THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

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

Paul Visendi Muhindira

B.Sc. Biochemistry & Chemistry University of Nairobi

M.Sc. Bioinformatics University of Nairobi

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2016

School of Agriculture and Food Sciences

Abstract

Next generation DNA sequencing technologies have led to an exponential growth in the number of genomes being sequenced. While generating whole genome shotgun (WGS) assemblies using next generation sequencing (NGS) is relatively fast and inexpensive, the application of this approach to the assembly of highly repetitive and complex genomes such as wheat results in inferior assemblies thus slowing efforts in identifying markers for crop improvement.

The wheat genome is large, highly repetitive and polyploid. Several approaches have been used to sequence and assemble the wheat genome to variable success. Published approaches such as whole chromosome shotgun (WCS) and whole genome shotgun (WGS) have resulted in draft assemblies that are incomplete, fragmented or only represent a subset of the targeted genomic region. BAC by BAC approaches offer the most accurate assemblies although BAC by BAC approaches are expensive and labour intensive.

This thesis presents the application of a novel BAC sequencing approach which combines indexed pools of BACs, Illumina paired read sequencing, a sequence assembler specifically designed for complex BAC assembly, and a custom bioinformatics pipeline. The approach is demonstrated by sequencing and assembling BAC cloned fragments from bread wheat chromosome arm 7DS. This approach enables the generation of accurate scalable and reproducible assemblies cost effectively compared to traditional BAC by BAC approaches.

Rigorous assembly validation prior to gene annotation and onward analysis is critical in genome sequencing projects but often missing. This thesis demonstrates rigorous assembly validation of bread wheat chromosome arm 7DS BAC assemblies using multiple independent platforms. Novel approaches for *de novo* assembly validation are also presented. The BAC assemblies were successfully validated using BAC end sequences

(BES), mate pair read mappings and orientations, BAC minimum tilling path (MTP), and a chromosome arm 7DS BioNano optical map.

Annotation of chromosome arm 7DS BAC assemblies revealed 5,063 more genes calls when compared to a 7DS chromosome arm shotgun assembly. The chromosome arm 7DS BAC assemblies were less fragmented as reflected in the lengths of scaffolds, isoforms and genes called compared to the 7DS whole chromosome arm shotgun assembly. A depletion in specific core gene families with domains containing amino acid repeats such as coiled-coil proteins and golgins was observed in the 7DS whole chromosome arm shotgun assemblies when compared to the 7DS whole chromosome arm BAC by BAC assembly.

This thesis has demonstrated that accurate genome assemblies of large complex polyploid genomes are possible using Illumina short reads despite repeats posing a challenge. This thesis similarly has demonstrated that the choice of an assembly approach not only affects the contiguity and accuracy of an assembly, but also the number and quality of structural and functional gene annotations and also impacts on the resultant gene families.

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. 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.

2. Helena Staňková, Alex Hastie, Saki Chan, Jan Vrána, Zuzana Tulpová, Marie Kubaláková, **Paul Visendi**, Satomi Hayashi, Mingcheng Luo, Jacqueline Batley, David Edwards, Jaroslav Doležel, Hana Šimková BioNano Genome Mapping of Individual Chromosomes Supports Physical Mapping and Sequence Assembling in Complex Plant Genomes, **2016**. **(Accepted in Plant Biotechnology Journal).**

3. Bayer, P. E.; Ruperao, P.; Mason, A. S.; Stiller, J.; Chan, C.-K. K.; Hayashi, S.; Long, Y.; Meng, J.; Sutton, T.; **Visendi, P**.; Varshney, R. K.; Batley, J.; Edwards, D. High-resolution skim genotyping by sequencing reveals the distribution of crossovers and gene conversions in Cicer arietinum and Brassica napus. Theor. Appl. Genet. **2015**, 128, 1039–1047.

4. Gou, L.; Hattori, J.; Fedak, G.; Balcerzak, M.; Sharpe, A.; **Visendi, P**.; Edwards, D.; Tinker, N.; Wei, Y.-M.; Chen, G.-Y.; Ouellet, T. Development and Validation of –Expressed Molecular Markers Specific for the Long Arm of Chromosome 7E. Crop Science **2015**.

5. 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. Gene loss in the fungal canola pathogen Leptosphaeria maculans. Functional & Integrative Genomics **2014**, 15, 189–196.

6. 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. Identification and characterization of more than 4 million intervarietal SNPs across the group 7 chromosomes of bread wheat. Plant Biotechnol. J. **2014**, 13, 97–104.

7. 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. A chromosomal genomics approach to assess and validate the desi and kabuli draft chickpea genome assemblies. Plant Biotechnol. J. **2014**, 12, 778–786.

8. Visendi, P.; Batley, J.; Edwards, D. Next Generation Characterisation of Cereal Genomes for Marker Discovery. Biology **2013**, *2*, 1357–1377.

9. Berkman, P. J.; **Visendi, P.;** Lee, H. C.; Stiller, J.; Manoli, S.; Lorenc, M. T.; Lai, K.; Batley, J.; Fleury, D.; Šimková, H.; Kubaláková, M.; Weining, S.; Doležel, J.; Edwards, D. Dispersion and domestication shaped the genome of bread wheat. Plant Biotechnology Journal **2013**, *11*, 564–571.

10. Lorenc, M. T.; Hayashi, S.; Stiller, J.; Lee, H.; Manoli, S.; Ruperao, P.; **Visendi, P**.; Berkman, P. J.; Lai, K.; Batley, J.; Edwards, D. Discovery of Single Nucleotide Polymorphisms in Complex Genomes Using SGSautoSNP. Biology **2012,** 1, 370–382.

Book chapters:

1. Visendi, P.; Batley, J.; Edwards, D. Next Generation Sequencing and Germplasm Resources. In *Managing, sequencing and mining genetic resources*; Tuberosa, R.; Graner, A.; Frison, E., Eds.; Genomics of Plant Genetic Resources**, 2014**; Vol. 1, pp. 369–390.

Publications included in this thesis

1. 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. –partially incorporated in chapter 2 and 3.

Contributor	Statement of contribution
Paul Visendi Muhindira **(Candidate)**	Developed assembly method (100%). Coded the assembly pipeline (100%). Did the analysis (100%) Wrote the manuscript (100%).
Paul J Berkman	Conceived and designed sequencing

	experiments (10%). Critiqued the manuscript (20%).
David Edwards	Conceived and designed sequencing experiments (40%). Critiqued the manuscript (40%).
Jacqueline Batley	Conceived and designed sequencing experiments (20%). Critiqued the manuscript (20%).
Hana Šimková	Conceived and designed sequencing experiments (20%). Critiqued the manuscript (20%).
Jaroslav Doležel	Conceived and designed sequencing experiments (10%).
Helena Staňková	Prepared libraries for sequencing (50%)
Satomi Hayashi	Prepared libraries for sequencing (50%)
Agnieszka A. Golicz	Performed data processing and quality checks (20%)
Philipp E. Bayer	Performed data processing and quality checks (20%).
Pradeep Ruperao	Performed data processing and quality checks (20%).
Bhavna Hurgobin	Performed data processing and quality checks (20%).
Juan Montenegro	Performed data processing and quality checks (10%)
Chon-Kit Kenneth Chan	Performed data processing and quality checks (10%).

2. Helena Staňková, Alex Hastie, Saki Chan, Jan Vrána, Zuzana Tulpová, Marie Kubaláková, **Paul Visendi**, Satomi Hayashi, Mingcheng Luo, Jacqueline Batley, David Edwards, Jaroslav Doležel, Hana Šimková BioNano Genome Mapping of Individual Chromosomes Supports Physical Mapping and Sequence Assembling in Complex Plant Genomes. **(Accepted in Plant Biotechnology Journal).** –partially incorporated in chapter 2 and 3.

Contributor	Statement of contribution
Paul Visendi Muhindira **(Candidate).**	Developed the assembly method (100%) Coded the assembly pipeline (100%) Assembled BACs (100%). Wrote manuscript (10%)
Alex Hastie	Generated the 7DS optical map (50%), wrote the manuscript (5 %). Conceived and designed the experiments (10%).
Saki Chan	Generated the 7DS optical map (50%), Wrote the manuscript (5 %). Conceived and designed experiments (10%)
Mingcheng Luo	Conceived and designed experiments

	(10%), Wrote the manuscript (5 %).
Jacqueline Batley	Conceived and designed experiments (10%), Wrote the manuscript (5 %)..
David Edwards	Conceived and designed experiments (10%), Wrote the manuscript (5 %).
Jaroslav Doležel	Conceived and designed experiments (10%), Wrote the manuscript (5 %).
Hana Šimková	Conceived and designed experiments (30%) Wrote the manuscript (30%).
Helena Staňková	Prepared BAC libraries (20%) Wrote the manuscript (30%).
Jan Vrána	Prepared BAC libraries (20%).
Zuzana Tulpová	Prepared BAC libraries (20%).
Marie Kubaláková	Prepared BAC libraries (20%).
Satomi Hayashi	Prepared BAC libraries (20%).

3. Visendi, P.; Batley, J.; Edwards, D. Next Generation Characterisation of Cereal Genomes for Marker Discovery. Biology **2013**, *2*, 1357–1377. –partially incorporated in chapter 1.

Contributor	Statement of contribution
Paul Visendi Muhindira **(Candidate)**	Wrote the manuscript (100%)
Jacqueline Batley	Critiqued and edited manuscript (50%)
David Edwards	Critiqued and edited manuscript (50%)

4. Visendi, P.; Batley, J.; Edwards, D. Next Generation Sequencing and Germplasm Resources. In *Managing, sequencing and mining genetic resources*; Tuberosa, R.; Graner, A.; Frison, E., Eds.; Genomics of Plant Genetic Resources**, 2014**; Vol. 1, pp. 369–390. – partially incorporated in chapter 1.

Contributor	Statement of contribution
Paul Visendi Muhindira **(Candidate)**	Wrote the manuscript (100%)
Jacqueline Batley	Critiqued and edited manuscript (50%).
David Edwards	Critiqued and edited manuscript (50%).

Contributions by others to the thesis

Juan Montenegro contributed other wheat chromosome arm assemblies for SNP and RNA-seq mapping, Paul Berkman provided Sugarcane BACs for analysis, Jaroslav Doležel and Hana Šimková provided BAC clones for sequencing and the 7DS optical map used in this thesis. Dave Edwards and Jacqueline Batley helped with editing of all chapters.

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

None

Acknowledgements

I thank my family Indira and Rachel for their patience, kindness and unwavering support, my mother Emma for her encouragement, the Edwards group and the University of Queensland for the facilities and stimulating academic environment. I would also like to thank my supervisors Dave Edwards, Jacqueline Batley and Jiri Stiller for their supervision, help and support. I would also like to thank Paul Berkman for his continued mentorship. Many thanks to the Australian Department of Foreign Affairs and Trade (DFAT) through Australia Awards for the funding to undertake this PhD.

I finally would like to thank the Son of Man, Jesus for his grace, love and provision during this PhD as Apostle Paul stated;

> “For by him were all things created, that are in heaven, and that are in earth, visible and invisible, whether *they be* thrones, or dominions, or principalities, or powers: all things were created by him, and for him”

and more aptly as King Solomon once said;

> “And further, by these, my son, be admonished: of making many books there is no end; and much study is a weariness of the flesh.”

Keywords

bioinformatics, *triticum aestivum*, chromosome arm 7ds, bac assembly, optical map, genomics, wheat

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060408 Genomics 40%,
ANZSRC code: 060799 Plant Biology not elsewhere classified 30%,
ANZSRC code: 080301 Bioinformatics Software 30 %.

Fields of Research (FoR) Classification

FoR code: 0601, Biochemistry and Cell Biology, 60%
FoR code: 0699, Other Biological Sciences, 20%
FoR code: 0803, Computer software, 20%

This thesis is dedicated to my daughter,
Indira Eve Migarusha Muhindira.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AED: Annotation Edit Distance

ANOVA: Analysis of Variance

BAC: Bacterial Artificial Chromosome

BASH: Bourne Again Shell

BES: BAC End Sequence

BLASR: Basic Local Alignment with Successive Refinement

BLAST: Basic Local Alignment Search Tool

bp: base pair

BWA: Burrows-Wheeler Aligner

CGAL: Computing Genome Assembly Likelihood

CMAP: Consensus Map

CMOS: Complementary Metal-Oxide Semiconductors

DBG: de Bruijn graph

DH: double haploid

DNA: Deoxyribonucleic acid

dNTPs: deoxyribonucleotide triphosphate

dsDNA: double stranded DNA

emPCR: emulsion PCR

EST: Expressed Sequence Tags

FPC: Fingerprinted Contigs

FR: Forward Reverse

GABI: Genome Assembly by Bayesian Inference

Gbp: giga base pair

GC: Guanine-Cytosine content

GBS: Genotyping by sequencing

GPS: Genome Parsing Suite

GPU: Graphics Processing Unit

HPC: High Performance Computing

Kbp: kilo base pair

IWGSC: International Wheat Genome Sequencing Consortium

Mb: mega base pair

MP: Mate pair

MPI: Message Passing Interface

MTP: Minimum Tilling Path

MYA: million years ago

NGS: Next Generation Sequencing

OLC: Overlap Layout Consensus

ONT: Oxford Nanopore Technologies

PAC: P1-derived Artificial Chromosome

PE: Paired End

PCR: Polymerase Chain Reaction

PFGE: Pulse Field Gel Electrophoresis

PGM: Personal Genome Machine

QTL: Quantitative Trait Loci

RF: Reverse Forward

RNA: Ribonucleic acid

RNA-seq: Ribonucleic Acid Sequence

SBS: Sequencing by Synthesis

SMRT: single molecule real-time sequencing technology

SNP: Single Nucleotide Polymorphism

SRA: Short Read Archive

ssDNA: single stranded DNA

TES: Transposable Elements

TGS: Third Generation Sequencing

TREP: Triticeae Repeats

WCS: Whole Chromosome Shotgun

WGP: Whole Genome Profiling

WGS: Whole Genome Shotgun

YAC: Yeast Artificial Chromosomes

ZMW: Zero-mode waveguide

1 Introduction

Volatile climatic changes and weather conditions continue to erode gains made in food productivity globally resulting in food shortages and rising food prices. Urgent measures are required to identify and improve cereal crop varieties that are resistant to diverse and extreme changes in climate such as higher temperatures and less rainfall. Cereals constitute over 60% of the world's food sources. In the African continent, cereals constitute 46% of the diet and 26% in Western Europe (www.FAOstat.fao.org). Global wheat production forecast for 2014 was 702 million tons, which constitutes 28% of the total cereal production estimate of 2,458 million tons in the same year. While world wheat requirements were forecast to grow by 1.7 percent to 699 million tons, climatic conditions continue to hinder production. The significance of wheat and cereals in general as a global factor food source thus cannot be overstated.

Attempts to improve crop yields have involved the use of high-resolution genetic markers in crop breeding programs for the identification of new high yielding crop varieties. The development of such markers is hampered by the complexity posed by the genomes of many cereal crops. This complexity hinders the generation of accurate and complete genome assemblies that would act as reference assemblies enabling accelerated accurate location, identification, characterization of agronomic traits on the wheat genome through genetic mapping.

Polyploidy is a major contributor to genome complexity and is widespread in plants (Masterson, 1994). Plant genomes have a high repeat content and have complex evolutionary histories in addition to polyploidy that further confound analysis. To adequately characterize cereal genomes, access to a complete and accurate genome assembly is a necessity. This thesis presents a methodology optimised for the generation of such assemblies and demonstrated by the assembly of the short arm of chromosome 7D of bread wheat.

1.1 The *Triticum aestivum* (bread wheat) genome

Triticum aestivum (bread wheat) is a monocot belonging to the Poaceae subfamily and tribe *Triticeae*, which includes barley and rye. The bread wheat genome is approximately 17 Gbp. It is 3.4 times larger than that of barley (*Hordeum vulgare*, 5 Gbp), twice as large as that of rye (*Secale cereale*, 9 Gbp) and approximately 1.5 times that of Oat (*Avena sativa*, 11 Gbp). It has been estimated that tetraploid wheat was domesticated approximately 10,000 years ago with its success as a cultivated crop attributed to genome plasticity as a result of polyploidy (Dubcovsky and Dvořák, 2007). Polyploidy confers the new species with an ability to adapt to diverse environmental conditions.

Polyploidy occurs by the hybridization followed by chromosome doubling of two or more genomes in one organism. Polyploids are classified into two main types; autopolyploids and allopolyploids. Autopolyploids arise within the same species by doubling of homologous genomes and allopolyploids arise by interspecific hybridization followed by doubling of genomes. Polyploidization in plant evolutionary history has been shown to confer adaptive advantages in subsequent generations following a polyploidization event (Parisod *et al.*, 2010). The mechanisms underlying such changes are unknown but are thought to involve several processes. These include gene silencing as a result of DNA methylation, formation of recombinant genomes through hybridization, activation/inactivation of transposable elements and the elimination of coding and non-coding sequences. Recent studies suggest polyploidization conferred increased salinity tolerance in *Arabidopsis* (Chao *et al.*, 2013).

In bread wheat, polyploidization events have resulted in rapid genomic changes in the F1 hybrids and subsequent changes spread sporadically over multiple generations (Feldman and Levy, 2005). Polyploidization in bread wheat resulted from the hybridization of three genomes of *Triticum urartu, Aegilopsis tauchii* and a species similar to *Aegilops speltoides* in two polyploidization events (Baum *et al.*, 2009). Thus, bread wheat has six sets of chromosomes (allohexaploid), which originate from these three diploid ancestors, each with seven sets of chromosomes, through a series of divergence and polyploidization events (Figure 1-1). The diploid donors are understood to have diverged from a shared ancestor between 2.5 and 6 million years ago (MYA). The two diploid donors, *Triticum urartu* (AuAu) and an unidentified species (BB) that is highly similar to *Aegilops speltoides,* following a polyploidization event, combined their genomes around 0.5 and 3 MYA. This event produced the allotetraploid genome of *Triticum turgidum* (AuAuBB) (Chantret *et al.*, 2005; S., Huang *et al.*, 2002). After this first polyploidization, it is believed that the domestication of wheat then followed. During this time, the second polyploidization event occurred resulting in the further combination of *T. turgidum* (AuAuBB) and *Aegilops tauschii* (DD). This polyploidization event produced the allohexaploid *T. aestivum* (AuAuBBDD). Despite having three seperate genomes, *T. aestivum* functions like a diploid organism due to Ph locus (Martinez-Perez *et al.*, 2001; Gill, 2004). Ph genes enables the correct pairing and segregation of chromosomes during meiosis in polyploids. Correct pairing of chromosomes during meiosis ensures genetic stability and fertility of the polyploidy species. When compared to the allotetraploid wheat (wild emmer), hexaploid wheat shows better adaptation to aluminium, frost, salt tolerance, low pH and resistance to pests and diseases (Dubcovsky and Dvořák, 2007).

