y., DREISIGACKER, S., CROSSA, J., SÁNCHEZ-VILLEDA, H., SORRELLS, M. & JANNINK, J.-L. 2012b. Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Gen.,* 5**,** 103-113.

POLAND, J. A., BROWN, P. J., SORRELLS, M. E. & JANNINK, J.-L. 2012c. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLoS ONE,* 7**,** e32253.

PONT, C., MURAT, F., GUIZARD, S., FLORES, R., FOUCRIER, S., BIDET, Y., QURAISHI, U. M., ALAUX, M., DOLEŽEL, J., FAHIMA, T., BUDAK, H., KELLER, B., SALVI, S., MACCAFERRI, M., STEINBACH, D., FEUILLET, C., QUESNEVILLE, H. & SALSE, J. 2013. Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. *The Plant Journal,* 76**,** 1030-1044.

POP, M., KOSACK, D. S. & SALZBERG, S. L. 2004. Hierarchical Scaffolding With Bambus. *Genome Research,* 14**,** 149-159.

PRASAD, M., VARSHNEY, R. K., ROY, J. K., BALYAN, H. S. & GUPTA, P. K. 2000. The use of microsatellites for detecting DNA polymorphism, genotype identification and genetic diversity in wheat. *Theoretical and Applied Genetics,* 100**,** 584-592.

PRINCE, V. E. & PICKETT, F. B. 2002. Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet,* 3**,** 827-837.

QI, L. L., ECHALIER, B., CHAO, S., LAZO, G. R., BUTLER, G. E., ANDERSON, O. D., AKHUNOV, E. D., DVOŘÁK, J., LINKIEWICZ, A. M., RATNASIRI, A., DUBCOVSKY, J., BERMUDEZ-KANDIANIS, C. E., GREENE, R. A., KANTETY, R., LA ROTA, C. M., MUNKVOLD, J. D., SORRELLS, S. F., SORRELLS, M. E., DILBIRLIGI, M., SIDHU, D., ERAYMAN, M., RANDHAWA, H. S., SANDHU, D., BONDAREVA, S. N., GILL, K. S., MAHMOUD, A. A., MA, X.-F., MIFTAHUDIN, GUSTAFSON, J. P., CONLEY, E. J., NDUATI, V., GONZALEZ-HERNANDEZ, J. L., ANDERSON, J. A., PENG, J. H., LAPITAN, N. L. V., HOSSAIN, K. G., KALAVACHARLA, V., KIANIAN, S. F., PATHAN, M. S., ZHANG, D. S., NGUYEN, H. T., CHOI, D.-W., FENTON, R. D., CLOSE, T. J., MCGUIRE, P. E., QUALSET, C. O. & GILL, B. S. 2004. A Chromosome Bin Map of 16,000 Expressed Sequence Tag Loci and Distribution of Genes Among the Three Genomes of Polyploid Wheat. *Genetics,* 168**,** 701-712.

QUAIL, M. A., SMITH, M., COUPLAND, P., OTTO, T. D., HARRIS, S. R., CONNOR, T. R., BERTONI, A., SWERDLOW, H. P. & GU, Y. 2012. A tale of three next

generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics,* 13**,** 341.

QUARRIE, S. A., STEED, A., CALESTANI, C., SEMIKHODSKII, A., LEBRETON, C., CHINOY, C., STEELE, N., PLJEVLJAKUSIĆ, D., WATERMAN, E., WEYEN, J., SCHONDELMAIER, J., HABASH, D. Z., FARMER, P., SAKER, L., CLARKSON, D. T., ABUGALIEVA, A., YESSIMBEKOVA, M., TURUSPEKOV, Y., ABUGALIEVA, S., TUBEROSA, R., SANGUINETI, M.-C., HOLLINGTON, P. A., ARAGUÉS, R., ROYO, A. & DODIG, D. 2005. A high-density genetic map of hexaploid wheat (Triticum aestivum L.) from the cross Chinese Spring × SQ1 and its use to compare QTLs for grain yield across a range of environments. *Theoretical and Applied Genetics,* 110**,** 865-880.

R CORE TEAM 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing.

RABINOWICZ, P. D., SCHUTZ, K., DEDHIA, N., YORDAN, C., PARNELL, L. D., STEIN, L., MCCOMBIE, W. R. & MARTIENSSEN, R. A. 1999. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat Genet,* 23**,** 305-8.

RAMAN, H., DALTON-MORGAN, J., DIFFEY, S., RAMAN, R., ALAMERY, S., EDWARDS, D. & BATLEY, J. 2014. SNP markers-based map construction and genome-wide linkage analysis in Brassica napus. *Plant Biotechnology Journal,* 12**,** 851-860.

RAMSEY, J. & SCHEMSKE, D. W. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu Rev Ecol Syst.,* 29.

RAND, A. C., JAIN, M., EIZENGA, J. M., MUSSELMAN-BROWN, A., OLSEN, H. E., AKESON, M. & PATEN, B. 2017. Mapping DNA methylation with high-throughput nanopore sequencing. *Nat Meth,* 14**,** 411-413.

RASHEED, A., XIA, X., OGBONNAYA, F., MAHMOOD, T., ZHANG, Z. & MUJEEB-KAZI, A. 2014. Genome-wide association for grain morphology in synthetic hexaploid wheats using digital imaging analysis. *BMC Plant Biol,* 14.