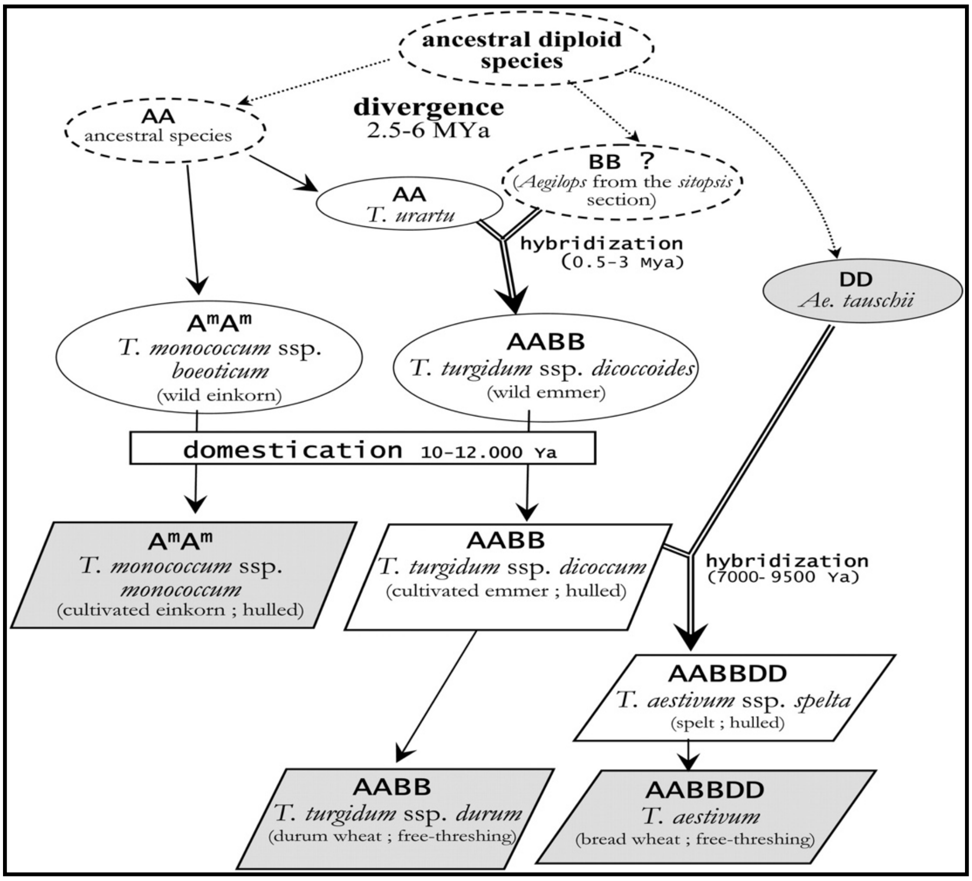
ancestral diploid species
divergence
2.5-6 MYa
AA
ancestral species
BB ?
(Aegilops from the sitopsis section)
AA
T. urartu
hybridization
(0.5-3 Mya)
DD
Ae. tauschii
AmAm
T. monococcum ssp. boeoticum
(wild einkorn)
AABB
T. turgidum ssp. dicoccoides
(wild emmer)
domestication 10-12.000 Ya
AmAm
T. monococcum ssp. monococcum
(cultivated einkorn ; hulled)
AABB
T. turgidum ssp. dicoccum
(cultivated emmer ; hulled)
hybridization
(7000- 9500 Ya)
AABBDD
T. aestivum ssp. spelta
(spelt ; hulled)
AABB
T. turgidum ssp. durum
(durum wheat ; free-threshing)
AABBDD
T. aestivum
(bread wheat ; free-threshing)

Figure 1-1: The polyploid ancestry of wheat. Image from (Chantret *et al.*, 2005).

In addition to the rich polyploid ancestry, bread wheat has also been shown to have a high structural and gene order conservation (synteny) with related grasses such as rice (*Oryza sativa)* (International Rice Genome Sequencing Project, 2005), sorghum (*Sorghum bicolor*) (Paterson *et al.*, 2009) and *Brachypodium* (Schmutz, Rokhsar, *et al.*, 2010). Wheat diverged from *Brachypodium* ~25-30 MYA (Bossolini *et al.*, 2007) while rice diverged from *Brachypodium* ~45 MYA, and sorghum diverged earlier ~50 MYA (Schmutz, Rokhsar, *et al.*, 2010). Given that *Brachypodium* and rice show synteny in ~ 80% of their genes (Wicker *et al.*, 2011), bread wheat should thus show > 80% synteny with *Brachypodium*. This syntenic relationship, in addition to previous genomic studies in wheat using the Chinese spring wheat cultivar has been exploited to validate draft wheat genome assemblies (Paux *et al.*, 2008; Gill, 2004).

Despite the synteny observed among wheat and its progenitors, chromosome-specific rearrangements have been observed. Chromosomal rearrangements include; duplications, inversions, translocations and deletions. Rearrangements result from the breakage of a chromosomal segment at two locations and the re-joining of these ends resulting in a different chromosomal segment or portion. Several inversions, rearrangements and translocations in wheat have been documented. Wheat chromosome 4A long arm (4AL) has translocations from chromosome 5A long arm (5AL), chromosome 7B short arm (7BS) and an inversion (Devos *et al.*, 1995; Miftahudin, 2004). Other translocations have been documented as a result of crosses between wheat and related grasses including rye (An *et al.*, 2013; Fu *et al.*, 2014; C., J., Liu *et al.*, 1992). Linkage analyses with molecular markers have provided insights into the evolutionary origins of these translocation and inversion events (J., C., Nelson *et al.*, 1995; C., J., Liu *et al.*, 1992; Devos *et al.*, 1995) as a result of polyploidy. Analysis suggests that the 4A/ 7B translocation occurred at the tetraploid level as it is in *T. durum* (2n=4x=28, genomes AABB). The 4AL/5AL translocation on the other hand occurred at the diploid level as it is in *T. monococcum* (2n = 2x = 14, genome AA) (Devos *et al.*, 1995; Jian Ma *et al.*, 2013).

In wheat, between 94,000 and 96,000 genes were identified using whole genome shotgun (WGS) assemblies (Brenchley *et al.*, 2012), while using isolated chromosome arm assemblies by the International Wheat Genome Sequencing Consortium (IWGSC), identified 124,201 genes (IWGSC *et al.*, 2014). The variation in gene counts suggests that a considerable proportion of genomic features have not been studied due to incomplete reference genomes, variation on sequencing platforms and the use of different gene annotation methods. Although there is structural and gene conservation between the three homoeologous chromosomes (Berkman *et al.*, 2013; Brenchley *et al.*, 2012; IWGSC *et al.*, 2014), there is evidence of differential gene loss across the sub genomes A, B and D (Berkman *et al.*, 2013; IWGSC *et al.*, 2014). It is estimated that between 10,000 and 16,000 genes have been lost in hexaploid wheat compared with its three diploid progenitors. This gene loss estimate is consistent with earlier studies of gene loss in synthesized wheat polyploids (Ozkan *et al.*, 2001) and the erosion of genetic diversity during wheat domestication (Haudry *et al.*, 2007).

The bread wheat genome contains approximately 80% repetitive DNA sequences (Wicker *et al.*, 2011). These repetitive sequences have been found to comprise mainly of transposable elements (TEs) (Gulick *et al.*, 2009; Choulet *et al.*, 2010) which further confound accurate sequencing, assembly, gene assignment and read mapping. TEs compose large proportions of most plant genomes and have been directly implicated in genome size expansion in maize where TEs were shown to be nested in high frequencies in inter-genic regions (SanMiguel *et al.*, 1996). In *Oryza australiensis*, a wild relative to rice, the amplification of TEs and not polyploidization has been solely attributed to its genome size doubling compared to its closest relatives (Piegu *et al.*, 2006).

TEs have been shown to influence gene expression through epigenetic mechanisms as a result of disruption or amplification of gene segments (Lippman *et al.*, 2004). The distribution of transposons across the three sub-genomes differs. Class I elements (retro-transposons) are more abundant in the A sub-genome (A > B > D), while class II elements (DNA transposons) are more abundant in the D sub-genome (D > B > A) (IWGSC *et al.*, 2014). Group 7 wheat chromosomes showed low polymorphism in the D sub genome relative to the A and B (A > B > D) (Berkman *et al.*, 2013). Genes on specific wheat sub genomes encode several disease resistance traits. These include stripe, stem and leaf rust resistance genes on chromosome 1B (Q., Huang *et al.*, 2014; Y., Zhou *et al.*, 2013), 2A (Bariana and McIntosh, 1993), 2B (P., Luo *et al.*, 2009), 2D (Tsilo *et al.*, 2008), 3D (Q., Li *et al.*, 2010) and 6B (Dadkhodaie *et al.*, 2010). Powdery mildew resistance genes are encoded on chromosome 1A (Hsam *et al.*, 2014) and 6D (Hongqi Ma *et al.*, 2011) while QTLs containing Fusarium head blight resistance genes have been found on 5A and 3B (Tamburic-Ilincic, 2012), 7A (Jayatilake *et al.*, 2011) and 7D (Cativelli *et al.*, 2013).

Agronomic traits for yield and grain quality have been associated with several chromosomes. Genes influencing baking and milling quality have been identified on chromosome 1D (Dumur *et al.*, 2009) and chromosome 3B and 4(Carter *et al.*, 2011) while traits associated with grain quality and colour have been identified on chromosome 3A (Crawford and Francki, 2012) and 7A (Crawford and Francki, 2013). Genes influencing yield which include grain size, number and density of grains in a spike, kernel sizes, plant sizes and height have been identified on chromosome 3A (Rustgi *et al.*, 2013), 2D (Nalam *et al.*, 2007; Guizhi Zhang *et al.*, 2015), 4A (L., Liu *et al.*, 2010), 4B (Kadam *et al.*, 2012), 6A (Spielmeyer *et al.*, 2007; Simmonds *et al.*, 2014) and 7D (Röder *et al.*, 2007). QTLs influencing pre harvest sprouting have been identified (Cabral *et al.*, 2014) and more specifically on chromosome 2B (Somyong *et al.*, 2014) and 2D (Apotikar *et al.*, 2011). Other traits identified include Aluminium tolerance on chromosome 2D (Anioł, 1995), and chloride and frost tolerance on 5A and 5B (Iehisa *et al.*, 2014; Genc *et al.*, 2014; Kocsy *et al.*, 2010; Vítámvás *et al.*, 2011).

1.2 The *Triticum aestivum* (bread wheat) chromosome arm 7DS

The 381 Mb (Safár *et al.*, 2010) short arm of chromosome 7D has several genes and QTLs of agronomic importance. These include QTLs for grain yield (X., Q., Huang *et al.*, 2004; Röder *et al.*, 2007), resistance genes for the Russian wheat aphid *Diuraphis noxia* (Peng *et al.*, 2008; X., M., Liu *et al.*, 2005; Miller *et al.*, 2001; Castro *et al.*, 2004), powdery mildew, leaf rust, yellow rust and stem rust (Krattinger *et al.*, 2009). *Septoria tritici* blotch resistance genes have also been identified on chromosome 7DS (Adhikari *et al.*, 2004; Arraiano *et al.*, 2001) in addition to Fusarium head blight resistance genes (Cativelli *et al.*, 2013).

Among the group 7 chromosomes of sub-genomes A, B and D, the D genome has been found to contain the highest proportion of genes. The average level of synteny with *B. distachyon*, *O. sativa*, and *S. bicolor* is slightly greater on the D genome at 58%, compared to A at 51% and B at 50% (IWGSC *et al.*, 2014). Three syntenic regions have been identified between chromosome 7DS and chromosomes 1 and 3 of *B. distachyon* (Figure 1-2), (Schmutz, Rokhsar, *et al.*, 2010; Berkman *et al.*, 2011), with 7DS having between 1,222 (IWGSC *et al.*, 2014) and 1,488 (Berkman *et al.*, 2011) syntenic genes with *B. distachyon*.

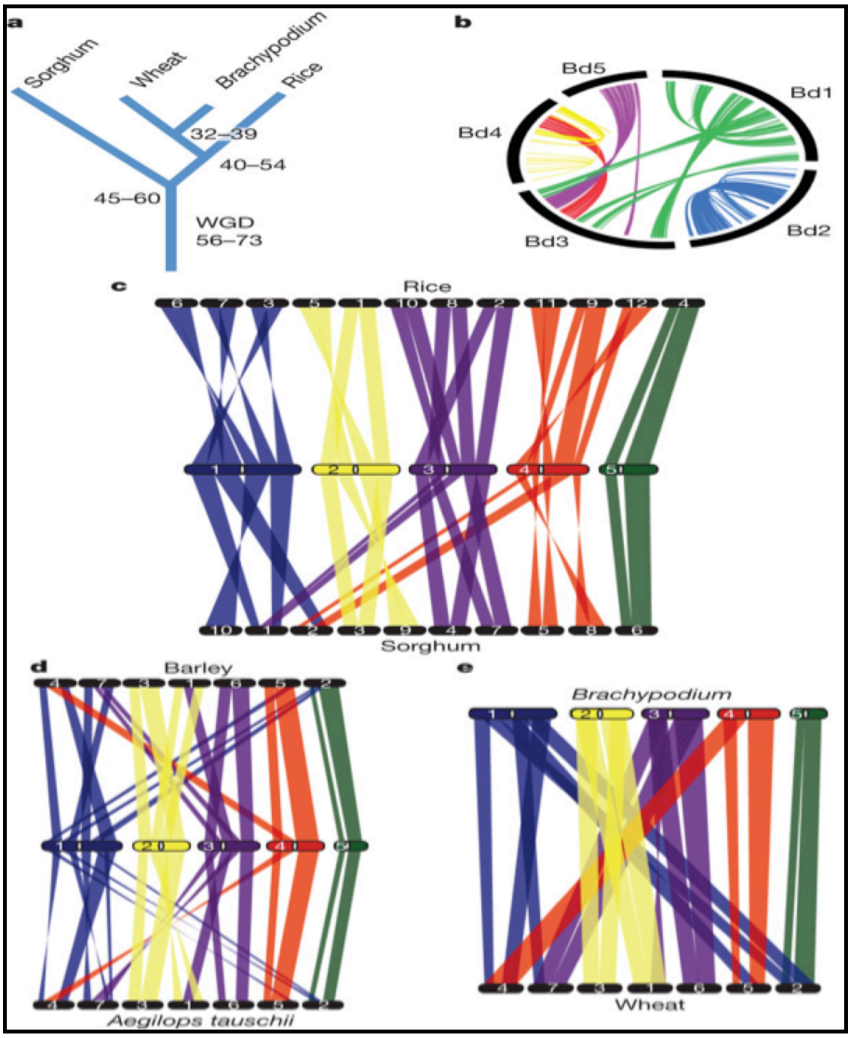

Figure 1-2: Polyploid history and synteny of the bread wheat genome to rice, sorghum and *Brachypodium*. Image from (Vogel et al. 2010).

Gene loss in wheat has been estimated at between 10,000 and 16,000 genes in hexaploid bread wheat compared with the three wheat diploid progenitors (Brenchley *et al.*, 2012; Ozkan *et al.*, 2001), but differential gene loss across the sub genomes has been observed, being slightly lower on the D genome compared to the A and B. Desirable traits have been introduced into wheat through the generation of translocation lines. Translocations have been observed involving chromosome 7DS in wheat barley hybrids (Kruppa *et al.*, 2013). The generation of chromosome-specific deletion lines is a pre-requisite for chromosome sorting through flow cytometry (see 1.4.1). Six deletion lines have been obtained for chromosome 7DS (Endo and Gill, 1996).

1.3 Current DNA sequencing technologies

DNA sequencing technologies continue to evolve rapidly from the 1970's following the development of Sanger sequencing (Sanger *et al.*, 1977). Sanger sequencing is still the most accurate sequencing method available that also offers long read lengths of between 800 -1000bps. The main disadvantage compared to other platforms is that it is expensive, time-consuming and laborious and thus has been replaced with next generation sequencing technologies (NGS). NGS technologies offer higher throughput, but shorter read lengths compared to Sanger sequencing. Attempts at increasing the read lengths of NGS platforms have resulted in the development of paired-end sequencing, which leverages the short read lengths. Mate pair (MP) libraries have been introduced to span greater genomic distances. Attempts have been made to merge further overlapping paired reads into longer single reads (Ruan *et al.*, 2013; Rodrigue *et al.*, 2010; Magoc and Salzberg, 2011; B., Liu *et al.*, 2012).

Although some NGS platforms produce sequenced reads with a relatively low read error rate, attempts at sequencing of longer reads has resulted in increased error rates and sequencing bias. Third generation sequencing (TGS) technologies are rapidly being developed and improved to address this limitation. TGS technologies are targeted at the generation of accurate longer reads. This would significantly advance genome assembly efforts. Such technologies include Pacific Biosciences, Ion Torrent and Oxford Nanopore discussed in section 1.3.2 to 1.3.4 below.

1.3.1 Illumina

Illumina (www.illumina.com) have a range of NGS platforms. Illumina platforms are based on sequencing by synthesis (SBS) (Bentley *et al.*, 2008). The current Illumina NGS platforms are the MiSeq, HiSeq, NextSeq and HiSeq X series. The choice of a sequencing platform to use depends on the scale and sequencing throughput needs. Currently, the MiSeq produces between 0.3-15 Gb of 2x300 bps reads in 5-55 hrs. per sequencing run, while the NextSeq produces between 20-120 Gb 2x150 bps reads in 15 - 30 hrs. The HiSeq and HiSeq X series produce more data at 10 - 1500 Gb and 900 - 1800 Gb both at read lengths (2x150 bps). The high data output and low error rates makes this the preferred sequencing platform for *de novo* genome sequencing, diversity studies, re-sequencing and SNP discovery despite shorter read lengths. Barcoding or indexing of samples with short identifying oligonucleotide enables the pooling of several samples within a single sequencing run.

Illumina SBS involves the sequencing of a DNA molecule one nucleotide at a time using DNA polymerase and nucleotide substrates. This involves two steps, template preparation and sequencing. Template preparation involves the enrichment and amplification of the DNA sample using commercially available sample preparation kits. The template preparation depends on the sequencing approach being used, for example paired-end (see 1.3.1.1 below) or mate-pair (see 1.3.1.2 below). During template preparation, the double-stranded DNA (dsDNA) molecules are denatured into single-stranded DNA (ssDNA) molecules. Sequencing primers are then attached followed by iterative amplification of the ssDNA molecules resulting in the formation of clusters of identical molecules on a flow cell (Figure 1-3 below). This is referred to as bridge amplification. Bridge amplification involves cycles of annealing, extension and denaturation. During amplification, biases may be introduced in the cluster formation due DNA polymerases binding preferentially to GC-rich DNA templates (Multer and Boynton, 1995).

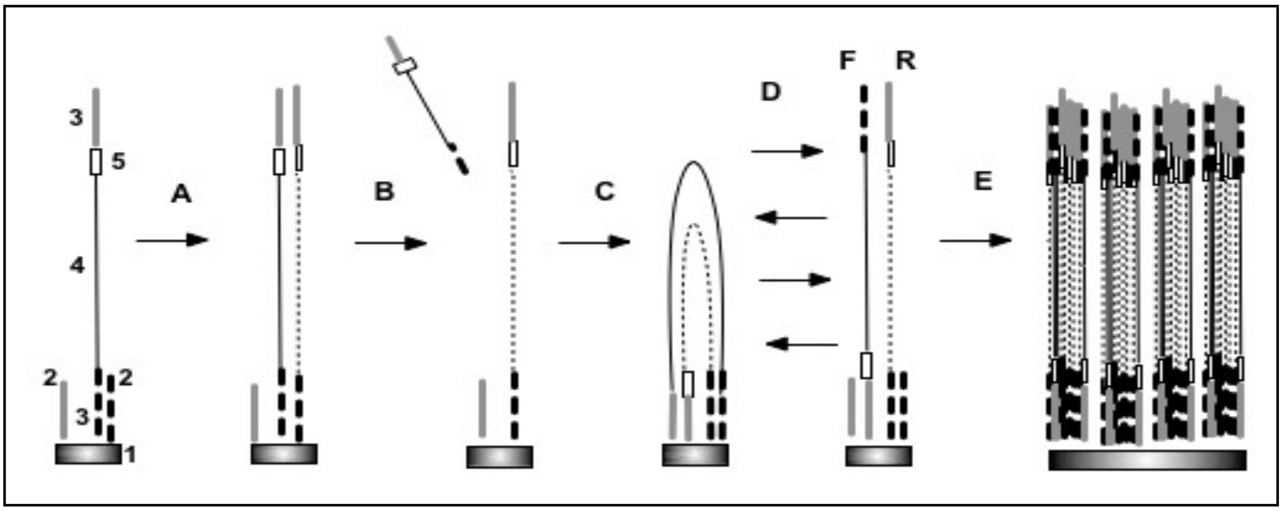

Figure 1-3: Illumina bridge amplification. The Illumina flow cell (1) has two types of oligonucleotides (2) attached, which are complimentary to sequencing adapters (3) attached to ssDNA molecule (4). If multiple samples are run, an index is used (5) to tag reads. DNA polymerase synthesizes a new strand from the template (A) and the dsDNA template is denatured and original template washed off (B). The newly synthesized strand bends and binds to the complimentary oligonucleotide on the flow cell followed by another polymerase reaction resulting in a dsDNA molecule (C). The dsDNA molecule is denatured and the previous step is repeated multiple times (D). This results in clusters of amplified DNA strands (E). The forward strands (F) are then washed off prior to sequencing.

The clonally amplified DNA fragments generated during bridge amplification form a template from which new DNA fragments (reads) are synthesized. During sequencing the DNA molecules are read and extended during the extension phase of the sequencing cycles (Figure 1-4), fluorescently tagged Deoxyribonucleotide triphosphate (dNTPs), (A, C, T, G) are added to the growing DNA strand. With each addition of the fluorescently labelled dNTPs, a fluorescent signal is captured followed by removal of the fluorescent group on the dNTPs allowing for the extension and addition of a new dNTPs. Images captured are compiled, and the DNA sequence is determined by analysing the fluorescent signal during each extension step. Ideally, all DNA molecules in a cluster should be the same sequence and the extension steps should occur at the same loci on all the DNA molecules in the cluster, but this does not always occur. As a result, some molecules get sequenced a base later than others resulting in noise when distinguishing fluorescent signals and poor base calls. This is commonly known as phasing and is mostly observed as a reduction of sequence quality at end of Illumina reads. Despite these limitations, Illumina currently dominates the NGS field given the high data output and relatively low error rates and is the sequencing platform of choice for this thesis.

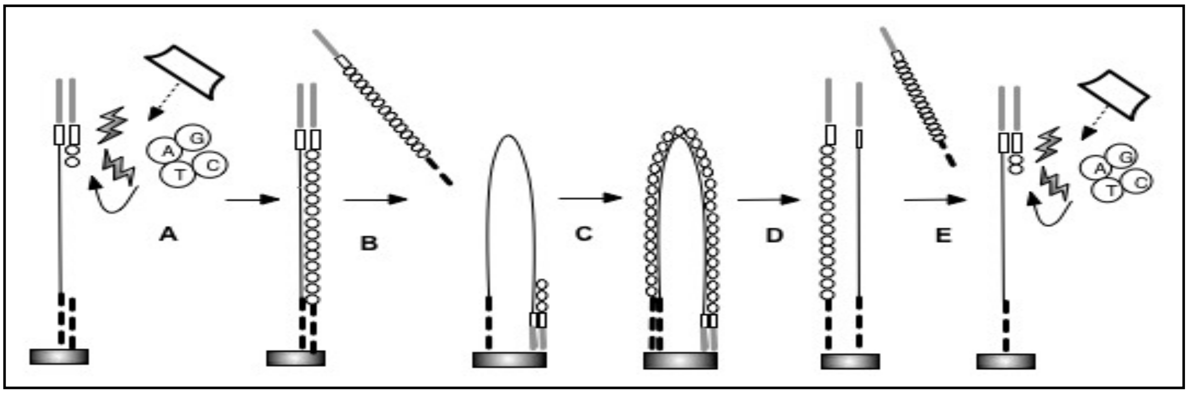

Figure 1-4: SBS. (A) Sequencing of the first read begins by attachment of primers to amplified ssDNA molecules. Fluorescently tagged nucleotides are added a base at a time complimentary to the ssDNA. With each addition the nucleotide is excited with a light source and the fluorescent signal emitted captured. (B) Sequenced product is washed off and the 3' end of the template is de-protected. The template bends and (C) DNA polymerase synthesis the second stand forming a dsDNA bridge. (D) The dsDNA molecule is linearized, its 3' end blocked and original template washed away. (E) Read two sequencing begins as in (A). Steps A - E occur in parallel in all clusters on the flow cell.

1.3.1.1 Paired end libraries

Illumina paired-end sequencing produces reads with short insert sizes usually between 300 bps to 500 bps. During the generation of paired-end libraries, DNA is fragmented and size selected followed by adapter ligation of the dsDNA fragments (Figure 1-5). Ligated fragments are denatured and anchored onto a flow cell and amplified through bridge amplification as previously discussed (Figure 1-3) above. Amplified fragments in the clusters are then sequenced by synthesis as discussed above (Figure 1-4). The resultant reads have a forward and reverse (FR) orientation when mapped to a reference.

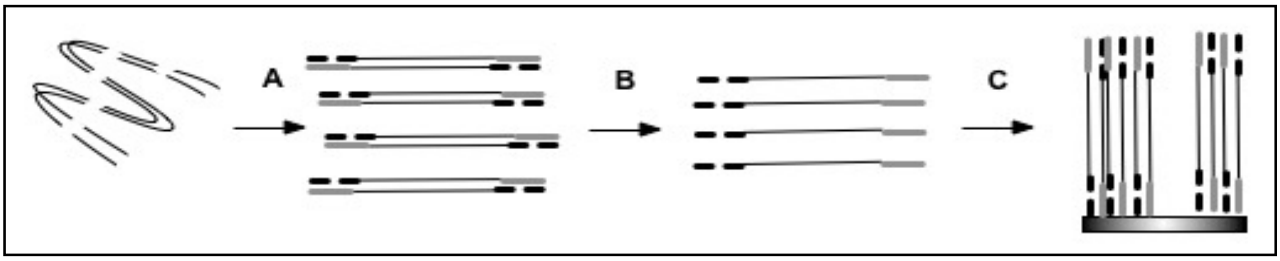

Figure 1-5: Paired-end read generation. (A) DNA is digested and dsDNA fragments ligated with adapters. (B) Ligated fragments are denatured and bound to the sequencing flow cell. (C) Attached ssDNA molecules undergo bridge amplification resulting in clusters followed by sequencing.

1.3.1.2 Mate pair libraries

For genome scaffolding, genome rearrangement and variant detection purposes, long insert libraries of several Kilobases (Kb) are used. Here we describe the Illumina mate pair protocol utilized in the generation of the mate pair datasets used in this thesis. A DNA sample is fragmented and ends labelled with a biotin tagged (biotinylated) junction adaptor (Figure 1-6). The junction adapters bind to each other circularizing the DNA fragment into a closed loop. Linear DNA is digested, and the circularized DNA is sheared and size selected resulting in only fragments that contain the biotinylated adapters. These final fragments are made up of a pair of sequences originating from regions several Kbs apart from the same DNA strand and joined by biotin adapters. External sequencing adapters are then ligated to the ends of the fragments and sequenced using the Illumina paired-end sequencing method described in 1.3.1.1. Sequenced reads have an insert size that is determined by the length of the circularised molecule providing several kilobases spanning distances for analysis using the paired-end sequencing protocol.

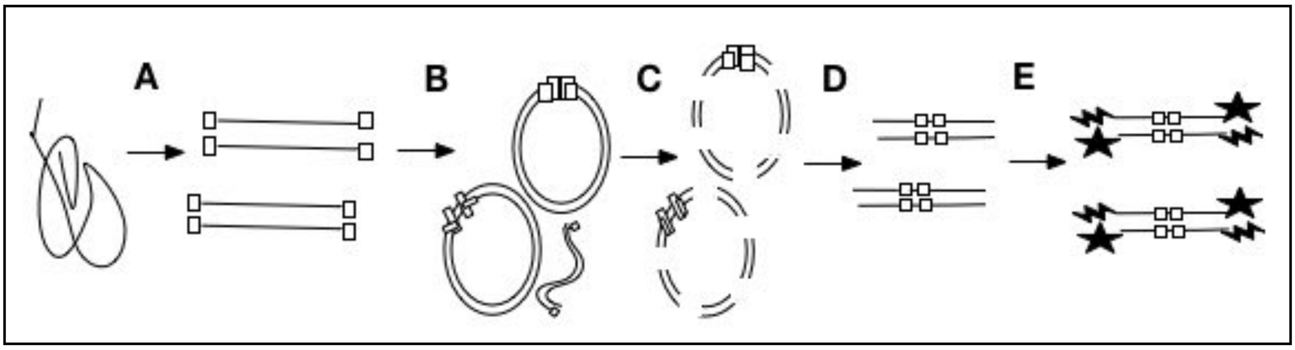

Figure 1-6: Mate pair generation. (A), DNA fragment is digested and fragment ends labelled with biotin adapters. (B), DNA fragments are circularised using biotin adapters. (C), Non-circularised DNA is digested and circularized DNA sheared. (D), Biotin labelled DNA is size selected. (E) Illumina sequencing adapters are ligated to DNA strands.

The mate pair protocol results in sequenced read pairs having various orientations when mapped to an assembly (Figure 1-7). The proportion of reads with different orientations depends on the efficiency of the mate pair protocol and library preparation. Additional processing of MP reads is required prior to mapping. The read processing depends on the position of the junction adapter in the read.

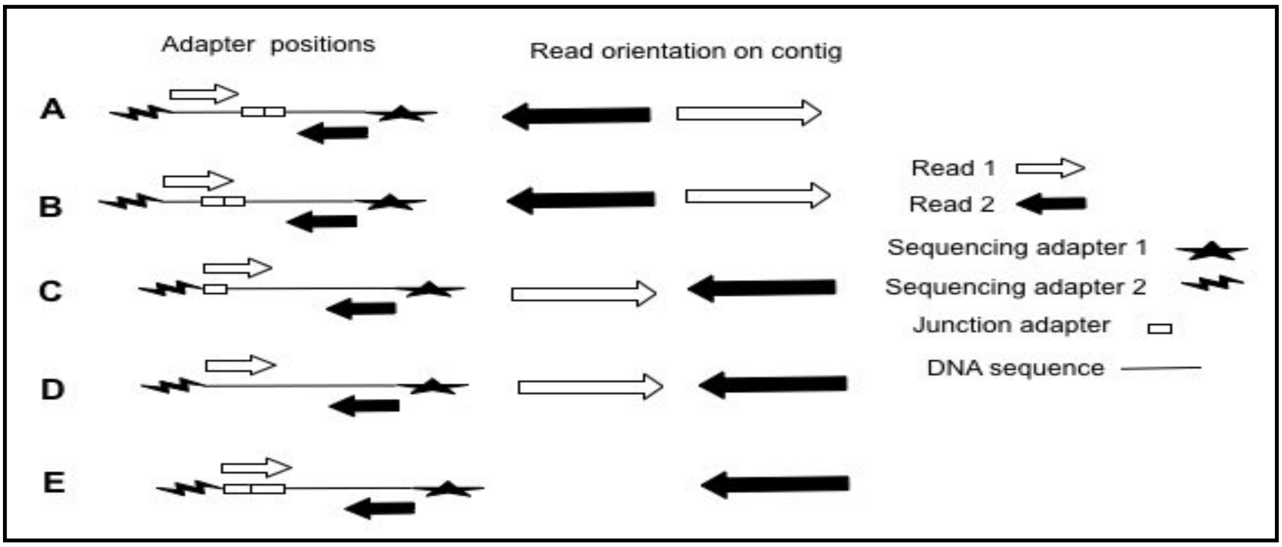

Figure 1-7: MP adapter positions and mapping orientations. Positions of junction adapters determine the mapping orientation of mate pairs on the assembly.

If junction adapters are found at positions in **A** or **B** on sequenced reads (Figure 1-7) above, they result in a reverse forward (RF) read mapping orientation when mapped to a reference. When a read runs into the junction adapter as in **B**, (Figure 1-7), the adapter sequence is trimmed prior to read mapping. Where the junction adapter appears at the 5' end of a read as shown in **C**, the adapter is also trimmed off. Junction adapters may also be absent as in **D**. In **C** and **D**, sequenced pairs map in a forward-reverse (FR) orientation as paired-end reads. The insert size of pairs in **D** may reflect those of a MP library or a PE library. In the case of **E** where a big proportion of the read is composed of a junction adapter, the read is discarded depending on a minimum read length threshold set resulting in an orphaned mate. Existing methods that pre-process MP datasets discard reads in cases C and E above. Technical challenges involved in the making of mate pair libraries result in variable levels of contaminants. Several tools have been developed to pre-process mate pair libraries, but most discard sequences that appear on the 3' of the junction adapter (O'Connell *et al.*, 2014; H., Jiang *et al.*, 2014; Leggett *et al.*, 2014). Newer tools are being developed to optimize the pre-processing of mate pair libraries such as NxTrim (O'Connell *et al.*, 2014), Cutadapt (Martin, 2011) and NextClip (Leggett *et al.*, 2014).

1.3.2 Pacific Biosciences

Pacific Biosciences' sequencing platform is a single molecule SBS approach that is based on real-time imaging of fluorescently labelled nucleotides. The single molecule real-time (SMRT) sequencing technology (Eid *et al.*, 2009) does not require the DNA template to be amplified as it is sensitive enough to detect the addition of a single nucleotide during sequencing. Prior to sequencing, DNA is fragmented and adapters ligated. The SMRT sequencing adapters are unique in that they have a hairpin structure that allows dsDNA molecules to be circularized. A sequencing primer attaches to the ssDNA portion of the SMRT adaptors allowing DNA polymerase to bind. Depending on the length of the dsDNA molecule, the sequencing run can result in sequencing of the molecule several times for short molecules or once for longer molecules. The sequencing reaction occurs in zero-mode waveguide (ZMW) wells on the SMRT cell (Levene, 2003; Korlach *et al.*, 2008). As the DNA molecule is synthesized, each nucleotide added is recorded using an optical system that reads individual ZMW cells detecting the fluorescence of the incorporated nucleotide in real time (Lundquist *et al.*, 2008).

Nucleotide extension does not require the termination of the sequencing reaction for a tagged nucleotide to be identified prior to the incorporation of the next nucleotide. As a result, continuous real-time recording of incorporated nucleotides is achieved. The continuous real-time recording of incorporated bases may be a contributing factor to the high single pass error rate (10% –15%) that is observed. As errors are randomly distributed insertion/deletions with few miscalled bases, their effect is moderated by the generation of consensus base calls using ~40x coverage per base. Such an approach has been shown to increase the accuracy of base calls to 99.999% by using the Quiver software (Chin *et al.*, 2013). In addition, BLASR, a BLAST like read mapper was developed that takes into account the long, relatively inaccurate reads (Chaisson and Tesler, 2012).

1.3.3 Ion torrent

Ion Torrent's next-generation sequencing platform, the Personal Genome Machine (PGM) employs a non-optical sequencing technique to identify nucleotides during sequencing. Non-optical sequencing exploits complementary metal-oxide semiconductor (CMOS) circuitry that is the basis of modern day electronics. Ion Torrent has developed an electronic semiconductor sensor chip able to detect Hydrogen ions (H+) released by DNA polymerase during DNA template strand extension (Sakurai and Husimi, 1992). Given the reduced costs associated with library preparation and the sequencing instrument development, the PGM offers a low cost, scalable sequencing platform.

Genomic DNA is prepared for sequencing via emulsion PCR (emPCR) (Nakano *et al.*, 2003). DNA is sheared and ligated to adapters. The adapter-ligated DNA is amplified on beads. Two sequencing primers complementary to library adapters are used to select and only amplify those DNA molecules that contain both primers. DNA molecules amplified on beads are then enriched and loaded into the sequencing chip.