RATEL, D., RAVANAT, J. L., CHARLES, M. P., PLATET, N., BREUILLAUD, L. & LUNARDI, J. 2006. Undetectable levels of N6-methyl adenine in mouse DNA: cloning and analysis of PRED28, a gene coding for a putative mammalian DNA adenine methyltransferase. *FEBS Lett,* 580.

RAVEL, C., PRAUD, S., MURIGNEUX, A., CANAGUIER, A., SAPET, F., SAMSON, D., BALFOURIER, F., DUFOUR, P., CHALHOUB, B., BRUNEL, D., BECKERT, M. & CHARMET, G. 2006. Single-nucleotide polymorphism frequency in a set of selected lines of bread wheat (Triticum aestivum L.). *Genome,* 49**,** 1131-1139.

REBETZKE, G. J., CHAPMAN, S. C., MCINTYRE, C. L., RICHARDS, R. A., CONDON, A. G., WATT, M. & VAN HERWAARDEN, A. F. 2009. Grain Yield Improvement in Water-Limited Environments. *Wheat Science and Trade.* Wiley-Blackwell.

REIF, J. C., ZHANG, P., DREISIGACKER, S., WARBURTON, M. L., VAN GINKEL, M., HOISINGTON, D., BOHN, M. & MELCHINGER, A. E. 2005. Wheat genetic diversity trends during domestication and breeding. *Theor Appl Genet,* 110**,** 859-64.

REYNOLDS, M., BONNETT, D., CHAPMAN, S. C., FURBANK, R. T., MANÈS, Y., MATHER, D. E. & PARRY, M. A. J. 2011. Raising yield potential of wheat. I. Overview of a consortium approach and breeding strategies. *Journal of Experimental Botany,* 62**,** 439-452.

RHOADS, A. & AU, K. F. 2015. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics,* 13**,** 278-289.

RIZZON, C., PONGER, L. & GAUT, B. S. 2006. Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice. *PLoS Comput Biol,* 2**,** e115.

ROBBINS, A. M. 2009. *Dwarfing genes in spring wheat: An agronomic comparison of Rht-B1, Rht-D1, and Rht8.* Montana State University-Bozeman, College of Agriculture.

ROBERTS, R. J., CARNEIRO, M. O. & SCHATZ, M. C. 2013. The advantages of SMRT sequencing. *Genome Biol,* 14.

RÖDER, M. S., KORZUN, V., WENDEHAKE, K., PLASCHKE, J., TIXIER, M.-H., LEROY, P. & GANAL, M. W. 1998. A Microsatellite Map of Wheat. *Genetics,* 149**,** 2007-2023.

ROTHBERG, J. M. & LEAMON, J. H. 2008. The development and impact of 454 sequencing. *Nat Biotech,* 26**,** 1117-1124.

ROUGEMONT, J., AMZALLAG, A., ISELI, C., FARINELLI, L., XENARIOS, I. & NAEF, F. 2008. Probabilistic base calling of Solexa sequencing data. *BMC Bioinformatics,* 9**,** 431.

ROULI, L., MERHEJ, V., FOURNIER, P. E. & RAOULT, D. 2015. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections,* 7**,** 72-85.

RUPERAO, P., CHAN, C.-K. K., AZAM, S., KARAFIÁTOVÁ, M., HAYASHI, S., ČÍŽKOVÁ, J., SAXENA, R. K., ŠIMKOVÁ, H., SONG, C., VRÁNA, J., CHITIKINENI, A., VISENDI, P., GAUR, P. M., MILLÁN, T., SINGH, K. B., TARAN, B., WANG, J., BATLEY, J., DOLEŽEL, J., VARSHNEY, R. K. & EDWARDS, D. 2014. A chromosomal genomics approach to assess and validate the desi and kabuli draft chickpea genome assemblies. *Plant Biotechnology Journal,* 12**,** 778-786.

ŠAFÁŘ, J., BARTOŠ, J., JANDA, J., BELLEC, A., KUBALÁKOVÁ, M., VALARIK, M., PATEYRON, S., WEISEROVA, J., TUSKOVA, R. & ČÍHALÍKOVÁ, J. 2004. Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. *Plant J,* 39.

ŠAFÁŘ, J., ŠIMKOVÁ, H., KUBALÁKOVÁ, M., ČÍHALÍKOVÁ, J., SUCHÁNKOVÁ, P., BARTOŠ, J. & DOLEŽEL, J. 2010a. Development of chromosome-specific BAC resources for genomics of bread wheat. *Cytogenet Genome Res,* 129**,** 211-23.

ŠAFÁŘ, J., ŠIMKOVÁ, H., KUBALÁKOVÁ, M., ČÍHALÍKOVÁ, J., SUCHÁNKOVÁ, P., BARTOŠ, J. & DOLEŽEL, J. 2010b. Development of Chromosome-Specific BAC Resources for Genomics of Bread Wheat. *Cytogenetic and Genome Research,* 129**,** 211-223.

SAINTENAC, C., JIANG, D. & AKHUNOV, E. D. 2011. Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Genome Biol.,* 12.

SAKAMURA, T. 1918. Kurze Mitteilung über die Chromosomenzahlen und die Verwandtschaftsverhältnisse der Triticum-Arten. *Botanical Magazine,* 32**,** 151-154.

SALLAM, A. H., ENDELMAN, J. B., JANNINK, J. L. & SMITH, K. P. 2015. Assessing genomic selection prediction accuracy in a dynamic barley breeding population. *Plant Genome,* 8.

SALMELA, L., WALVE, R., RIVALS, E. & UKKONEN, E. 2017. Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics,* 33**,** 799-806.