Current sequencing chips output paired-end reads of 200 - 400 bp and 1 - 2 Gb of sequences per run on the Ion PGM chips. The newer Ion Proton chips produce 200 bp paired-end reads with up to 10 Gb per run. Conservative error estimates for substitutions are ~0.1% but are higher for 5-mer homopolymer regions estimated at ~3.5% (Merriman *et al.*, 2012). High error rates have been attributed to inaccurate detection and distinction of bases incorporated during sequencing. Incorporated bases are not identified based on the emission of a specific fluorescent signal as with Illumina SBS but on the release of H+. Released H+ do not differentiate individual dNTPs, dNTPs have to be added in a cyclic predetermined order, and when a base is incorporated, a pH change is observed based on the emitted H+. Longer homopolymer regions do not always result in an increase in the pH proportional to the number of bases as would be expected (Buermans and Dunnen, 2014). Attempts to compensate for these effects include the use of specialized algorithms in base calling and read mapping that take into account these limitations (Golan and Medvedev, 2013).

1.3.4 Oxford Nanopore

Oxford Nanopore Technologies (ONT) has introduced a miniature nanopore disposable DNA sequencer, the MinION (Mikheyev and Tin, 2014) which outputs on average, reads of length of 5.4 kilobases. Nanopore sequencing is based on the detection of nucleotides of a DNA strand in a nanopore (Kasianowicz *et al.*, 1996). Nanopore sequencing offers clear advantages over other sequencing platforms in that DNA amplification is not required, and base detection does not involve fluorescent labelling. The nanopore consists of seven α-haemolysin membrane proteins that form a haemolysin pore into which a single DNA strand is inserted (Hall *et al.*, 2010). Enzymes are utilized to direct DNA strands into the nanopore (strand sequencing) or to attach the DNA to the pore followed by cleaving one base at a time (exonuclease sequencing).

During library preparation, the DNA is sheared and adapters ligated, incubated to bind the transporter proteins, and loaded onto the sequencer. Differential changes in the electrical signal as the four bases are passed through the pore enable the identification of respective nucleotides (Stoddart *et al.*, 2009). The electrical signals are analysed using a hidden Markov model to determine the base detected (Timp *et al.*, 2012). Further unpublished data by ONT suggests that nanopores can distinguish between a cytosine and its methylated form (5-methylcytosine). Despite the clear advantages offered by ONT, the high error rate, currently estimated at ~10% may hinder uptake of this as a sequencing platform of choice. Preliminary evaluation of read accuracy using lambda phage showed only 10% of the reads mapped to the reference (Mikheyev and Tin, 2014).

1.3.5 Roche 454 pyrosequencing

The Roche/454 sequencing platform was the first to integrate and commercialise pyrosequencing on their GS20 system (Margulies *et al.*, 2005). Pyrosequencing involves the use of bioluminescence to measure inorganic pyrophosphates released enzymatically during the addition of a single dNTP. Different dNTP's will emit different levels of bioluminescence enabling the identification of the incorporated dNTP (Ronaghi *et al.*, 1996; Ronaghi *et al.*, 1998). Through repeated cycles of incorporation of dNTPs, whereby the DNA strand is sequenced by the synthesis of the complimentary strand, bioluminescence peaks are recorded as flowgrams revealing the DNA sequence (Margulies *et al.*, 2005). The pyrosequencing platform was further optimized to use utilize emPCR for DNA template amplification. Amplified fragments are then loaded onto fabricated Pico Titer Plates (PTP) for complimentary strand synthesis.

Improvements on the GS20 resulted in the release of the Genome Sequencer (GS) FLX capable of sequencing 12.5 million bases in 1 hr with read lengths of up to 400 bps long compared to 100 bps produced by the previous GS20 system. The GS FLX titanium system also produces paired reads. The latest improvements to the GS FLX system (the GS FLX+) can produce 30 million bases in 1 hr with read lengths of up to 1 kb (http://454.com/products/gs-flx-system/index.asp). While the GS FLX system offers the longest read lengths with highest per base accuracy, sequencing of homopolymer regions still poses a challenge. Homopolymer regions of greater than six nucleotides often result in errors, leading to insertions or deletions in the sequenced DNA.

1.3.6 SOLiD

The Support Oligonucleotide Ligation Detection (SOLiD) platform based on Sequencing by Ligation (SBL) was commercialized by applied biosystems (http://www.appliedbiosystems.com). SBL uses DNA ligase with a single or two base encoded fluorescently labelled probes that are hybridized to the complimentary DNA strand of a DNA template. Ligated probes are then fluorescently imaged to determine the DNA sequence through repeated cycles (Landegren *et al.*, 1988). The use of two base probes results in fluorescent imaging of each base twice, providing improved accuracy in the detection of SNPs in addition to error detection. Two base encoding is a unique feature of the SOLiD system (Valouev *et al.*, 2008; Metzker, 2010) compared to other sequencing platforms. The current SOLiD platform, the 5500 W system, has improved sequencing chemistry that does not rely on emPCR for amplification as libraries are amplified directly on the flow cell. While the negation of emPCR amplification has resulted in reduced library preparation time and costs, the SOLiD platform still lags behind other sequencing platforms in read lengths. The current maximum available read lengths are 75 bps.

1.4 Approaches in *de novo* sequencing and assembly of complex polyploid genomes

The main contributors to plant genome complexity are the genome size, polyploidy and repetitive DNA sequences. These attributes significantly increase the cost of sequencing and impact the quality of the resulting assemblies and downstream analysis. Most land plants have been shown to have had a polyploidization event in their evolutionary history (Soltis *et al.*, 2004). Plants also vary in genome size by several orders of magnitude (Gaut and Ross-Ibarra, 2008) as a result of the prevalence of TEs (Tenaillon *et al.*, 2010). High proportions of TEs may explain the relative larger genomes sizes of plants being greater than those of animals (Lockton and Gaut, 2005). Currently, no single approach addresses this complexity.

Sequencing of diploid relatives (Young *et al.*, 2011; Shulaev *et al.*, 2010; X., Xu *et al.*, 2011) offers insights and a basis for further analysis of complex polyploid crops as the genome complexity is reduced. Heterozygous alleles present in a genome pose challenges to genome assembly algorithms. Specifically, such alleles introduce single nucleotide polymorphisms (SNPs) or small indels in sequenced reads. Such variation within reads from the same loci during assembly result in multiple alternatives by which the genome can be reconstructed at that locus. This greatly complicates the assembly process (Kajitani *et al.*, 2014). For sequencing and assembly, doubled haploids offer reduced heterozygosity, which has been shown to improve genome assembly (Hong Zhang *et al.*, 2014).

An accurate *de novo* genome assembly not only enables the precise and high-resolution identification of markers for improved crop breeding, but also offers insights into plant evolution and comparative genomic studies. An example of such success was in establishing the evolutionary history of soybean (*Glycine max*) as a result of sequencing followed by a high-quality genome assembly of soybean

(Schmutz, Cannon, *et al.*, 2010). An in-depth understanding of the polyploid history offers insights into its domestication. Several approaches have been exploited in the sequencing of complex genomes. These include whole genome shotgun (WGS) sequencing, isolation of chromosomes followed by WGS of individual chromosomes and BAC-by-BAC approaches.

1.4.1 Chromosome isolation

Complexity reduction through the sequencing of isolated chromosome arms in bread wheat has made the sequencing of the 17 Gb genome possible. While chromosome isolation can be done by micro dissection (Stein *et al.*, 1998), flow cytometry is much more efficient and large chromosomes can be isolated and purified in a short time (Doležel and Lucretti, 1995). Flow cytometry involves the high throughput characterization and sorting of molecules or cells using fluid dynamics and lasers. A population of molecules are stained by a DNA-binding fluorochrome and loaded into a flow cytometer where molecules in solution are passed through a single stream of fluid. The stream is further broken up into droplets that ideally contain a single molecule or cell. Droplets are passed in front of a laser beam. The fluorescence intensity emitted when the molecule passes through a laser beam is used to classify the DNA molecules, cells or chromosomes. The fluorescence intensity is a reflection of a molecule's relative DNA content.

The preparation of intact chromosome suspensions and inability to distinguish chromosomes of similar sizes hindered the application of flow cytogenetics in plants. However, the use of chromosome deletion and alien addition lines in polyploid wheat and other cereals has resolved these limitations (Kubaláková *et al.*, 2002). Polyploid genomes can tolerate the addition or removal of whole or partial chromosomes. This is referred to as aneuploidy.

Deletion and alien addition lines are formed when specific chromosomes from related ancestral species such as *Aegilops cylindrica* are introduced into wheat. In the case of the wheat Chinese Spring cultivar, chromosomal breaks are observed in the gametes that do not have the added chromosome. As a result, various chromosomal aberrations including deletions occur and are transmitted to the offspring. Such offspring are maintained as ditelosomic cytogenetic stocks. Ditelosomic offspring result in cases where one of the gametes has a chromosome that divided transversely at the centromere instead of longitudinally as a result of chromosomal aberrations. Such stocks have been developed for wheat (Endo and Gill, 1996). The use of ditelosomic lines has helped in the identification of wheat chromosomes, which in combination with flow cytometry has enabled the identification, sorting and extraction of high molecular weight DNA for specific wheat chromosomes. Other specialized protocols tailored for the preparation of BAC libraries from flow sorted chromosomes (Šimková *et al.*, 2003; Chalhoub *et al.*, 2004) have advanced the application of flow-sorted chromosomes in plant genomics. Sub-genomic BAC libraries specific for several wheat arms have been developed. All wheat chromosome arms have been sorted and subsequently WGS sequenced and assembled (IWGSC *et al.*, 2014). WGS assemblies are usually incomplete due to collapsed repeats and duplicated regions. BAC libraries offer a much higher sequence resolution.

1.4.2 Whole genome shotgun (WGS)

Whole-genome shotgun sequencing (WGS) is a rapid genome sequencing approach that involves the generation of one or several diverse insert size libraries from a template DNA generated from total genomic DNA. The reads are then sequenced and assembled. WGS is considered rapid because it eliminates steps involved in library preparation when using a BAC-by-BAC approach (see below). The assembly algorithm exploits the diverse insert sizes. Reads can then be linked that span across repetitive regions given the insert size is known. For non-repetitive genomes, a WGS would suffice, but for highly repetitive and polyploid genomes, WGS is confounded by repetitive elements during sequence assembly. For such genomes, accurate assembly of sequences using this approach is still problematic, as reads from different parts of the genome would be highly similar. Although WGS is a faster sequencing approach, the correct placement, ordering and orientation of assembled contigs limits the generation of high quality reference genomes.

Attempts at sequencing plant genomes using WGS approaches have been moderately successful. Examples in cereals include the sequencing of *Oryza sativa ssp. japonica (Nipponbare)* (International Rice Genome Sequencing Project, 2005)*, Oryza sativa ssp. Indica* (Goff, 2002), *Sorghum bicolor (L.) Moench* (J., Yu, 2002), *Zea mays (Palomero Toluqueno)* (popcorn) (Vielle-Calzada *et al.*, 2009), *Glycine max* (soybean) (Schmutz, Cannon, *et al.*, 2010), *Setaria italica* (Foxtail Millet) (Gengyun Zhang *et al.*, 2012) and wheat (Brenchley *et al.*, 2012). This approach has resulted in draft assemblies, and in the case of wheat, an orthologous genic assembly of the bread wheat genome (Brenchley *et al.*, 2012). This draft genome was of low accuracy due to repeats and homeologous sequences of the A, B and D genomes.

A common genome sequencing approach is to use bacterial artificial chromosomes (BACs) (Tao, 1998). Earlier approaches included the use of yeast artificial chromosomes (YACs) (Burke *et al.*, 1987) and P1-derived artificial chromosome (PACs) (Sternberg, 1990) to clone DNA fragments. By reducing the genome complexity into a manageable set of BAC clones referred to as a BAC library. BAC clones can be sequenced and reconstructed through *de novo* assembly. The large genome size of polyploids necessitates the cloning of large DNA fragments. While the size of DNA fragments (insert size) that can be cloned in BACs and PACs are smaller than YACs, BACs have replaced YACs and PACs and offer several advantages. These include stability in host cells, transformation efficiency and high replication producing large amounts of DNA.

Construction of BAC libraries is laborious and involves several steps. These include the isolation of DNA and size selection of the DNA molecules by gel electrophoresis. The size-selected DNA could be from an isolated chromosome or total DNA depending on the experimental design. The reduction of size-selected DNA into fragments by partial digestion of the DNA molecules with restriction endonucleases then follows. Partially digested fragments are then ligated into a suitable cloning vector, followed by the transformation of suitable host cells. Transformed cells are collected from colonies identified by incubation on agar plates with an appropriate antibiotic. Cells from transformed colonies are then stored at −80◦C.

The order of BAC clones in a BAC library, as they occur in the genome, is then determined. BAC clones are digested with restriction endonucleases and using the profiles of the resulting restriction fragments as fingerprints, overlapping clones are identified. The restriction profiles of the fingerprinted contigs (FPC) are edited with a computer program and contigs assembled with the aid of software. Redundant clones are discarded resulting in the creation of a minimum-tilling path (MTP) of overlapping clones. Depending on the genome size of the plant, the MTP can consist of several thousand clones, making sequencing individual clones costly. Clones can be pooled and indexed depending on the sequencing experimental design to reduce costs.

A BAC-by-BAC approach has been applied to sequence several complex crops including *Oryza sativa ssp. japonica (Nipponbare)* (International Rice Genome Sequencing Project, 2005)*, Zea mays (B73)* (Schnable *et al.*, 2009) *and Hordeum vulgare (Barley)* (Wise *et al.*, 2012). Following the successful development of BAC libraries of *Triticum urartu*, (AA) *Aegilops speltoides* (BB) and *Ae. Tauschii* (DD), the diploid donor genomes of bread wheat genomes A, B and D respectively (Akhunov *et al.*, 2005), a BAC-by-BAC approach was applied to the sequencing of isolated chromosome arms of bread wheat (Šimková *et al.*, 2011). The first isolated chromosome of wheat to be sequenced and assembled was chromosome 3B (Choulet *et al.*, 2014). Other isolated chromosome BAC libraries have been developed. Such libraries include 1D, 4D and 6D (Janda *et al.*, 2004), 1B (Janda *et al.*, 2006), 3B (Safár *et al.*, 2004) and 7D (Šimková *et al.*, 2011). To date, all wheat chromosomes BAC libraries are available (Safár *et al.*, 2010) and are at different stages of sequencing and assembly.

1.4.3 Overview of assembly algorithms

Genome assembly is analogous to solving a jigsaw puzzle where each jigsaw piece (a read) fits at a particular location in the final assembly. When solving the puzzle, a read may appear to fit in multiple places equally well. Furthermore, some of the reads may be identical; as such the correct placement of the reads is a challenge. A genome assembler merges overlapping reads into contigs and contigs into scaffolds. Contigs constitute a consensus of overlapped reads. Scaffolds are thus contigs that are ordered in the right orientation and distance. The accurate assembly of reads using algorithms to reconstruct genomes is an essential task that forms the basis of current genomic studies. There are several assembly algorithms and these differ both in the type of data sets they use and implementation of algorithms.

Genome assembly approaches are grouped into two broad categories: comparative assembly and *de novo* assembly. Given most polyploid genomes have not been sequenced before, due to their complexity and size, *de novo* assemblers are the choice for sequencing such genomes. If a related species has been assembled, a *de novo* assembly is still undertaken and a comparative assessment of both genomes done after assembly. Using a related species as a reference for assembly introduces bias from the difference between the two species and also from assembly errors inherent in the reference assembly. Assembly algorithmic challenges include polymorphic regions and genes, repeat elements, false positive and false negative read assignments, poor sequence quality, bias in sequencing and uneven coverage during library preparation of DNA for sequencing. Such challenges have resulted in a number of published genomes being fragmented, incomplete and erroneous (Alkan *et al.*, 2010). Assembly validation and downstream analysis is thus made difficult.

Genome assembly algorithms are categorised into three groups; de Bruijn graph (DBG), greedy algorithms and overlap layout consensus (OLC). Some algorithms combine the above approaches or have variations of these, but most implementations vary based on how errors and read pair information is used during assembly. The DBG algorithm (Idury and Waterman, 1995) works by cutting reads into kmers that are used to build an assembly graph. Kmers are sub strings of a string of a specified length. For example, a read GCAT four base pairs long can be decomposed into two three sequence read fragments of length three, GCA and CAT. The k in kmers is used to designate the length of the strings. The DBG consists of kmers in reads represented as nodes or vertices and neighbouring kmers are represented with edges between them. Edges link nodes based on overlaps. A simple overlap between reads would thus be represented by a DBG with single edges between nodes. Where the reads are repetitive, the DBG would be represented with multiple edges between nodes with some of the edges forming circles or bubbles. The assembled sequence is inferred by traversing the shortest path through all the vertices. Several assemblers based DBG are available and these include Velvet (Zerbino and Birney, 2008), Euler-USR (Chaisson *et al.*, 2008), AllPath-LG (Gnerre *et al.*, 2011), SOAPdenovo (R., Li *et al.*, 2010) and ABySS (Simpson *et al.*, 2009).

Greedy algorithms are based on read extension of maximum overlaps between reads. Reads are represented as nodes and overlaps are edges in the assembly graph. All reads are compared pairwise and reads with the longest overlaps are merged. The longest overlaps are not always the correct ones, and some reads are thus not extended. Several algorithms based on the greedy assembly approach include SHARCGS (Dohm *et al.*, 2007), PE-Assembler (Ariyaratne and Sung, 2011), VCAKE (Jeck *et al.*, 2007) and SSAKE (R., L., Warren *et al.*, 2007).

OLC algorithms work by generating all possible overlaps within the reads supplied and then generates a consensus sequence that is the assembly. Several assemblers based on OLC are also available, and these include Phrap (la Bastide and McCombie, 2002), CAP3 (Schatz *et al.*, 2010), PCAP (X., Huang and Yang, 2002), Arachne (Batzoglou, 2002), Phusion (Mullikin, 2002), Newbler (Margulies *et al.*, 2005) and the Celera Assembler (Myers, 2000).

While OLC algorithms need to call consensus sequences from several sequence alignments, DBG algorithms do not as the neighbouring kmers already include consensus sequences. Because DBG algorithms use kmers instead of whole reads to build the assembly graph, DBGs are more memory efficient when assembling large genomes than OLC algorithms (Z., Li *et al.*, 2012). Using kmers for graph building prevents an all versus all read comparison. In addition, the assembly graph usually results in large and entangled graphs that are difficult to resolve when assembling repetitive genomes. Although OLC algorithms are not as memory efficient, they compare all reads during graph building generating a truer representation of the genome. Furthermore, DBG algorithms can only store a single instance of a unique kmer at a time. Kmers, however, often occur multiple times in a genome. Such kmers are thus left out during the assembly process when using DBG algorithms.