SALMON, A., AINOUCHE, M. L. & WENDEL, J. F. 2005. Genetic and epigenetic consequences of recent hybridization and polyploidy in Spartina (Poaceae). *Molecular Ecology,* 14**,** 1163-1175.

SALZBERG, S. L. & YORKE, J. A. 2005. Beware of mis-assembled genomes. *Bioinformatics,* 21**,** 4320-4321.

SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences,* 74**,** 5463-5467.

SANTUARI, L., PRADERVAND, S., AMIGUET-VERCHER, A.-M., THOMAS, J., DORCEY, E., HARSHMAN, K., XENARIOS, I., JUENGER, T. E. & HARDTKE, C. S. 2010. Substantial deletion overlap among divergent Arabidopsis genomes revealed by intersection of short reads and tiling arrays. *Genome Biology,* 11**,** R4-R4.

SARKAR, P. & STEBBINS, G. L. 1956. Morphological Evidence Concerning the Origin of the B Genome in Wheat. *American Journal of Botany,* 43**,** 297-304.

SATYA, P., PASWAN, P. K., GHOSH, S., MAJUMDAR, S. & ALI, N. 2016. Confamiliar transferability of simple sequence repeat (SSR) markers from cotton (Gossypium hirsutum L.) and jute (Corchorus olitorius L.) to twenty two Malvaceous species. *3 Biotech,* 6**,** 65.

SAXENA, R. K., EDWARDS, D. & VARSHNEY, R. K. 2014. Structural variations in plant genomes. *Briefings in Functional Genomics,* 13**,** 296-307.

SCHATZ, M. C., MARON, L. G., STEIN, J. C., WENCES, A. H., GURTOWSKI, J., BIGGERS, E., LEE, H., KRAMER, M., ANTONIOU, E., GHIBAN, E., WRIGHT, M. H., CHIA, J.-M., WARE, D., MCCOUCH, S. R. & MCCOMBIE, W. R. 2014. Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. *Genome Biology,* 15**,** 1-16.

SCHIRRMEISTER, B. E., DALQUEN, D. A., ANISIMOVA, M. & BAGHERI, H. C. 2012. Gene copy number variation and its significance in cyanobacterial phylogeny. *BMC Microbiology,* 12**,** 177.

SCHNABLE, J. C., SPRINGER, N. M. & FREELING, M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad Sci U S A.,* 108.

SCHOEN, C., BLOM, J., CLAUS, H., SCHRAMM-GLUCK, A., BRANDT, P., MULLER, T., GOESMANN, A., JOSEPH, B., KONIETZNY, S., KURZAI, O., SCHMITT, C., FRIEDRICH, T., LINKE, B., VOGEL, U. & FROSCH, M. 2008. Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in Neisseria meningitidis. *Proc Natl Acad Sci U S A,* 105**,** 3473-8.

SCOTT, D. & ELY, B. 2015. Comparison of Genome Sequencing Technology and Assembly Methods for the Analysis of a GC-Rich Bacterial Genome. *Current Microbiology,* 70**,** 338-344.

SEARS, E. 1966. Nullisomic-Tetrasomic Combinations in Hexaploid Wheat. *In:* RILEY, R. & LEWIS, K. R. (eds.) *Chromosome manipulations and plant genetics.*

SEARS, E. R. 1969. Wheat cytogenetics. *Annual Review of Genetics,* 3**,** 451-468.

SEARS, E. R. & MILLER, T. E. 1985. THE HISTORY OF CHINESE SPRING WHEAT. *Cereal Research Communications,* 13**,** 261-263.

SEBAT, J., LAKSHMI, B., TROGE, J., ALEXANDER, J., YOUNG, J., LUNDIN, P., MÅNÉR, S., MASSA, H., WALKER, M., CHI, M., NAVIN, N., LUCITO, R., HEALY, J., HICKS, J., YE, K., REINER, A., GILLIAM, T. C., TRASK, B., PATTERSON, N., ZETTERBERG, A. & WIGLER, M. 2004. Large-Scale Copy Number Polymorphism in the Human Genome. *Science,* 305**,** 525-528.

SEEMANN, S. G. A. T. 2012. VelvetOptimiser is a multi-threaded Perl script for automatically optimising the three primary parameter options (K, -exp_cov, -cov_cutoff) for the Velvet de novo sequence assembler. *Victorian Bioinformatics Consortium.*

SEHGAL, S. K., LI, W., RABINOWICZ, P. D., CHAN, A., SIMKOVÁ, H. & DOLEŽEL, J. 2012. Chromosome arm-specific BAC end sequences permit comparative analysis of homoeologous chromosomes and genomes of polyploid wheat. *BMC Plant Biol,* 12.

SEMAGN, K., BJØRNSTAD, Å., SKINNES, H., MARØY, A. G., TARKEGNE, Y. & WILLIAM, M. 2006a. Distribution of DArT, AFLP, and SSR markers in a

genetic linkage map of a doubled-haploid hexaploid wheat population. *Genome,* 49**,** 545-555.

SEQUENCING PROJECTINTERNATIONAL RICE, G. 2005. The map-based sequence of the rice genome. *Nature,* 436**,** 793-800.

SHAKED, H., KASHKUSH, K., OZKAN, H., FELDMAN, M. & LEVY, A. A. 2001. Sequence Elimination and Cytosine Methylation Are Rapid and Reproducible Responses of the Genome to Wide Hybridization and Allopolyploidy in Wheat. *The Plant Cell,* 13**,** 1749-1759.