The outcome of an assembly will vary significantly depending on the assembly algorithm used. Assembly variation also results from the sequencing technology used and the nature of the genome being assembled (repeats, heterogeneity and size). Longer reads with fewer sequencing errors are ideal for the assembly process. Assembly algorithms have been shown to fall short of accurately assembling highly repetitive regions (Pevzner *et al.*, 2001). As a consequence, many genome assemblies consist of mis-assemblies (Salzberg and Yorke, 2005; Phillippy *et al.*, 2008). Several projects have attempted to benchmark assembly algorithms objectively. These include dnGASP (*de novo* Genome Assembly Project) [http://cnag.bsc.es/, GAGE (Genome Assembly Gold-standard Evaluations) (Salzberg *et al.*, 2012), and Assemblathon (Earl *et al.*, 2011; Bradnam *et al.*, 2013). The dnGASP and assemblathon approach used simulated reads to evaluate assembly algorithms while CAGE used bacterial, human and insect genomes. None of the projects have evaluated the performance of assembly algorithms on complex polyploidy plant genomes due to the complexity, size and high repeat content. Thus, a novel assembly algorithm SASSY was specifically designed to address this shortfall and tailored for the assembly of complex genomes (Kazakoff *et al.*, 2011) http://sassy.mikeimelfort.com.

The SASSY assembler (Kazakoff *et al.*, 2011) uses an improved OLC algorithm that exploits the advantages of DBG algorithms, as well as the accuracy of OLC algorithms. SASSY uses kmers to determine overlaps (a DBG approach) between reads and then generates an overlap graph from whole reads (an OLC approach) of the respective kmers. Furthermore, SASSY uses paired reads during its initial graph building stages. Given a read pair should be represented in the final assembly once, multiple instances of unique kmers from different reads can be stored, unlike DBG algorithms. Given repeats occur multiple times in a genome, especially in complex genomes such as wheat, SASSY was chosen as the preferred algorithm for assembly. We applied SASSY in this thesis to assemble the short arm of bread wheat chromosome 7D and sugarcane BACs.

1.5 Current assembly validation approaches

Once a genome has been assembled, validation of its accuracy is undertaken prior to further analysis. Completed genome assemblies have been shown to have false segmental duplications and collapsed repeats (Alkan *et al.*, 2010). False segmental duplications in assemblies have been attributed to the presence of heterozygous sequences from two haplotypes which when assembled result in separate contigs (Kelley and Salzberg, 2010), while collapsed repeats in assemblies have been partially attributed to short reads that do not span repeat regions.

There are two main approaches in validating a genome assembly. Reference based and non reference based approaches. Reference based approaches depend on the existence of a previously assembled genome, or a genome assembly of a closely related organism. Non reference based approaches are used where no appropriate reference genome is available for comparison. Evaluating the accuracy of *de novo* genome assemblies is challenging, as errors are determined based on the existence of an accurate reference genome, which is unavailable. Validating assemblies using reference-based approaches does not offer an objective evaluation due to assembly errors inherent in both assemblies in addition to differences that may be present between organisms when a closely related organism's genome assembly is used. Furthermore, different assembly algorithms have been shown to produce different assemblies with the same dataset (Earl *et al.*, 2011; Bradnam *et al.*, 2013). This confounds validation if the genomes were assembled using different assembly algorithms.

Generally, various assembly metrics are used to give an estimation of the accuracy of an assembly. These include the number of contigs and scaffolds in the assembly, where fewer are better. The longest, shortest, mean and median contig lengths. The larger lengths are considered better. The N50 of an assembly, defined as the length of which all contigs of that length or longer contain at least half of the total assembly size. A higher N50 is considered better. A variation of the N50 metric is the normalized N50 (Mäkinen *et al.*, 2012). The normalized N50 is an indicator of the assembly contiguity based on the known genome size rather than the total assembly size. The CC50 (correct contiguity) (Earl *et al.*, 2011) has also been used as an assembly evaluation metric and is a measure of the long-range continuity of the assembly. Higher values of the normalized N50 and CC50 are also considered better.

Published assembly evaluations methods have relied on the above metrics to evaluate genome assemblies. Evaluations such as Assemblathon 1 (Earl *et al.*, 2011; Bradnam *et al.*, 2013) used small simulated genomes to evaluate assemblies produced by 23 different assemblers. This evaluation showed that no single assembly algorithm greatly outperformed others. Conclusions from the Assemblathon 1 evaluations were unsatisfactory, as these may not apply to larger genomes with complex repeat structures. The GAGE evaluation (Genome Assembly Gold-standard Evaluations) (Salzberg *et al.*, 2012) partly addressed limitations of the Assemblathon 1 evaluations by using real datasets, but avoided large genomes. The dnGasp evaluation (http://cnag.bsc.es/) used both real and simulated data from large genomes while the Assemblathon 2 evaluated large genomes of snake, bird and fish (Bradnam *et al.*, 2013). Despite the above evaluations of genome assemblies and assemblers, evaluation of assemblies of large complex polyploid plant genomes has been lacking. Therefore, novel non reference based and comparative assembly approaches are required for the evaluation of complex polyploid plant genomes.

While assembly metrics do give an evaluation of how good an assembly is, they do not give an indication of how true on assembly is compared to the actual genome in terms of the assembly size, assembly contiguity, assembly accuracy at the sequence level and the gene content. Other approaches used for the evaluation of these aspects of a genome assembly are discussed in more detail below.

1.5.1 Assembly size

Genome sizes are used to evaluate the completeness of an assembly, especially in *de novo* assembly. The genome size of an organism can be estimated based on kmer analysis, flow cytometry or gel electrophoresis. To estimate a genome's size using kmers, the genome is sequenced to sufficient depth, and a kmer counting tool is used to estimate the kmer distribution frequency. Several kmer counting tools are available. These include Tallymer (Marcais and Kingsford, 2011), Jellyfish (Kurtz *et al.*, 2008), BFCounter (Melsted and Pritchard, 2011), DSK (Rizk *et al.*, 2013), KMC (Deorowicz *et al.*, 2013), Turtle (Roy *et al.*, 2014) and KAnalyze (Audano and Vannberg, 2014). A kmer size for the estimation of the genome size is determined from a frequency distribution based on the uniqueness ratio. The uniqueness ratio is the ratio of kmers occurring exactly once relative to all kmers in the sample to that of all kmer sizes (Kurtz *et al.*, 2008). The inflection point of such a distribution represents the optimal kmer size that would give the best resolution and highest information content. The genome size is then estimated using the following formulae (J., Kim *et al.*, 2014);

The genome coverage depth is calculated first:

kmer coverage depth × ***(average read length)*** × ***(average read length - k-mer size + 1)*** $^{-1}$.

This is followed with the calculation of the genome size as:

Genome size = (total bases) x (genome coverage depth) $^{-1}$.

The kmer coverage depth in the above formulae is established by taking the maximal peak of a kmer distribution plot of kmer depth vs kmer frequency. Kmer analysis has been used successfully to estimate the genome sizes of several *de novo* assemblies such as those of the Pacific oyster (Gerdol *et al.*, 2015) and giant panda (R., Li, Fan, *et al.*, 2009). DNA extraction and NGS library preparation methods, sample contamination, heterozygosity of an organism, its physical sequence coverage, depth of coverage and sequencing errors contribute to counts of unique kmers often leading to an over-representation of non-true kmers in a sample and consequently larger or smaller genome size estimates (Qingpeng Zhang *et al.*, 2014).

Small DNA fragments sizes (10 bps - 40,000 bps) can be determined using gel electrophoresis. The samples are loaded into wells on a gel. A DNA ladder is also loaded whose molecular size is known. An electric pulse is then applied to the gel enabling the migration of DNA molecules through the gel.

DNA fragments will separate through the gel based on their sizes. The actual size of a fragment can be estimated based on its position in the gel relative to that of a DNA ladder as a reference. Pulse Field Gel Electrophoresis (PFGE) is a variation of gel electrophoresis, where much larger DNA fragments can be separated (Cantor *et al.*, 1988; Schwartz and Cantor, 1984). Standard gel electrophoresis cannot adequately separate large DNA molecules, as fragment migration across a gel is independent of the DNA fragment size. Rather, fragment size is dependent on the net charge and frictional drag from the structure of the DNA fragment as it moves through the gel (Cantor *et al.*, 1988).

Flow cytometry is regarded as the standard method for the estimation of genome sizes of large DNA fragments (DeSalle *et al.*, 2005; J., Kim *et al.*, 2014). To estimate the size of a DNA fragment using flow cytometry, the fragment is first stained using a bisintercalating dye. Bisintercalation refers to the alternate binding of a dsDNA molecule at regular alternate positions. Bisintercalating dyes have been shown to bind to every second nucleotide in a dsDNA (Glazer and Rye, 1992). Such dyes when fluorescently tagged enable the accurate determination of the size of a DNA fragment based on intensity of fluorescence emission, as this is proportional to the dye bound nucleotides (Glazer and Rye, 1992).

The estimation of DNA fragment sizes using kmer methods provides rough estimates. Where the DNA may have a higher repeat content, heterozygous, or sequencing errors, increased sequencing depth increases errors (B., Liu *et al.*, 2013). Estimated sizes have been shown to differ from actual fragment sizes (B., Liu *et al.*, 2013). PFGE has been shown to have a 10% uncertainty (Guo *et al.*, 1992) in size estimation compared to flow cytometry at 2% uncertainty (Z., Huang *et al.*, 1996).

1.5.2 Genome co-linearity and gene content

Previously established genome co-linearity with closely related organisms has been used to evaluate, study and improve genome assemblies (International Rice Genome Sequencing Project, 2005; Paterson *et al.*, 2009; Schmutz, Rokhsar, *et al.*, 2010; Berkman *et al.*, 2013; Berkman *et al.*, 2011; Berkman *et al.*, 2012). Genome co-linearity between two species refers to stretches of sequence blocks or genes that are conserved across related species. Approximately 80% of the genes in *Brachypodium* and rice have been shown to be in conserved (Wicker *et al.*, 2010). A slightly lower estimate has been observed between *Ae. Tauschii,* rice and sorghum (M., C., Luo *et al.*, 2009).

Although genome co-linearity is an integral part of comparative studies where better characterized organisms are available, genome co-linearity detection methods are varied and produce different results (Altenhoff *et al.*, 2015). Genome co-linearity detection methods allow comparisons of diverse genomes by allowing a degree of variation between genomes of organisms during the detection and visualization of conserved blocks (syntenic blocks). Variation is a result of recombination, rearrangements (inversions, translocations and deletions), segmental duplications, horizontal transfer and whole genome duplication. Identified blocks are further evaluated by the order in which they appear in the compared genomes. By allowing for variation, high-level interspecies similarities can be studied given differences between individuals due to evolutionary and or sequencing and assembly errors are eliminated.

Genome co-linearity can be missed due to poor assemblies especially when comparing complex highly repetitive genome assemblies that often remain highly fragmented and incomplete at the end of genome sequencing projects (E., Bao *et al.*, 2014). In studying angiosperm evolution, recurring polyploidy, gene rearrangements, gene loss, repeat elements and differences in genome sizes further confound analysis (Bowers *et al.*, 2003). For example, *Arabidopsis thaliana* is a model organism for comparative studies of angiosperms due to its relatively small and well-characterized genome (The *Arabidopsis* Genome Initiative, 2000). However, the *A. thaliana* genome has undergone two whole genome duplications, (Bowers *et al.*, 2003), one whole genome triplication (Jaillon *et al.*, 2007) and 9 to 10 chromosomal rearrangements. Comparisons between *A. thaliana* and other angiosperms are thus complicated.

Gene and genome co-linearity has been well documented in plants despite the remarkable differences in size, and whole genome duplication events observed even within closely related species (Bennett and Smith, 1991; Bowers *et al.*, 2003). Chromosome numbers however, have been shown to be relatively static (Bennett and Smith, 1991), but whole genome duplications and gene loss have fractionated conserved genes (homologues) across related chromosomes (Bowers *et al.*, 2003). Homologues are categorised as orthologues and paralogues. Orthologues are two genes in different species that originated from a single gene in an ancestral species. Paralogues are genes that diverged through gene duplication within a species and can be orthologous to their unduplicated orthologue in the related species (Fitch, 1970; Sonnhammer and Koonin, 2002; Catchen *et al.*, 2009). Despite these challenges, gene count and order have been widely used as an indication of assembly quality (Mayer *et al.*, 2009; International Rice Genome Sequencing Project, 2005; Paterson *et al.*, 2009; Schmutz, Rokhsar, *et al.*, 2010).

The estimation of the total number of genes in a genome as an evaluation of the assembly is difficult and biased. Annotation errors, gene fragments from transposon activity among other factors (Bennetzen *et al.*, 2004; N., Jiang *et al.*, 2004; J., Lai *et al.*, 2005; Morgante *et al.*, 2005; Paterson *et al.*, 2009) lead to inaccurate gene count estimates. Analysis of chromosome 3B (Choulet *et al.*, 2010) showed that despite the expected conserved co-linear blocks between rice, sorghum and *Brachypodium*, a high number of homologues were outside co-linear blocks and were linked to transposon activity (Wicker *et al.*, 2011).

Comparison of *de novo* assemblies to related organisms offers inaccurate insights into the quality of an assembly and errors are likely to be propagated. Given gene estimates continue to increase with the improved quality of genome assemblies, and the complete repertoire of genes is yet to be determined, the use of gene content and genome co-linearity as an evaluation of the quality of a *de novo* assembly is thus highly biased as only the known genes and established synteny is used for evaluation. Highly accurate and contiguous assemblies are thus needed to identify accurate and complete gene conservation across related species.

1.5.3 Long range genome contiguity using physical maps

Physical maps are crucial in understanding the overall relationships between functional elements and structural variation detection in a genome. Such an understanding relies on accurate long-range contiguity of an assembly. Genome assemblies are usually highly fragmented. Gaps still exist in finished assemblies despite the use of paired-end and mate-pair reads to order, orientate and bridge contigs into scaffolds. Structural variations such as inversions, duplications and copy number variations are usually missed without the use of physical maps. In addition, genome maps help to further order scaffolds and identify gaps during genome finishing. There are several types of physical genome maps in use. Such maps are usually developed from an independent platform to that used for sequencing so as to both complement and validate sequencing and assembly.

Several sequencing projects have employed the use of physical maps for genome finishing including maize (S., Zhou *et al.*, 2009; Schnable *et al.*, 2009), *Arabidopsis* (The *Arabidopsis* Genome Initiative, 2000) and rice (International Rice Genome Sequencing Project, 2005). Several approaches have been used to generate physical maps. These include whole genome profiling (WGP) (Philippe *et al.*, 2012; van Oeveren *et al.*, 2011), optical mapping (Schwartz *et al.*, 1993; Teague *et al.*, 2010), hybridization (Singh *et al.*, 2006), radiation hybrid maps (Kumar *et al.*, 2015) and fingerprinting (Ding *et al.*, 2001;

M., Luo *et al.*, 2003; W., M., Nelson *et al.*, 2007; Z., Xu *et al.*, 2004; Marra *et al.*, 1997; Marra *et al.*, 1999; L., Lin *et al.*, 2010; Ragupathy *et al.*, 2011; Xiaojun Zhang *et al.*, 2009).

1.5.4 Restriction mapping of BAC clones

Fingerprinting is commonly applied to BAC clones and involves the digestion of BAC clones with restriction enzymes. Resulting restriction profiles are then compared and overlaps between fingerprints identified (Chen *et al.*, 2002; McPherson *et al.*, 2001). The SNaPshot approach was the most accurate and high-throughput approach for building physical maps (W., M., Nelson *et al.*, 2007; M., Luo *et al.*, 2003). SNaPshot was used to generate the maize and wheat chromosome 3B physical maps (Paux *et al.*, 2008; Wei *et al.*, 2007).

SNaPshot uses five restriction enzymes with fluorescent labels to digest DNA fragments. BAC clone DNA is first isolated in a 96-well plate and digested with five restriction endonucleases. Four of the restriction endonucleases recognize 6bps and generate 3' overhangs at the cut sites. The other endonuclease recognizes a 4bps motif and produces blunt ends. The four 3' end overhangs are then labelled with different fluorescent dyes. Five restriction enzymes offer high density and resolution during construction of the physical map. The restriction fragments are then sized using a capillary DNA analyser to produce fingerprints. The fragment fingerprints are then compared to re-generate overall restriction profile of the BAC clones and of overlapping clones. This is done by the use of the fingerprinted contigs (FPC) computer program (Soderlund *et al.*, 1997; Soderlund *et al.*, 2000).

Several limitations have hindered the detailed analysis of structural polymorphism in genomes using restriction mapping. These include the labour intensity and cost of restriction mapping when applied to large complex genomes and the short reads characteristic of NGS technologies which have been shown to be inadequate for capturing the structure of repetitive regions where most of a genome's variation lies (Lupski, 2007). The resolution of restriction mapping is approximately 8kb (Kidd *et al.*, 2008). Given the fragment size limitation in BAC cloning, inaccurate sizing of DNA fragments and challenges in resolution of highly repetitive regions when considering the order and number of repeats in a genome, high density maps are needed.

1.5.5 Optical mapping

Optical maps offer ordered physical maps devoid of amplification, cloning or hybridization errors. In addition, physical maps of large and complex eukaryotic genomes can be generated (Hastie *et al.*, 2013; Teague *et al.*, 2010; Zody *et al.*, 2006; S., Zhou *et al.*, 2007; S., Zhou *et al.*, 2009; Church *et al.*, 2009). The lack of amplification, cloning and hybridization errors makes optical maps suitable for validation and further improvement of genome assembly contiguity (Jing *et al.*, 1998; Z., Lai *et al.*, 1999; S., Zhou, 2003; J., Lin, 1999). Optical maps however have several limitations including high error rates resulting from DNA fragment sizing, uneven DNA stretching and low resolution at about one restriction site at every 20kb (Hastie *et al.*, 2013). Advances in optical mapping have addressed the above limitations through techniques that measure and stretch DNA fragments more accurately through the use of micro-channels, nanochannels and fluid dynamics (Chan, 2004).