SHAPIRO, R., SERVIS, R. E. & WELCHER, M. 1970. Reactions of Uracil and Cytosine Derivatives with Sodium Bisulfite. *Journal of the American Chemical Society,* 92**,** 422-424.

SHARP, P. J., KREIS, M., SHEWRY, P. R. & GALE, M. D. 1988. Location of β-amylase sequences in wheat and its relatives. *Theoretical and Applied Genetics.,* 75**,** 286-290.

SHEIKHIZADEH, S., SCHRANZ, M. E., AKDEL, M., DE RIDDER, D. & SMIT, S. 2016. PanTools: representation, storage and exploration of pan-genomic data. *Bioinformatics,* 32**,** i487-i493.

SIEDLER, H., MESSMER, M. M., SCHACHERMAYR, G. M., WINZELER, H., WINZELER, M. & KELLER, B. 1994. Genetic diversity in European wheat and spelt breeding material based on RFLP data. *Theoretical and Applied Genetics,* 88**,** 994-1003.

SIMÃO, F. A., WATERHOUSE, R. M., IOANNIDIS, P., KRIVENTSEVA, E. V. & ZDOBNOV, E. M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*.

SIMEÃO RESENDE, R. M., CASLER, M. D. & VILELA DE RESENDE, M. D. 2014. Genomic Selection in Forage Breeding: Accuracy and Methods. *Crop Science,* 54**,** 143-156.

ŠIMKOVÁ, H., JANDA, J., HŘIBOVÁ, E., ŠAFÁŘ, J., DOLEŽEL, J. 2007. Cot-based cloning and sequencing of the short arm of wheat chromosome 1B *Plant , soil and environment,* 53**,** 437-441.

SIMPSON, J. T., WONG, K., JACKMAN, S. D., SCHEIN, J. E., JONES, S. J. M. & BIROL, I. 2009. ABySS: A parallel assembler for short read sequence data. *Genome Research,* 19**,** 1117-1123.

SIMPSON, J. T., WORKMAN, R. E., ZUZARTE, P. C., DAVID, M., DURSI, L. J. & TIMP, W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Meth,* 14**,** 407-410.

SINGH, R., SHEORAN, S., SHARMA, P. & CHATRATH, R. 2011. Analysis of simple sequence repeats (SSRs)dynamics in fungus Fusarium graminearum. *Bioinformation,* 5**,** 402-404.

SMET, R., ADAMS, K. L., VANDEPOELE, K., MONTAGU, M. C., MAERE, S. & PEER, Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci U S A.,* 110.

SMIT, A., HUBLEY, R. & GREEN, P. 2015. RepeatMasker Open-4.0. *http://www.repeatmasker.org*.

SMIT, A., HUBLEY, R & GREEN, P. 2013-2015. *Repeat Masker Open-4.0* [Online]. Available: http://www.repeatmasker.org [Accessed].

SMITH, D. B. & FLAVELL, R. B. 1975. Characterisation of the wheat genome by renaturation kinetics. *Chromosoma,* 50**,** 223-242.

SMITH, D. B., RIMPAU, J. & FLAVELL, R. B. 1976. Interspersion of different repeated sequences in the wheat genome revealed by interspecies DNA/DNA hybridisation. *Nucleic Acids Research,* 3**,** 2811-2825.

SMITH, R. B. F. A. D. B. 1976. Nucleotide sequence organisation in the wheat genome. *Heredity,* 37**,** 231-252.

SOLTIS, P. S. & SOLTIS, D. E. 2012. *Polyploidy and genome evolution,* Berlin, Springer.

SOMERS, D. J., ISAAC, P. & EDWARDS, K. 2004a. A high-density microsatellite consensus map for bread wheat (Triticum aestivum L.). *Theoretical and Applied Genetics,* 109**,** 1105-1114.

SOMERS, D. J., ISAAC, P. & EDWARDS, K. 2004b. A high-density microsatellite consensus map for bread wheat (Triticum aestivum L.). *Theor Appl Genet,* 109.

SONAH, H., BASTIEN, M., IQUIRA, E., TARDIVEL, A., LÉGARÉ, G., BOYLE, B., NORMANDEAU, É., LAROCHE, J., LAROSE, S., JEAN, M. & BELZILE, F. 2013. An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. *PLOS ONE,* 8**,** e54603.

SONG, Q. J., SHI, J. R., SINGH, S., FICKUS, E. W., COSTA, J. M., LEWIS, J., GILL, B. S., WARD, R. & CREGAN, P. B. 2005. Development and mapping of microsatellite (SSR) markers in wheat. *Theoretical and Applied Genetics,* 110**,** 550-560.

SORRELLS, M. E., GUSTAFSON, J. P., SOMERS, D., CHAO, S., BENSCHER, D., GUEDIRA-BROWN, G., HUTTNER, E., KILIAN, A., MCGUIRE, P. E., ROSS, K., TANAKA, J., WENZL, P., WILLIAMS, K. & QUALSET, C. O. 2011a. Reconstruction of the Synthetic W7984 × Opata M85 wheat reference population. *Genome,* 54**,** 875-882.

SORRELLS, M. E., GUSTAFSON, J. P., SOMERS, D., CHAO, S., BENSCHER, D., GUEDIRA-BROWN, G., HUTTNER, E., KILIAN, A., MCGUIRE, P. E., ROSS, K., TANAKA, J., WENZL, P., WILLIAMS, K. & QUALSET, C. O. 2011b. Reconstruction of the synthetic W7984 x Opata M85 wheat reference population. *Genome,* 54**,** 875-82.