One such system using nanochannels is the Irys System $^{\mathrm{TM}}$ by BioNano Genomics (bionanogenomics.com). The Irys System$^{\mathrm{TM}}$ employs the use of 45 nm wide nanochannel arrays implemented on non-fluidic chips (Tegenfeldt *et al.*, 2004). When generating a genome map using the Irys System $^{\mathrm{TM}}$, the DNA sample is prepared using IrysPrep Reagent Kits $^{\mathrm{TM}}$. Sample preparation involves the extraction of high molecular weight dsDNA. The dsDNA molecules are then nicked on one strand at specific 7 bps recognition sites and then hybridized with fluorescently labelled nucleotides (Alexa 546 dUTP) (Xiao *et al.*, 2007; Das *et al.*, 2010). Unique long-range restriction patterns are thus created. The DNA sample is then pipetted onto an IrysChip$^{\mathrm{TM}}$ containing nanochannels (Reisner *et al.*, 2005). The dsDNA molecules are linearized as they are drawn into the nanochannels. The resulting linearized molecules are free from folds and overlaps as DNA molecules confined in the nanochannels are unable to fold (Tegenfeldt *et al.*, 2004; Reisner *et al.*, 2005). The Iryschip$^{\mathrm{TM}}$ is then scanned using the Irys instrument. Multiple nanochannels are scanned simultaneously, with the fluorescent signals recorded. The Iryschip$^{\mathrm{TM}}$ is then flushed, and more dsDNA molecules loaded for scanning. Each of the dsDNA's fluorescent signal is then compared and ordered to generate overlaps and consensus maps using proprietary assembly algorithms. The final consensus maps can then be compared to a reference genome assembly where one is available or with other consensus maps using the Irys visualization tool IrysView $^{\mathrm{TM}}$.

Visualization of the genome maps and comparisons with other genome maps enables the identification of structural variation between organisms, validation of assemblies and anchoring of contigs. Optical mapping offers better resolution than restriction mapping as genomic DNA is not fragmented as a result digestion with restriction enzymes. As a result, the physical order of restriction sites is maintained further improving accuracy and long range contiguity of the optical maps.

The Irys system has been used to study the human adenovirus and several strains of human rhinoviruses (Xiao *et al.*, 2007). The Irys system has also helped resolve the complex structure of the DUF1220-rich 1q21 region of the human genome (O'Bleness *et al.*, 2014), analysis of structural variation in the human genome (Cao *et al.*, 2014) and *de novo* assembly of *Aegilopsis tauchii* (Hastie *et al.*, 2013). The use of the optical maps for assembly scaffolding and error detection was demonstrated in *Aegilopsis tauchii*, where scaffolds were ordered into super scaffolds several megabases long (Appels *et al.*, 2015). Due to the high accuracy and long range contiguity offered by the Irys System^{TM}, a genome map for bread wheat chromosome 7DS (publication in print) was developed and used in the evaluation of chromosome 7DS BAC-by-BAC assembly in this thesis.

1.6 Summary

Cereal crop improvement is critical in order to meet the ever-increasing global food demands. Genome assemblies play a significant role in the identification of markers that can be used for crop improvement. The generation of complete and accurate genome assemblies is an active ongoing research area driven by evolving sequencing platforms and assembly algorithms. The large and complex genome of bread wheat presents a significant challenge for sequencing and assembly. Of the currently available sequencing platforms, no single platform has yet been evaluated on complex highly repetitive genomes (Visendi *et al.*, 2013). Pacific Biosciences offers longer reads with reduced per base accuracy (Carneiro *et al.*, 2012) in addition to no amplification during library preparation and thus a lack GC bias during sequencing. High coverage is however required for error correction by consensus base calling (Roberts *et al.*, 2013). Ion Torrent does offer longer reads than Illumina, but with lower per base accuracy due to phasing issues. Illumina sequencing platforms currently offers the most accurate platform and lowest per base cost for *de novo* genome sequencing despite shorter read lengths (Loman *et al.*, 2012). For sequencing approaches, BAC-by-BAC sequencing of isolated chromosome arms is the best approach despite the intensive labour and sequencing costs associated. In the absence of a single molecule third-generation sequencing platform that would offer very long reads alleviating the need to assemble, short high accuracy reads that need to be assembled and validated will suffice.

Following sequencing and assembly, evaluation of the accuracy and completeness of such assemblies is important. Although the evaluation of the accuracy of a genome assembly remains largely unclear

(Bradnam *et al.*, 2013), in these thesis, we present several approaches that when combined have enabled us to evaluate the 7DS wheat chromosome arm assembly.

This thesis presents novel sequence assembly approaches, and assembly evaluation methods. These were employed to accurately sequence, assemble and validate the bread wheat 7DS chromosome arm. These approaches are currently being applied to sequence other wheat chromosome arms.

2 Methodology for the sequencing and assembly of chromosome 7DS using a BAC-by-BAC approach

Given the size and complexity of the wheat genome, the generation of an accurate and complete genome assembly is a challenge and costly. By combining several approaches that reduce the genome complexity, such as chromosome sorting (see 1.4.1), and improve the resolution and accuracy of the genome assembly by using BAC-by-BAC sequencing approaches (see 1.4.3), we sought to establish a high-throughput and affordable method for the sequencing and assembly of complex polyploid genomes. We demonstrate this approach in the sequencing and assembly of BAC clones of the short arm of wheat chromosome 7D. We optimised this method using several approaches. For sequencing, we used the most accurate sequencing platform at the project initiation phase (Illumina).

2.1 Materials and methods

2.1.1 The selection of an appropriate genome assembler for wheat BACs

Different genome assemblers result in varied genome assemblies based on the underlying algorithm implementation, parameter settings used to run the assembly, the sequence coverage used, and the complexity of the genome being assembled (Bradnam *et al.*, 2013). Given the nature of genome sequencing projects, sequence datasets are prone to contamination, sequencing errors and genomic repeats that confound assembly. Recommendations on the selection of a genome assembler from the Assemblathon 2 project (see 1.4.4 for a detailed review) proposed the selection of an assembler based on the complexity of the genome to be assembled (Bradnam *et al.*, 2013). Given the large number of genome assemblers available, only the SASSY assembler (http://sassy.mikeimelfort.com) (Kazakoff *et al.*, 2011). has been specifically developed and tested for the assembly of complex highly repetitive plant BAC sequences. Furthermore, a comparison of all assemblers is not practical and would require each assembler's parameters to be first optimized for a particular assembly before the assemblies can be compared. Thus two assemblers; 1. Velvet, the most commonly used *de Bruijn* based short read assembler (Zerbino and Birney, 2008) for plant genomes and Ray, a recent parallel *de Bruijn* based assembler shown to be suitable for assembly of metagenomic or mixed samples (Boisvert *et al.*, 2010) were evaluated against SASSY, an improved overlap consensus assembler customised for BAC assemblies to determine their suitability for assembling wheat BAC clones.

To assess the three genome assemblers, seven non-overlapping single wheat BACs (A, B, C, E, F, G and H) were sequenced to high coverage (> 500 x) by Jacqueline Batley's lab. The seven wheat BACs were first filtered for *Escherichia coli* (*E. coli)* str. K-12 and vector sequences using Perl scripts and the short read mapper SOAP2 (R., Li, Yu, *et al.*, 2009). *E. coli* and vector filtered reads were then assembled with Velvet, Ray and SASSY on a Linux cluster with five CPUs and 23 Gb of memory. The assemblies were optimised by assessing the coverage and assembly parameters. Paired reads for each of the seven wheat BACs were randomly selected using a Python script to generate subsets. The subsets corresponded to approximately 15x, 30x, 90x, 180x, 270x, 360x, 450x and > 500x coverage for each of the seven BACs and were assembled with Velvet, Ray and SASSY. Velvet assemblies were optimised using VelvetOptimiser, a wrapper that determines the optimal assembly parameters for Velvet (Zerbino and Birney, 2008), whereas Ray assemblies were optimised using Ray's permissible k-mer values of 15, 18, 21, 24, 37, 30 and 32. SASSY assemblies were run with default parameters and only > 500 x sequence coverage, because SASSY has been shown to require high sequence coverage (http://sassy.mikeimelfort.com) (Kazakoff *et al.*, 2011). The assembly N50, number of contigs, longest contig sizes and total assembly size of the selected optimal assemblies of Velvet, Ray and SASSY were compared. Assembly N50 is a parameter used to determine the contiguity of an assembly. It is defined as the length for which all contigs of that length or longer contain at least half of the total assembly length.

2.1.2 Determination of optimal sequencing depth for BAC assemblies

Due to the repetitive nature of the wheat genome and its probable effect on sequence coverage and assembly, sugarcane was selected to determine optimal sequencing depth for assembly when using SASSY as it has a similar repeat content and due to availability of sugarcane BACS sequenced to high coverage (>10,000 x). The sugarcane genome has higher ploidy (~12 x) (D rsquo Hont, 2005) than wheat. Eleven individual sugarcane BACs (14D11, 163F13, 206K8, 213A5, 220G15, 222M11, 222M13, 241K5, 246N21, 256O12 and 44O7) from a BAC library of the sugarcane cultivar R570 (Tomkins *et al.*, 1999) were selected and sequenced to between 4,000 x and 10,000 x coverage, with an average insert size of 130 Kb. The sequenced read pairs were then split into subsets representing

200 x - 3,000 x coverage in 100 x increments. This resulted in 319 subsets, which were assembled using SASSY on a Linux cluster with five CPUs and 23G of memory.

Perl and BASH scripts were used to automate and enable multiple runs of assemblies. *E. coli* str. K-12 sequences were filtered using Perl scripts prior to assembly using the short read mapper SOAP2 (R., Li, Yu, *et al.*, 2009).Reads that were filtered out were selected based on exact matches to *E. coli* based on SOAP2 parameters -M 0 -r 1 -v 0. Statistics were calculated on reads incorporated in the assembly, assembly N50, number of contigs, longest contigs and total assembly length. A count of the number of unique reads in the dataset was done using the k-mer counting tool DSK (Rizk *et al.*, 2013), with the k-mer length set to 100 bp. Based on the assembly statistics, a model to predict the minimal coverage depth using local regression (LOESS) (Jacoby, 2000) was developed using the R statistical package (R version 3.0.2). LOESS is a robust non-parametric regression technique implemented in R version 3.0.2. LOESS fits linear regressions over a subset of localised data points while limiting over-fitting of data points. For modelling, datasets with coverage between 200 x and 1000 x were selected to reduce the influence of outliers. The resulting LOESS regression model had a smoothing parameter of 0.75, degrees of freedom (df) 2 and was based on 91 observations. Plots of sequence coverage versus assembly length were generated using the R package gg2plot (Wickham, 2009).

2.1.3 Determination of a BAC pooling strategy

A BAC pooling strategy was necessary due to the high sequencing depth required for assembling BACs with SASSY. This demanded the evaluation of the accuracy of individual BACs when assembled in pools. Seven non-overlapping single wheat BACs (A, B, C, E, F, G and H) were sequenced to high coverage (> 500 x). Simulated pools were made by combining sequenced reads of the single BACs, resulting in pools of up to seven BACs (i.e. A, AB, ABC, ABCE, ABCEF, ABCEFG and ABCEFGH). Simulated pools of four BACs were also generated by combining sequenced reads of single BACs, resulting in four pools of four BACs (i.e. ABCE, BCEF, CEFG, EFGH). The resulting pools were then assembled with SASSY using default parameters. Assembled contigs of single BACs and the corresponding pooled BACs were compared to evaluate the percentage of sequence identity. Pooled BAC assemblies were further evaluated using BAC end sequences (BES). The BES for each of the BACs were sequenced by Hana Simkova's lab using the Sanger sequencing platform. Sequenced BES were base called using the Phrep software version 071220 (Ewing *et al.*, 1998). BES mappings onto assemblies were performed using the Basic Local Alignment Search Tool (BLAST) sequence mapping software (Altschul *et al.*, 1990).

To evaluate the assembly accuracy of contigs using BES mapping positions, while accounting for multiple mapping positions due to repeats, it was necessary to weight BES mapping positions with contig edges. This was as a result of BES having multiple best mapping positions on the assemblies. For any set of contigs (e.g. ***X, Y, N***) with multiple hits of a single BES ***B,*** with ***B*** being either a forward or reverse BES with BLAST hits with bit scores ($\boldsymbol{h_1, h_2... h_n}$) occurring randomly on the contigs (***X, Y, N***), all contigs had to be of lengths greater than the BES ***B*** and the BES ***B*** > 120 bp. For each hit ($\boldsymbol{h_1, h_2... h_{en}}$) there are two distances ($\boldsymbol{d_1, d_2}$) representing distances from the left edges $\boldsymbol{d_1}$ and right edges $\boldsymbol{d_2}$ of contigs (***X, Y, N***). The correct location and position of the BES ***B*** on either contig ***X***, ***Y*** or ***N*** is determined by the highest score (***S***) of all the hits. ***S*** is obtained by subtracting the shortest distance ($\boldsymbol{d_s}$) (which is the least of either $\boldsymbol{d_1}$ or $\boldsymbol{d_2}$ for each hit) from the hits' bit scores. BLASTN bit scores were used as they incorporate both percentage of identity and aligned length. The above algorithm was implemented in Perl. BES mapping positions on contigs were visualised using R scripts implemented with the R software package ggbio (Yin *et al.*, 2012). Contigs from single BAC assemblies were compared to contigs from pooled assemblies using the dot plot tool MUMmer (Kurtz *et al.*, 2004).

2.1.4 Optimisation of SASSY pooled BAC assemblies

The effects of varying library preparation kits and SASSY parameters on BAC assemblies were explored. By comparing assemblies of the same BACs that were sequenced using different library preparation kits and assembled with different SASSY parameters, the optimal combination of library preparation kit and assembly parameters was determined. Jacqueline Batley's lab sequenced 17 pools of four wheat chromosome 7DS BACs on the Illumina Hiseq 2000, using the three Illumina DNA library preparation kits TruSeq, Nextera and Nextera XT. The kits are designed by Illumina for use on their sequencing platforms. Six pools were prepared for sequencing using the Nextera kit, seven pools using the TruSeq kit, and four pools using the Nextera XT kit. Sequenced reads from each of the 17 pools were filtered for *E. coli* and vector sequences using Perl scripts and the short read mapper SOAP2, and assembled at read lengths of 61 bp, 71 bp, 81 bp, 91 bp and 100 bp. The resulting 85 assemblies were evaluated to determine the optimal kit and assembly parameters. Assembly metrics were based on the number of contigs, longest contig size, total assembly size and assembly N50.

The SASSY manual (http://sassy.mikeimelfort.com) recommends three sets of parameters for the assembly of low, medium and high sequence coverage. These parameters are the naive offsets and extension offsets (2, 10 for low coverage; 8, 16 for high coverage; and defaults of 4, 12). An offset is defined as maximum allowable length (bp) of the non-overlapping parts of two overlapped reads. Naive offsets are offsets set during the initial steps of the assembly where contigs are first generated. Extension offsets are lengths set for the contig extension phase of the assembly algorithm. To establish the optimal assembly parameter for naive offsets and extension offsets, 96 pools of four BACs from bread wheat chromosome 7DS minimal tiling path (MTP) (Šimková *et al.*, 2011) were sequenced using the TruSeq DNA library preparation kit. The sequenced pools had an average coverage per BAC of 690 x and were assembled with three sets of assembly parameters (2, 10), (4, 12) and (8, 16). The resulting assemblies were evaluated based on the average of assembly length, number of contigs, and N50, as well as comparisons to the MTP size estimate. Statistical analysis was performed using the ANOVA and paired t tests implemented statistical packages in R version 3.1.2.

2.1.5 Automation of bread wheat chromosome arm 7DS BAC assembly

The bread wheat chromosome arm 7DS MTP consists of 4,608 BAC clones. This necessitated the development of an automated assembly process. The 7DS BAC library was constructed, BACs fingerprinted, and contigs assembled using FPC software by Hana Simkova's lab (Šimková *et al.*, 2011). This resulted in an MTP of 4,608 BACs ordered into 931 MTP contigs. The BAC sequencing strategy involved sequencing 96 pools indexed on a single lane. Each pool consisted of four non-overlapping BACs sequenced on the Illumina HiSeq 2000 using the TruSeq library preparation kit. The resulting paired end reads were assembled using SASSY. The 4,608 BAC clones were pooled into 1,152 pools, which were indexed and sequenced in 96 well plates. This resulted in the sequencing of a total of 12 plates. The assembly process was automated by developing an assembly pipeline (SASSY pipeline) using Perl scripts. The SASSY pipeline involved several steps (**Figure 2-1** below).

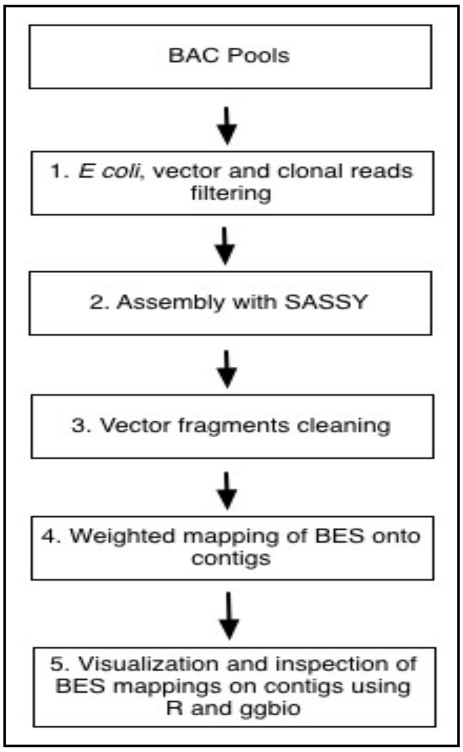
BAC Pools
1. *E coli*, vector and clonal reads filtering
2. Assembly with SASSY
3. Vector fragments cleaning
4. Weighted mapping of BES onto contigs
5. Visualization and inspection of BES mappings on contigs using R and ggbio

Figure 2-1: SASSY assembly pipeline steps.

2.1.6 Mate pair scaffolding of assembled BAC pools

To further order and scaffold BAC assemblies, mate pair (MP) reads were used. To generate the MP libraries, DNA from pooled BACs was used. The MP libraries were prepared by Jacqueline Batley's lab using the Nextera MP kit with insert sizes 6 - 10 kb. Pools were indexed to run on 12 lanes on the Illumina HiSeq 2000, corresponding to the 384 well (see 2.1.5). The resulting sequences were filtered to remove adapter contaminants. The MP reads were processed using the NxTrim tool (O'Connell *et al.*, 2014) to remove sequencing and junction adapters introduced during MP library preparation (see 1.3.1.2). External adapters were first removed using Cutadapt version 1.7. (Martin, 2011) prior to using NxTrim, as the NxTrim software does not remove external adapters. Cutadapt has been noted as having the highest specificity and sensitivity in the identification of adapter sequences when compared to other available tools (O'Connell *et al.*, 2014), but it does not pre-process junction adapters. Only reads longer than 35 bp were retained post adapter filtering.

To evaluate NxTrim's applicability in optimal pre-processing of MP libraries, processed MP reads were mapped to their respective BAC pools using BLAST (Altschul *et al.*, 1990). To scaffold respective pools, only read pairs that mapped with 100% identity across the full length of the mapped adapter filtered read and within the same pool were used. Pools were scaffolded using SSPACE

(Boetzer *et al.*, 2011) with scaffolding parameters k (number of links) and insert size determined based on observed mappings of adapter filtered MP reads to assemblies.

2.1.7 Generation of chromosome arm 7DS super scaffold

The high repeat content characteristic of the wheat genome necessitated the development of an approach for the reconstruction of the 7DS super scaffold to high accuracy. This approach utilises the MTP of overlapping BACs, their respective BES and overlaps between BACs coupled with manual curation. The 7DS MTP consisted of 931 MTP contigs, each with between one and 40 overlapping BACs. Assembled BAC pools for all plates 1 - 12 and all 4,608 specific BAC BES were used. Each BAC had both a reverse and forward BES.

A super scaffold was generated using a combination of overlaps of BAC assemblies and BES within a MTP contig. Mappings of BES to their respective pools during the assembly process were used to assign contigs within a pooled assembly to the corresponding BAC (see step 4 in **Figure 2-1**). The super scaffold was then reconstructed in a multi-step process. Using two BACs **A** and **B** as an example (where BACs **A** and **B** overlap in the MTP and had been assembled in pools EF**A**G and H**B**JK), contigs from the assembled BAC pools EF**A**G and H**B**JK were compared using BLAST (Altschul *et al.*, 1990) to identify overlaps. Contigs sharing overlaps were merged. BES were used to assign contigs in pools EF**A**G and H**B**JK to their corresponding BACs **A** and **B**. Contigs belonging to the same BAC identified by correctly paired BES were joined with 100 Ns if these had not been joined due to overlaps. This was done with manual curation for all 931 MTP contigs consisting of 4,608 BACs.

2.2 Results

The results presented in this section follow the order of the materials and methods presented in section 2.1 above.

2.2.1 A suitable genome assembler identified for wheat BACs

Assemblies of seven wheat BACs (A, B, C, E, F, G and H) using the Ray assembler at sequence coverage between 15 x to > 510 x at k-mer values 15 - 32 were analysed. Within the Ray assemblies, the assembly with highest average N50 (33 Kb; Table 2-1) was achieved at a sequence coverage of 180 x and a k-mer value of 30 (Figure 2-2).

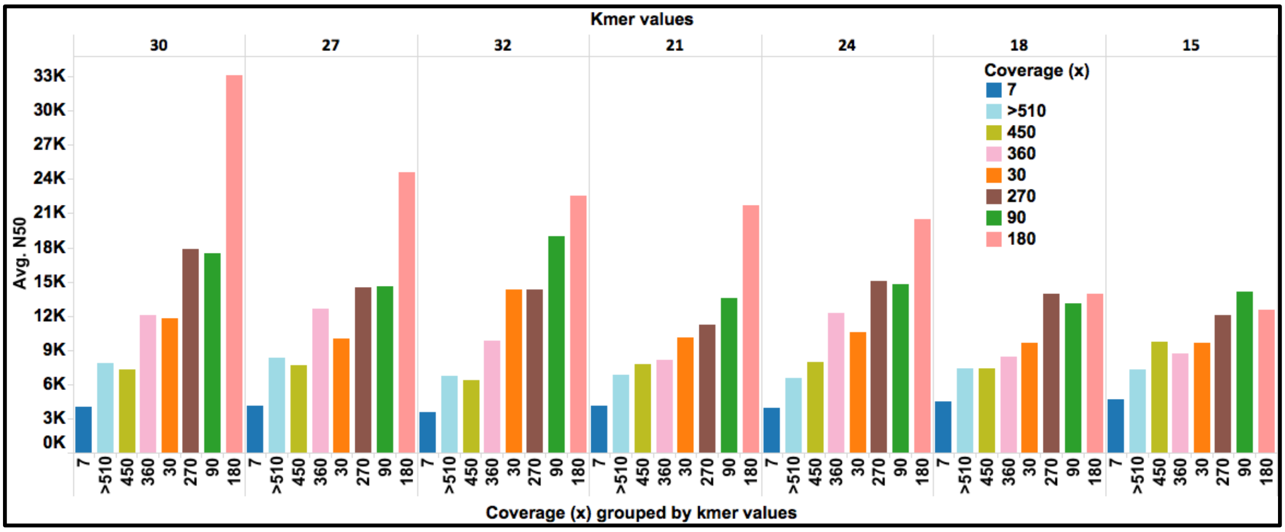

Figure 2-2: Ray assembly average N50 by coverage grouped by k-mer for seven wheat BACs.

Table 2-1: Ray assembly average N50 (Kb) by coverage 15 x to > 510 x and k-mer values 15 to 32 for seven wheat BACs.

	K-mer						
Coverage	**15**	**18**	**21**	**24**	**27**	**30**	**32**
15x	0.6	0.6	0.5	0.5	0.5	0.4	0.4
30x	9	9	10	10	10	11	14
90x	14	13	13	14	14	17	19
180x	12	14	21	20	24	**33**	22
270x	12	13	11	15	14	17	14
360x	8	8	8	12	12	12	9
450x	9	7	7	7	7	7	6
>510x	7	7	6	6	8	7	6

Assemblies of seven wheat BACs (A, B, C, E, F, G and H) using the Velvet assembler at sequence coverage between 15 x to > 510 x were analysed (Table 2-2). Within the Velvet assemblies, the assembly with the highest average N50 (33 Kb; Table 2-2) was achieved at a sequence coverage of 30 x (Figure 2-3).

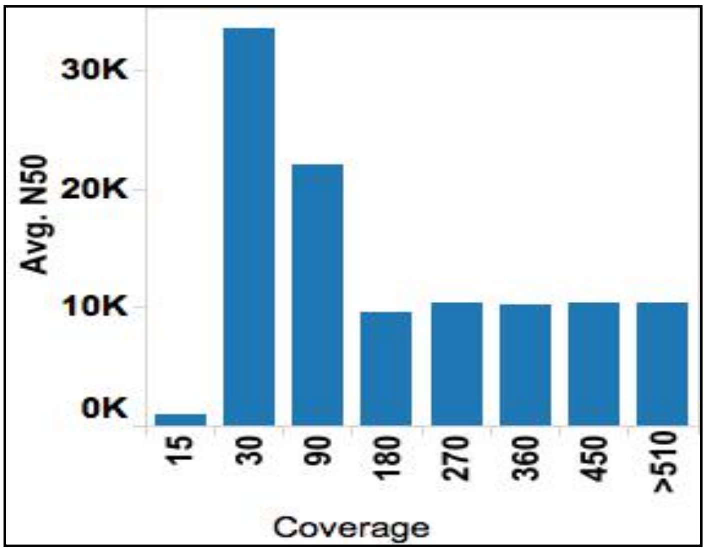

Figure 2-3: Velvet assembly average N50 (Kb) for coverage 15 x to > 510 x for seven wheat BACs.

Table 2-2: Velvet assembly average N50 (Kb) at various coverages for seven wheat BACs.

	Coverage x							
BAC	**15**	**30**	**90**	**450**	**180**	**270**	**360**	**>510**
A	1	63	4	4	4	4	4	5
B	1	6	9	8	8	8	8	8
C	1	3	4	6	6	5	6	5
E	1	64	10	9	9	9	8	9
F	1	4	25	9	6	6	9	9
G	1	86	89	25	21	28	23	25
H	1	5	9	9	9	9	9	9
Average N50	1	**33**	22	10	9	10	10	10

A comparison of Velvet, Ray and SASSY assemblies was done using assembly parameters producing the highest average N50. As previously observed, the Ray assembly with the highest average N50 was at k-mer of 30 and sequence coverage of 180 x (Table 2-1), whereas the Velvet assembly was at a coverage of 30 x (Table 2-2). A comparison of the three assemblers showed SASSY assemblies had the longest scaffold sizes, highest N50s and least number of contigs compared to Velvet and Ray assemblies (Figure 2-4,Table 2-3).

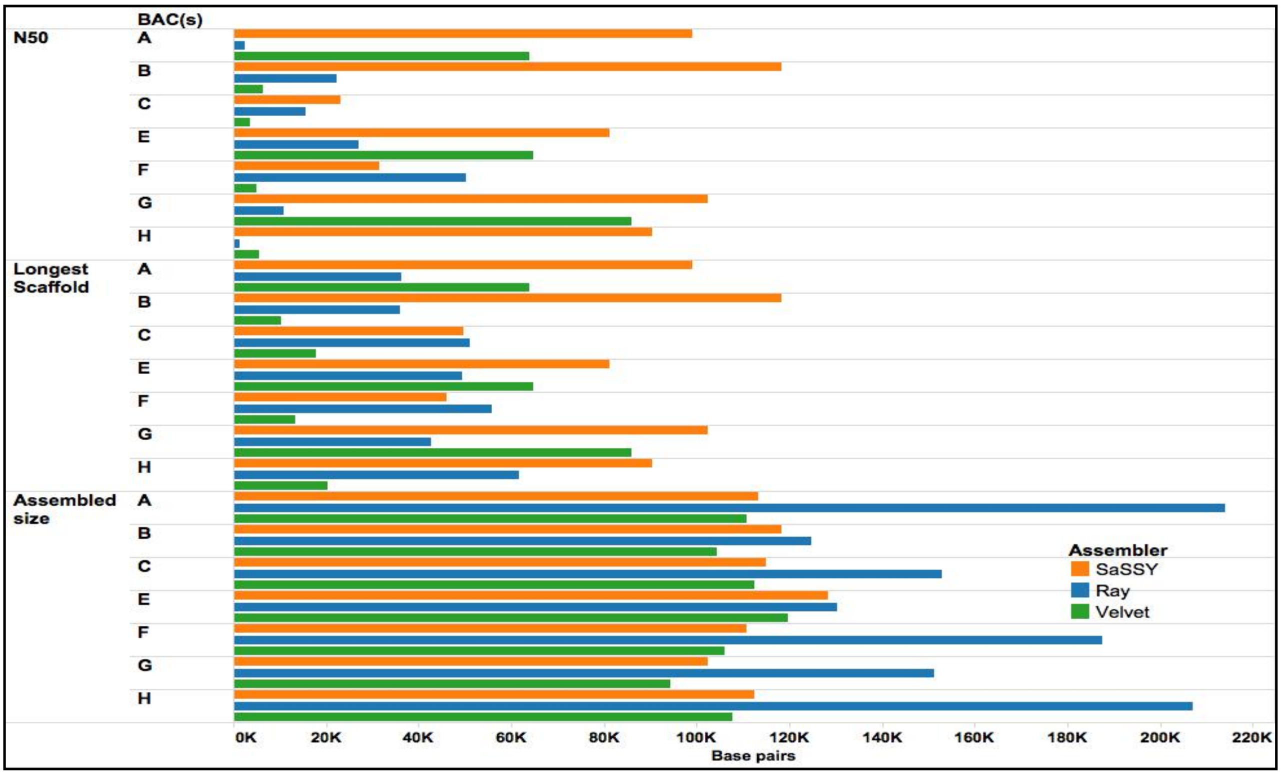

Figure 2-4: A comparison of N50, longest scaffold and assembly size of SASSY, Ray and Velvet assemblies for seven wheat BACs.

Table 2-3: A comparison of N50, longest scaffolds, assembly size and number of contigs for seven wheat BACs assembled with SASSY, Ray and Velvet.

		BACs							
Assembler	**Parameter**	**A**	**B**	**C**	**E**	**F**	**G**	**H**	**Average**
SASSY	N50 kb	99	118	23	81	32	102	90	**78**
	Longest scaffold Kb	99	118	50	81	46	102	90	**84**
	Assembled size Kb	113	118	115	128	111	102	113	**801**
	MTP kb	80	128	99	95	101	85	115	**703**
	No of contigs	4	1	7	4	5	1	4	**4**
Ray	N50 kb	3	22	16	27	50	11	2	**19**
	Longest scaffold Kb	36	36	51	50	56	43	62	**48**
	Assembled size Kb	214	125	153	130	187	151	207	**1,167**
	MTP kb	80	128	99	95	101	85	115	**703**
	No of contigs	625	123	241	23	276	348	449	**298**
Velvet	N50 kb	64	7	4	65	5	86	5	**34**
	Longest scaffold Kb	64	10	18	65	13	86	20	**39**
	Assembled size Kb	111	104	113	120	106	94	108	**756**
	MTP kb	80	128	99	95	101	85	115	**703**
	No of contigs	120	58	127	64	69	34	125	**85**

For the seven BAC assemblies, SASSY assemblies had an average N50 of 78 Kb compared to the Ray and Velvet assemblies with 19 Kb and 34 Kb respectively (Table 2-3). The longest scaffold sizes on average for SASSY, Ray and Velvet assemblies were 84 Kb, 48 Kb, and 39 Kb, respectively (Table 2-3). All three assemblers had on average, larger total assembled bases than the average MTP size estimate. On average, SASSY, Velvet and Ray assemblers had assembly sizes of 801 Kb, 1,167 Kb and 756 Kb respectively compared to an average MTP estimate of all the seven BACs of 703 Kb (Table 2-3). The average number of contigs per BAC for SASSY, Ray and Velvet assemblies was 4, 298, and 85 contigs respectively (Table 2-3). Two BACs (B and H) were assembled into a single contig with the SASSY assembler (Table 2-3), but no BACs could be assembled into single contigs using Ray or Velvet.

2.2.2 An optimal sequencing depth for BAC assembly

With a similar repeat content to wheat and the availability of BACs sequenced to very high coverage (> 10,000 x), sugarcane BAC clones were used to estimate the optimal sequencing depth required to assemble BACs using SASSY. Sequencing wheat BACs to a similar depth was costly. A summary of assembly statistics for each of the 11 sugarcane BACs, representing a total of 319 datasets, assembled with various stepped coverage suggested coverage was a factor in the outcome of an assembly. A plot showing assembly size versus increase in coverage using the R statistical software (gg2plot package) confirmed this. For each of the BACs, the assembly length increased to 450 x, levelled off between 450 x and 900 x, and then varied upwards after 900 x whilst remaining above ~100 Kb (**Figure 2-5** below).

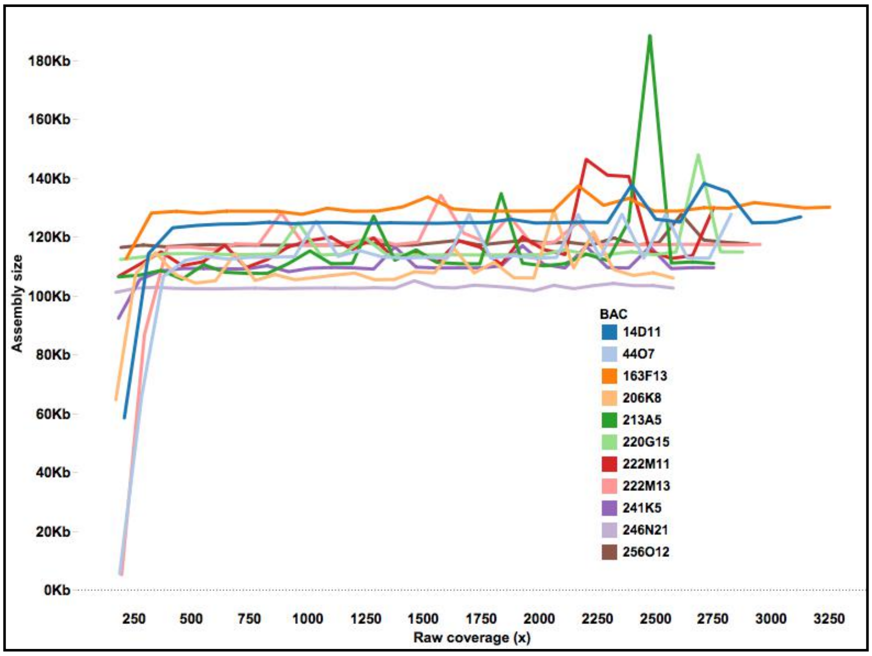

Figure 2-5: Optimal coverage: Assembly size vs. coverage for 11 sugarcane BACs.

No BACs had any sequences assembled below 50 x. The coverage required to assemble a 120 Kb BAC was predicted to be 680 x, with a standard error of 224 x (i.e. a range of 456 x - 904 x coverage), within a 95% confidence interval. The results were consistent with a visual inspection of the assembly size versus coverage plot **(Figure 2-5)** and the regression model **(Figure 2-6** below).

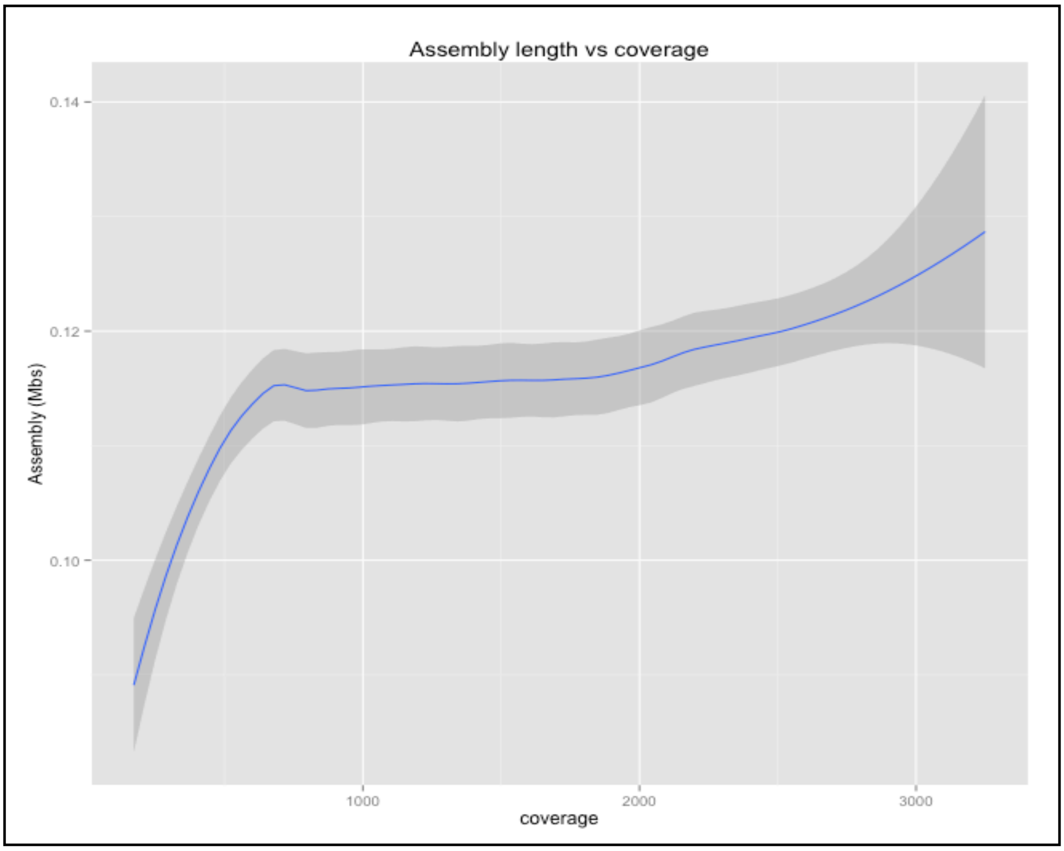

Figure 2-6: Local regression model (LOESS) of assembly size versus coverage.