SOURDILLE, P., SINGH, S., CADALEN, T., BROWN-GUEDIRA, G., GAY, G., QI, L., QI, L. L., DUFOUR, P., MURIGNEUX, A. & BERNARD, M. 2004. Microsatellite-based deletion bin system for the establishment of genetic-physical map relationships in wheat (Triticum aestivum L.). *Funct Integr Genomics,* 4.

SPRINGER, N. M., YING, K., FU, Y., JI, T., YEH, C.-T., JIA, Y., WU, W., RICHMOND, T., KITZMAN, J., ROSENBAUM, H., INIGUEZ, A. L., BARBAZUK, W. B., JEDDELOH, J. A., NETTLETON, D. & SCHNABLE, P. S. 2009a. Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. *PLoS Genet,* 5**,** e1000734.

SPRINGER, N. M., YING, K., FU, Y., JI, T. M., YEH, C. T., JIA, Y., WU, W., RICHMOND, T., KITZMAN, J., ROSENBAUM, H., INIGUEZ, A. L., BARBAZUK, W. B., JEDDELOH, J. A., NETTLETON, D. & SCHNABLE, P. S. 2009b. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *Plos Genetics,* 5.

STAJICH, J. E., BLOCK, D., BOULEZ, K., BRENNER, S. E., CHERVITZ, S. A., DAGDIGIAN, C., FUELLEN, G., GILBERT, J. G., KORF, I. & LAPP, H. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res,* 12.

STAMATAKIS, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics,* 22**,** 2688-2690.

STANKE, M. & MORGENSTERN, B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research,* 33**,** W465-W467.

STEBBINS, G. L. 1947. Types of polyploids: their classification and significance. *Adv Genet.,* 1.

STEPHENSON, P., BRYAN, G., KIRBY, J., COLLINS, A., DEVOS, K., BUSSO, C. & GALE, M. 1998. Fifty new microsatellite loci for the wheat genetic map. *Theoretical and Applied Genetics,* 97**,** 946-949.

STRNADOVA V, B. A., GONZALES J, JEGEKLA S, CHAPMAN J, GILBERT JR, ET AL. 2014. Efficient and accurate clustering for large-scale genetic mapping.

SUN, C., HU, Z., ZHENG, T., LU, K., ZHAO, Y., WANG, W., SHI, J., WANG, C., LU, J., ZHANG, D., LI, Z. & WEI, C. 2017. RPAN: rice pan-genome browser for ~3000 rice genomes. *Nucleic Acids Research,* 45**,** 597-605.

SUZUKI, R. & SHIMODAIRA, H. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics,* 22**,** 1540-1542.

SWANSON-WAGNER, R. A., EICHTEN, S. R., KUMARI, S., TIFFIN, P., STEIN, J. C., WARE, D. & SPRINGER, N. M. 2010. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Research,* 20**,** 1689-1699.

TAE, H., KARUNASENA, E., BAVARVA, J. H., MCIVER, L. J. & GARNER, H. R. 2014. Large scale comparison of non-human sequences in human sequencing data. *Genomics,* 104**,** 453-458.

TAHILIANI, M., KOH, K. P., SHEN, Y., PASTOR, W. A., BANDUKWALA, H. & BRUDNO, Y. 2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science,* 324.

TAN, S., ZHONG, Y., HOU, H., YANG, S. & TIAN, D. 2012. Variation of presence/absence genes among Arabidopsis populations. *BMC Evolutionary Biology,* 12**,** 86.

TANG, N., JIANG, Y., HE, B.-R. & HU, Y.-G. 2009. The Effects of Dwarfing Genes (Rht-B1b, Rht-D1b, and Rht8) with Different Sensitivity to GA3 on the

Coleoptile Length and Plant Height of Wheat. *Agricultural Sciences in China,* 8**,** 1028-1038.

TATARINOVA, T. V., CHEKALIN, E., NIKOLSKY, Y., BRUSKIN, S., CHEBOTAROV, D., MCNALLY, K. L. & ALEXANDROV, N. 2016. Nucleotide diversity analysis highlights functionally important genomic regions. *Scientific Reports,* 6**,** 35730.

TATE, J. A., JOSHI, P., SOLTIS, K. A., SOLTIS, P. S. & SOLTIS, D. E. 2009. On the road to diploidization? Homoeolog loss in independently formed populations of the allopolyploid Tragopogon miscellus (Asteraceae). *BMC Plant Biology,* 9**,** 80.

TAUTZ, D. & RENZ, M. 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Research,* 12**,** 4127-4138.

TAUTZ, D. & SCHLÖTTERER, C. 1994. Simple sequences. *Current Opinion in Genetics & Development,* 4**,** 832-837.

TESTER, M. & LANGRIDGE, P. 2010. Breeding Technologies to Increase Crop Production in a Changing World. *Science,* 327**,** 818-822.

TETTELIN, H., MASIGNANI, V., CIESLEWICZ, M. J., DONATI, C., MEDINI, D., WARD, N. L., ANGIUOLI, S. V., CRABTREE, J., JONES, A. L., DURKIN, A. S., DEBOY, R. T., DAVIDSEN, T. M., MORA, M., SCARSELLI, M., MARGARIT Y ROS, I., PETERSON, J. D., HAUSER, C. R., SUNDARAM, J. P., NELSON, W. C., MADUPU, R., BRINKAC, L. M., DODSON, R. J., ROSOVITZ, M. J., SULLIVAN, S. A., DAUGHERTY, S. C., HAFT, D. H., SELENGUT, J., GWINN, M. L., ZHOU, L., ZAFAR, N., KHOURI, H., RADUNE, D., DIMITROV, G., WATKINS, K., O'CONNOR, K. J. B., SMITH, S., UTTERBACK, T. R., WHITE, O., RUBENS, C. E., GRANDI, G., MADOFF, L. C., KASPER, D. L., TELFORD, J. L., WESSELS, M. R., RAPPUOLI, R. & FRASER, C. M. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pan-genome”. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 13950-13955.

TETTELIN, H., RILEY, D., CATTUTO, C. & MEDINI, D. 2008. Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology,* 11**,** 472-477.

THE GENOMES PROJECT, C. 2015. A global reference for human genetic variation. *Nature,* 526**,** 68-74.

TILMAN, D., BALZER, C., HILL, J. & BEFORT, B. L. 2011. Global food demand and the sustainable intensification of agriculture. *Proc Natl Acad Sci U S A.,* 108.

TISCHLER, G. & MYERS, E. W. 2017. Non Hybrid Long Read Consensus Using Local De Bruijn Graph Assembly. *bioRxiv*.

TORADA, A., KOIKE, M., MOCHIDA, K. & OGIHARA, Y. 2006. SSR-based linkage map with new markers using an intraspecific population of common wheat. *Theoretical and Applied Genetics,* 112**,** 1042-1051.

TRAPNELL, C., PACHTER, L. & SALZBERG, S. L. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics,* 25.

TRAPNELL, C., ROBERTS, A., GOFF, L., PERTEA, G., KIM, D., KELLEY, D. R., PIMENTEL, H., SALZBERG, S. L., RINN, J. L. & PACHTER, L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protocols,* 7**,** 562-578.

TRICK, M., ADAMSKI, N. M., MUGFORD, S. G., JIANG, C. C., FEBRER, M. & UAUY, C. 2012. Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat. *BMC Plant Biol,* 12.

TSUNEWAKI, K. 2009. Plasmon analysis in the Triticum-Aegilops complex. *Breeding Science,* 59**,** 455-470.

UCHIMURA, Y., WYSS, M., BRUGIROUX, S., LIMENITAKIS, J. P., STECHER, B., MCCOY, K. D. & MACPHERSON, A. J. 2016. Complete Genome Sequences of 12 Species of Stable Defined Moderately Diverse Mouse Microbiota 2. *Genome Announcements,* 4.

VALOUEV, A., ICHIKAWA, J., TONTHAT, T., STUART, J., RANADE, S., PECKHAM, H., ZENG, K., MALEK, J. A., COSTA, G., MCKERNAN, K., SIDOW, A., FIRE, A. & JOHNSON, S. M. 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res,* 18**,** 1051-63.

VAN BELKUM, A., SCHERER, S., VAN ALPHEN, L. & VERBRUGH, H. 1998. Short-Sequence DNA Repeats in Prokaryotic Genomes. *Microbiology and Molecular Biology Reviews,* 62**,** 275-293.

VAN HEESCH, S., KLOOSTERMAN, W. P., LANSU, N., RUZIUS, F.-P., LEVANDOWSKY, E., LEE, C. C., ZHOU, S., GOLDSTEIN, S., SCHWARTZ, D. C., HARKINS, T. T., GURYEV, V. & CUPPEN, E. 2013. Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing. *BMC Genomics,* 14**,** 257.

VASER, R., SOVIĆ, I., NAGARAJAN, N. & ŠIKIĆ, M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research,* 27**,** 737-746.

VERNIKOS, G., MEDINI, D., RILEY, D. R. & TETTELIN, H. 2015b. Ten years of pan-genome analyses. *Current Opinion in Microbiology,* 23**,** 148-154.

VILLA, T. C. C., MAXTED, N., SCHOLTEN, M. & FORD-LLOYD, B. 2005. Defining and identifying crop landraces. *Plant genetic resources: characterization and utilization,* 3**,** 373-384.

VISENDI, P., BATLEY, J. & EDWARDS, D. 2013. Next generation characterisation of cereal genomes for marker discovery. *Biology,* 2**,** 1357-1377.

VOS, P., HOGERS, R., BLEEKER, M., REIJANS, M., LEE, T. V. D., HORNES, M., FRITERS, A., POT, J., PALEMAN, J., KUIPER, M. & ZABEAU, M. 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research,* 23**,** 4407-4414.

VYAHHI, N., PYSHKIN, A., PHAM, S. & PEVZNER, P. A. 2012. From de Bruijn Graphs to Rectangle Graphs for Genome Assembly. *In:* RAPHAEL, B. & TANG, J. (eds.) *Algorithms in Bioinformatics: 12th International Workshop, WABI 2012, Ljubljana, Slovenia, September 10-12, 2012. Proceedings.* Berlin, Heidelberg: Springer Berlin Heidelberg.

WANG, F., JIANG, L., CHEN, Y., HAELTERMAN, N. A., BELLEN, H. J. & CHEN, R. 2015. FlyVar: a database for genetic variation in Drosophila melanogaster. *Database: The Journal of Biological Databases and Curation,* 2015**,** bav079.

WANG, J., LUO, M.-C., CHEN, Z., YOU, F. M., WEI, Y., ZHENG, Y. & DVORAK, J. 2013. Aegilops tauschii single nucleotide polymorphisms shed light on the origins of wheat D-genome genetic diversity and pinpoint the geographic origin of hexaploid wheat. *New Phytologist,* 198**,** 925-937.

WANG, S., WONG, D., FORREST, K., ALLEN, A., CHAO, S., HUANG, B. E., MACCAFERRI, M., SALVI, S., MILNER, S. G., CATTIVELLI, L., MASTRANGELO, A. M., WHAN, A., STEPHEN, S., BARKER, G., WIESEKE,

R., PLIESKE, J., INTERNATIONAL WHEAT GENOME SEQUENCING, C., LILLEMO, M., MATHER, D., APPELS, R., DOLFERUS, R., BROWN-GUEDIRA, G., KOROL, A., AKHUNOVA, A. R., FEUILLET, C., SALSE, J., MORGANTE, M., POZNIAK, C., LUO, M.-C., DVORAK, J., MORELL, M., DUBCOVSKY, J., GANAL, M., TUBEROSA, R., LAWLEY, C., MIKOULITCH, I., CAVANAGH, C., EDWARDS, K. J., HAYDEN, M. & AKHUNOV, E. 2014. Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnology Journal,* 12**,** 787-796.

WANG, Y., WANG, X. & PATERSON, A. H. 2012. Genome and gene duplications and gene expression divergence: a view from plants. *Annals of the New York Academy of Sciences,* 1256**,** 1-14.

WANJUGI, H., COLEMAN-DERR, D., HUO, N., KIANIAN, S. F., LUO, M.-C., WU, J., ANDERSON, O. & GU, Y. Q. 2009. Rapid development of PCR-based genome-specific repetitive DNA junction markers in wheat. *Genome,* 52**,** 576-587.

WARBURTON, M. L., CROSSA, J., FRANCO, J., KAZI, M., TRETHOWAN, R., RAJARAM, S., PFEIFFER, W., ZHANG, P., DREISIGACKER, S. & GINKEL, M. V. 2006. Bringing wild relatives back into the family: recovering genetic diversity in CIMMYT improved wheat germplasm. *Euphytica,* 149**,** 289-301.

WEBER, J. L. & WONG, C. 1993. Mutation of human short tandem repeats. *Human Molecular Genetics,* 2**,** 1123-1128.

WEISENFELD, N. I., YIN, S., SHARPE, T., LAU, B., HEGARTY, R., HOLMES, L., SOGOLOFF, B., TABBAA, D., WILLIAMS, L., RUSS, C., NUSBAUM, C., LANDER, E. S., MACCALLUM, I. & JAFFE, D. B. 2014. Comprehensive variation discovery in single human genomes. *Nat Genet,* 46**,** 1350-1355.

WENDEL, J. F. 2000. Genome evolution in polyploids. *Plant Mol Biol.,* 42.

WENDEL, J. F. & DOYLE, J. J. 2005. Polyploidy and evolution in plants. *In:* HENRY, R. J. (ed.) *Plant diversity and evolution.* Wallingford, UK: CABI Publishing.

WENDEL, J. F., JACKSON, S. A., MEYERS, B. C. & WING, R. A. 2016. Evolution of plant genome architecture. *Genome Biology,* 17.

WILLIAMS, R. C. 1989. Restriction fragment length polymorphism (RFLP). *American Journal of Physical Anthropology,* 32**,** 159-184.

WINFIELD, M. O., ALLEN, A. M., BURRIDGE, A. J., BARKER, G. L. A., BENBOW, H. R., WILKINSON, P. A., COGHILL, J., WATERFALL, C., DAVASSI, A., SCOPES, G., PIRANI, A., WEBSTER, T., BREW, F., BLOOR, C., KING, J., WEST, C., GRIFFITHS, S., KING, I., BENTLEY, A. R. & EDWARDS, K. J. 2015. High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnology Journal,* 14**,** 1195-1206.

WINFIELD, M. O., WILKINSON, P. A., ALLEN, A. M., BARKER, G. L. A., COGHILL, J. A. & BURRIDGE, A. 2012. Targeted re-sequencing of the allohexaploid wheat exome. *Plant Biotechnol J,* 10.

WOODHOUSE, M. R., CHENG, F., PIRES, J. C., LISCH, D., FREELING, M. & WANG, X. 2014. Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proc Natl Acad Sci U S A.,* 111.

WU, Q.-H., CHEN, Y.-X., ZHOU, S.-H., FU, L., CHEN, J.-J., XIAO, Y., ZHANG, D., OUYANG, S.-H., ZHAO, X.-J., CUI, Y., ZHANG, D.-Y., LIANG, Y., WANG, Z.-Z., XIE, J.-Z., QIN, J.-X., WANG, G.-X., LI, D.-L., HUANG, Y.-L., YU, M.-H., LU, P., WANG, L.-L., WANG, L., WANG, H., DANG, C., LI, J., ZHANG, Y., PENG, H.-R., YUAN, C.-G., YOU, M.-S., SUN, Q.-X., WANG, J.-R., WANG, L.-X., LUO, M.-C., HAN, J. & LIU, Z.-Y. 2015. High-Density Genetic Linkage Map Construction and QTL Mapping of Grain Shape and Size in the Wheat Population Yanda1817 × Beinong6. *PLOS ONE,* 10**,** e0118144.

WU, Y., BHAT, P. R., CLOSE, T. J. & LONARDI, S. 2008. Efficient and Accurate Construction of Genetic Linkage Maps from the Minimum Spanning Tree of a Graph. *PLoS Genetics,* 4**,** e1000212.

XU, X., LIU, X., GE, S., JENSEN, J. D., HU, F., LI, X., DONG, Y., GUTENKUNST, R. N., FANG, L., HUANG, L., LI, J., HE, W., ZHANG, G., ZHENG, X., ZHANG, F., LI, Y., YU, C., KRISTIANSEN, K., ZHANG, X., WANG, J., WRIGHT, M., MCCOUCH, S., NIELSEN, R., WANG, J. & WANG, W. 2012. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotech,* 30**,** 105-111.

XU, Y. & CROUCH, J. H. 2008. Marker-Assisted Selection in Plant Breeding: From Publications to Practice All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage

and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher. *Crop Science,* 48**,** 391-407.

XU, Y., ZHONG, L., WU, X., FANG, X. & WANG, J. 2009. Rapid alterations of gene expression and cytosine methylation in newly synthesized Brassica napus allopolyploids. *Planta,* 229**,** 471-83.

XUE, S., ZHANG, Z., LIN, F., KONG, Z., CAO, Y., LI, C., YI, H., MEI, M., ZHU, H., WU, J., XU, H., ZHAO, D., TIAN, D., ZHANG, C. & MA, Z. 2008. A high-density intervarietal map of the wheat genome enriched with markers derived from expressed sequence tags. *Theoretical and Applied Genetics,* 117**,** 181-189.

YANDELL, M. & ENCE, D. 2012. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet,* 13.

YAO, W., LI, G., ZHAO, H., WANG, G., LIAN, X. & XIE, W. 2015. Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biology,* 16**,** 1-20.

YOSHIHARA, K., TAJIMA, A., ADACHI, S., QUAN, J., SEKINE, M., KASE, H., YAHATA, T., INOUE, I. & TANAKA, K. 2011. Germline copy number variations in BRCA1-associated ovarian cancer patients. *Genes, Chromosomes and Cancer,* 50**,** 167-177.

YU, J.-K., LA ROTA, M., KANTETY, R. V. & SORRELLS, M. E. 2004. EST derived SSR markers for comparative mapping in wheat and rice. *Molecular Genetics and Genomics,* 271**,** 742-751.

YU, P., WANG, C., XU, Q., FENG, Y., YUAN, X., YU, H., WANG, Y., TANG, S. & WEI, X. 2011. Detection of copy number variations in rice using array-based comparative genomic hybridization. *BMC Genomics,* 12**,** 372.

ZEGEYE, H., RASHEED, A., MAKDIS, F., BADEBO, A. & OGBONNAYA, F. C. 2014. Genome-wide association mapping for seedling and adult plant resistance to stripe rust in synthetic hexaploid wheat. *PLoS One,* 9.

ZERBINO, D. R. & BIRNEY, E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research,* 18**,** 821-829.

ZERBINO, D. R., MCEWEN, G. K., MARGULIES, E. H. & BIRNEY, E. 2009. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PLoS One,* 4**,** e8407.

ZEVEN, A. C. 1998. Landraces: A review of definitions and classifications. *Euphytica,* 104**,** 127-139.

ZHANG, L.-M., LUO, H., LIU, Z.-Q., ZHAO, Y., LUO, J.-C., HAO, D.-Y. & JING, H.-C. 2014. Genome-wide patterns of large-size presence/absence variants in sorghum. *Journal of Integrative Plant Biology,* 56**,** 24-37.

ZHAO, Y., JIA, X., YANG, J., LING, Y., ZHANG, Z., YU, J., WU, J. & XIAO, J. 2014. PanGP: A tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics,* 30**,** 1297-1299.

ZHAO, Y., LI, Z., LIU, G., JIANG, Y., MAURER, H. P., WÜRSCHUM, T., MOCK, H.-P., MATROS, A., EBMEYER, E., SCHACHSCHNEIDER, R., KAZMAN, E., SCHACHT, J., GOWDA, M., LONGIN, C. F. H. & REIF, J. C. 2015. Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. *Proceedings of the National Academy of Sciences,* 112**,** 15624-15629.

ZHENG, X., LEVINE, D., SHEN, J., GOGARTEN, S. M., LAURIE, C. & WEIR, B. S. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics,* 28**,** 3326-3328.

ZIETKIEWICZ, E., RAFALSKI, A. & LABUDA, D. 1994. Genome Fingerprinting by Simple Sequence Repeat (SSR)-Anchored Polymerase Chain Reaction Amplification. *Genomics,* 20**,** 176-183.

ZIMIN, A., STEVENS, K. A., CREPEAU, M. W., HOLTZ-MORRIS, A., KORIABINE, M., MARÇAIS, G., PUIU, D., ROBERTS, M., WEGRZYN, J. L., DE JONG, P. J., NEALE, D. B., SALZBERG, S. L., YORKE, J. A. & LANGLEY, C. H. 2014. Sequencing and Assembly of the 22-Gb Loblolly Pine Genome. *Genetics,* 196**,** 875-890.

ZIMIN, A. V., PUIU, D., LUO, M.-C., ZHU, T., KOREN, S., MARÇAIS, G., YORKE, J. A., DVOŘÁK, J. & SALZBERG, S. L. 2017a. 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**,** 787-792.

ZIMIN, A. V., STEVENS, K. A., CREPEAU, M. W., PUIU, D., WEGRZYN, J. L., YORKE, J. A., LANGLEY, C. H., NEALE, D. B. & SALZBERG, S. L. 2017b. An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *GigaScience,* 6**,** 1-4.

ŻMIEŃKO, A., SAMELAK, A., KOZŁOWSKI, P. & FIGLEROWICZ, M. 2014. Copy number polymorphism in plant genomes. *Theoretical and Applied Genetics,* 127**,** 1-18.