2.2.3 The BAC pooling strategy

The accuracy of pooled BACs assembled with SASSY was evaluated, to assess whether the integrity of individual BACs in pooled assemblies was maintained. Seven wheat BACs sequenced to the estimated optimal sequencing depth > 450 x, as suggested from results of the determination of optimal sequencing depth in section 2.2.1, had a mean raw coverage of 844 x and a mean coverage of 716 x after filtering *E. coli* and vector sequences. Sequential pools of two to seven BACs were assembled. Assembly sizes of single BAC assemblies (A, B, C, E, F, G and H) agreed with the respective pooled BACs and MTP estimates (**Figure 2-7**, **Table 2-4** below). With every sequential addition of a BAC into the simulated pool, the assembly size increased accordingly. The assembly size increase was also consistent with the sum of the individual and pooled BACs for the seven BACs (A, B, C, E ,F ,G and

H) and the two to seven BAC pools (i.e. AB, ABC, ABCE, ABCEF, ABCEFG and ABCEFGH) (**Figure 2-7**, **Table 2-4** below).

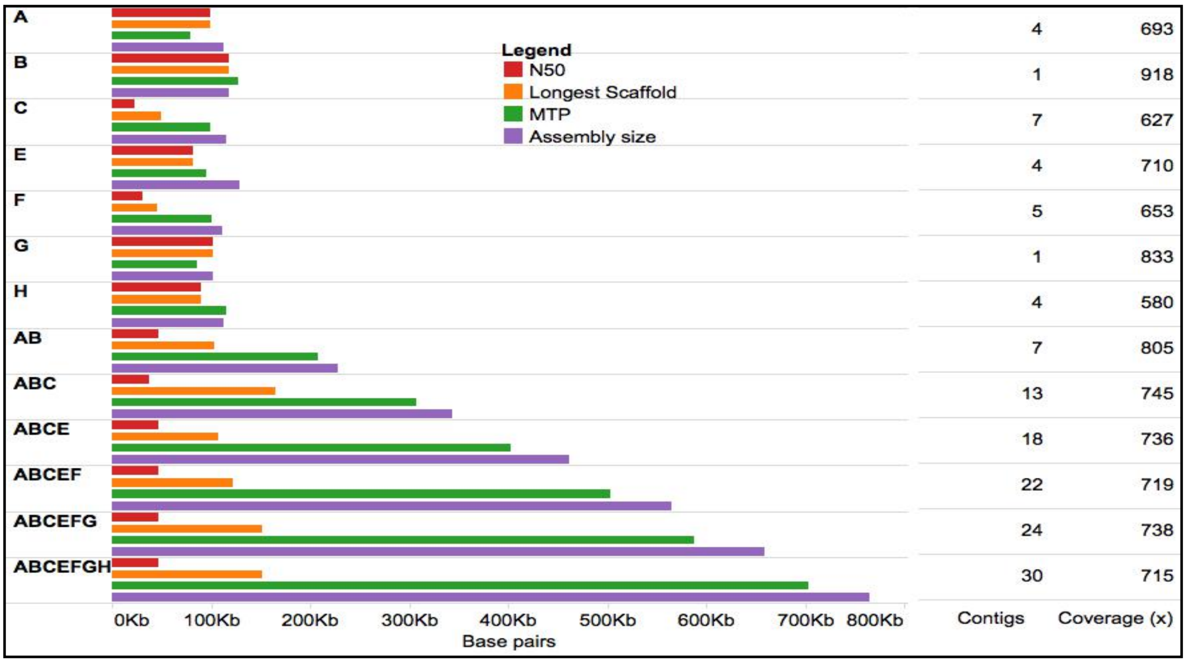

Figure 2-7: N50, longest scaffold, MTP sizes, assembly size and number of contigs of single BACs and simulated BAC pool assemblies.

Differences obtained between the sum of assembly sizes of individual BACs compared to the assembly size of the pooled BACs showed that they differed by between 3 Kb and 37 Kb with larger pools having a bigger difference. **Table 2-4** below).

Table 2-4: N50, longest scaffold *(a)*, MTP sizes, sum of single BAC assemblies *(b)*, assembly size, difference between single and pooled assembly sizes *(c)* and number of contigs of single BACs and simulated BAC pool assemblies.

BAC(s)	**N50 Kb**	**a (Kb)**	**MTP kb**	**b (Kb)**	**Asembly size Kb**	**c (kb)**	**No scaffolds**	**Cov (x)**
A	99	99	80	113	113	0	4	693
B	118	118	128	118	118	0	1	918
C	23	50	99	115	115	0	7	627
E	81	81	95	128	128	0	4	710
F	32	46	101	111	111	0	5	653
G	102	102	85	102	102	0	1	833
H	90	90	115	113	113	0	4	580
AB	47	104	207	232	228	3	7	805
ABC	38	165	307	347	344	3	13	745
ABCE	47	107	402	475	461	14	18	736
ABCEF	47	122	503	586	564	22	22	719
ABCEFG	47	151	588	688	659	29	24	738
ABCEFGH	47	151	703	801	764	37	30	715

There was an increase in the number of contigs assembled among simulated pools compared to single BAC assemblies. The N50 and longest contigs did not show such an increase (**Figure 2-7** above). BACs B and G each assembled into a single contig (**Figure 2-7**, **Table 2-4** above). Assemblies of the four BAC pools (ABCE, BCEF, CEFG and EFGH) had a mean contig count of 5.3 per BAC, and the average N50 of the pools was 50.3 Kb. In all four pools, the BES mapped at the edges of contigs as expected.

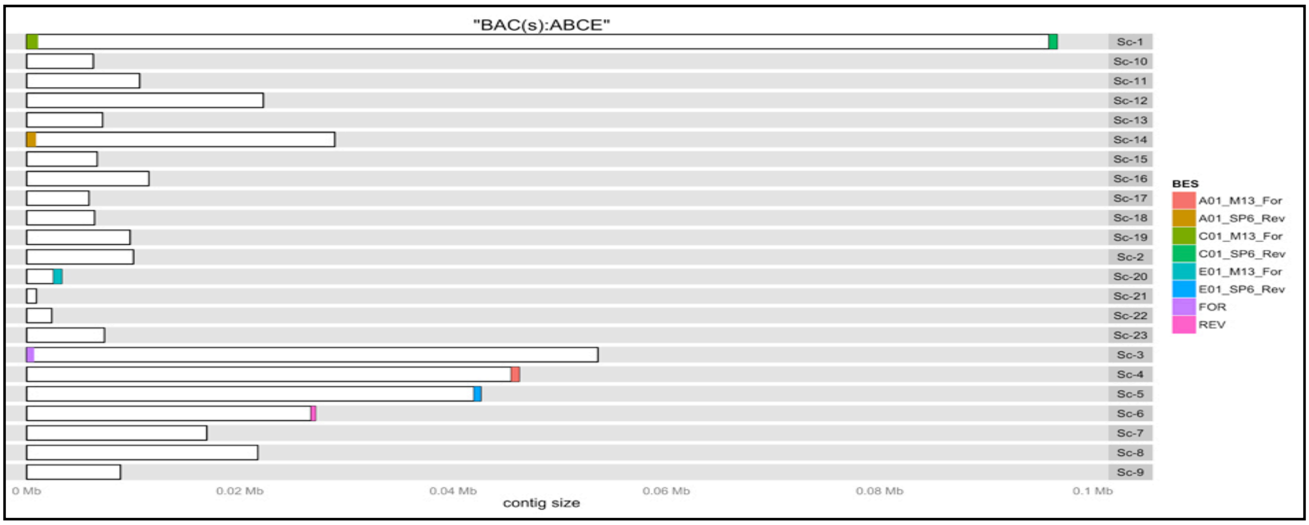

Figure 2-8: BES mappings of pool ABCE.

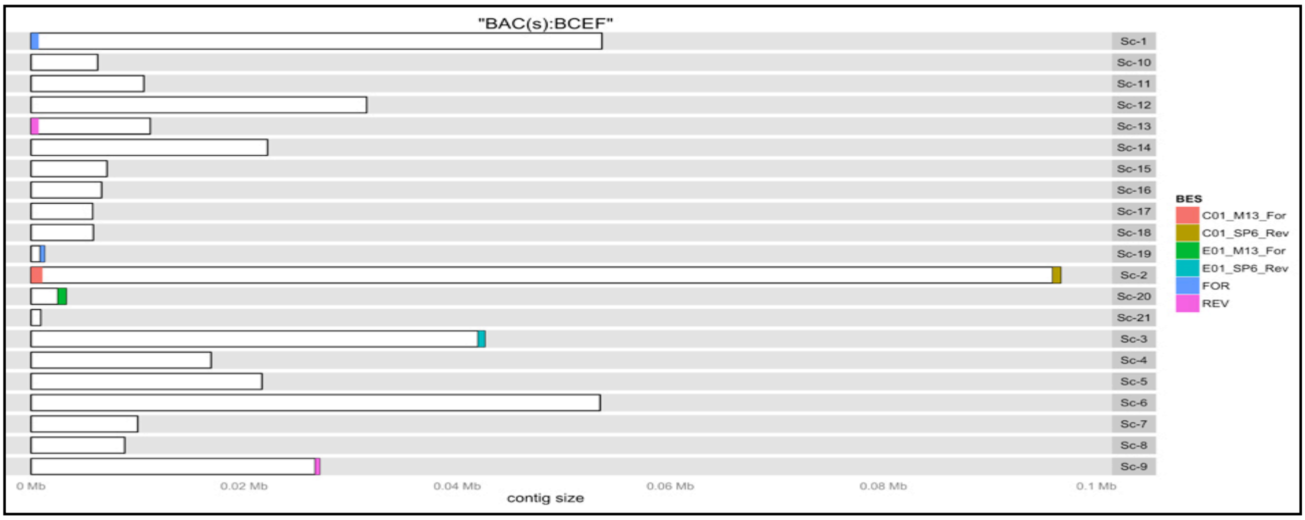

Figure 2-9: BES mappings of pool BCEF.

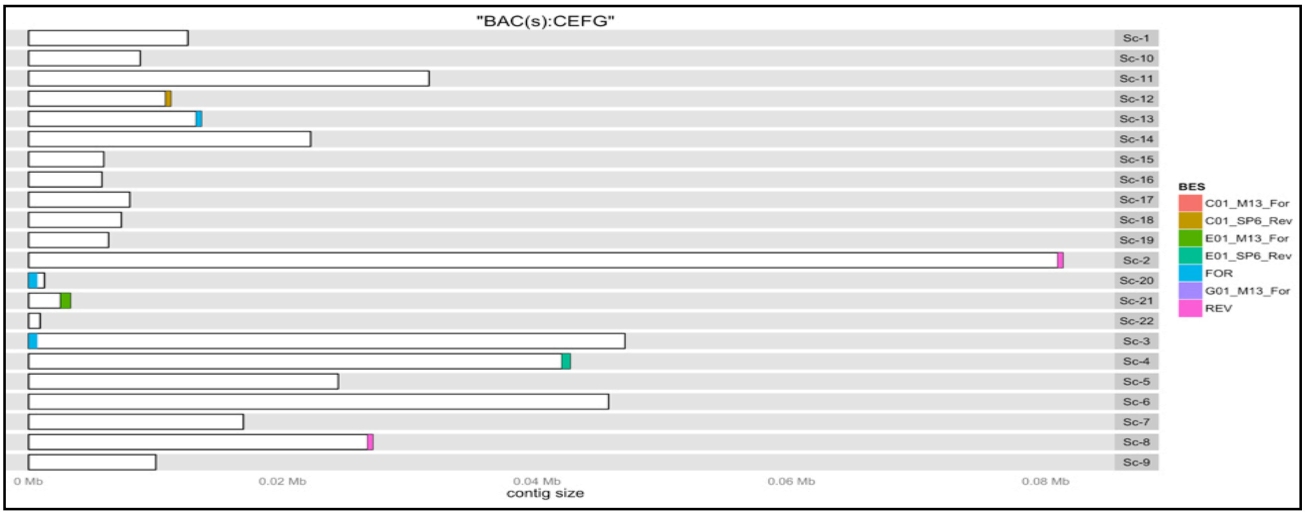

Figure 2-10: BES mappings of pool CEFG.

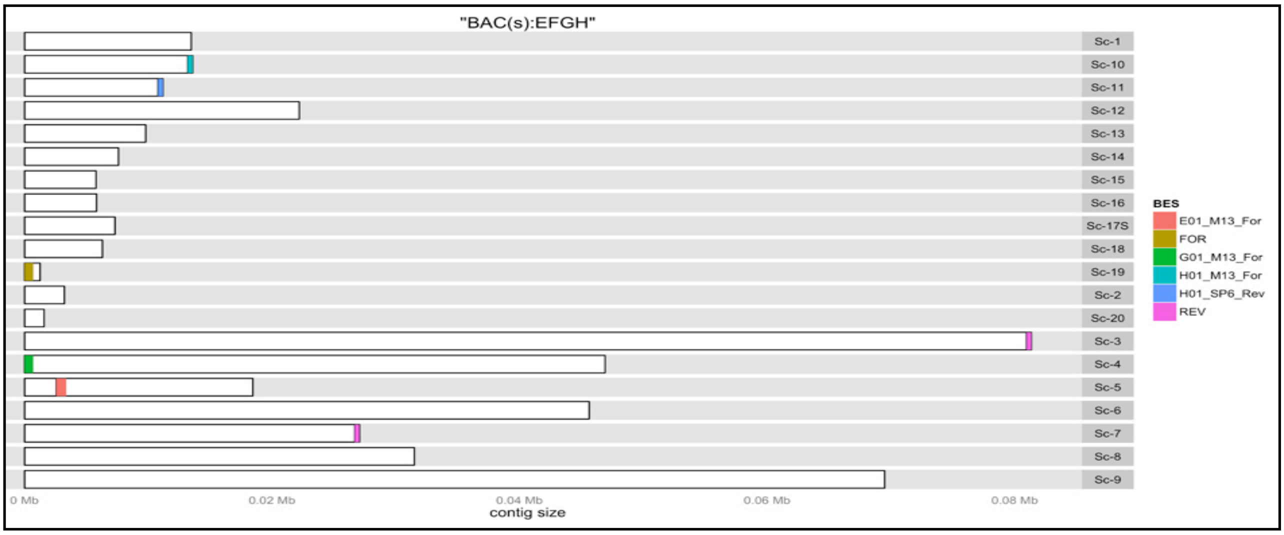

Figure 2-11: BES mappings of pool EFGH.

The contigs within pooled assemblies (ABCE, BCEF, EFGH) were compared using MUMmer to those of their corresponding single BACs (A, B, C, E), (B, C, E, F) and (E, F, G, H). MUMmer plots showed that contigs from un-pooled assemblies were collinear to those of pooled assemblies with repeat sequences within contigs showing regions with inversions with sequence similarity of < 100% (yellow and green segments in Figure 2-12 - Figure 2-15).

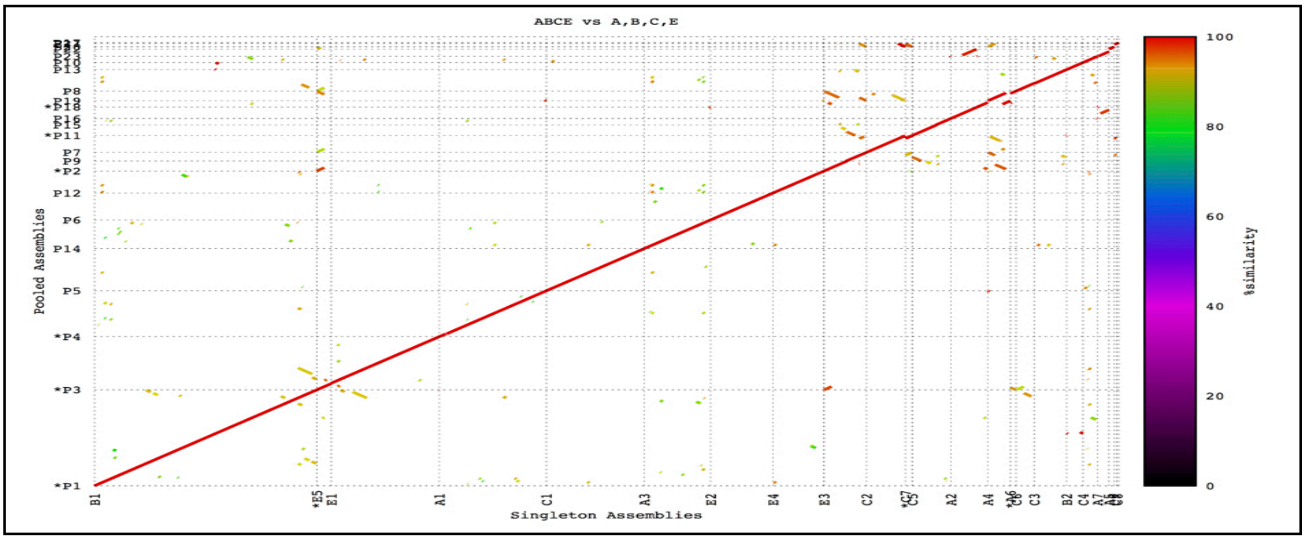

Figure 2-12: MUMmer plot of contigs of pool ABCE (y axis) versus contigs of BACs A, B, C, E (x-axis).

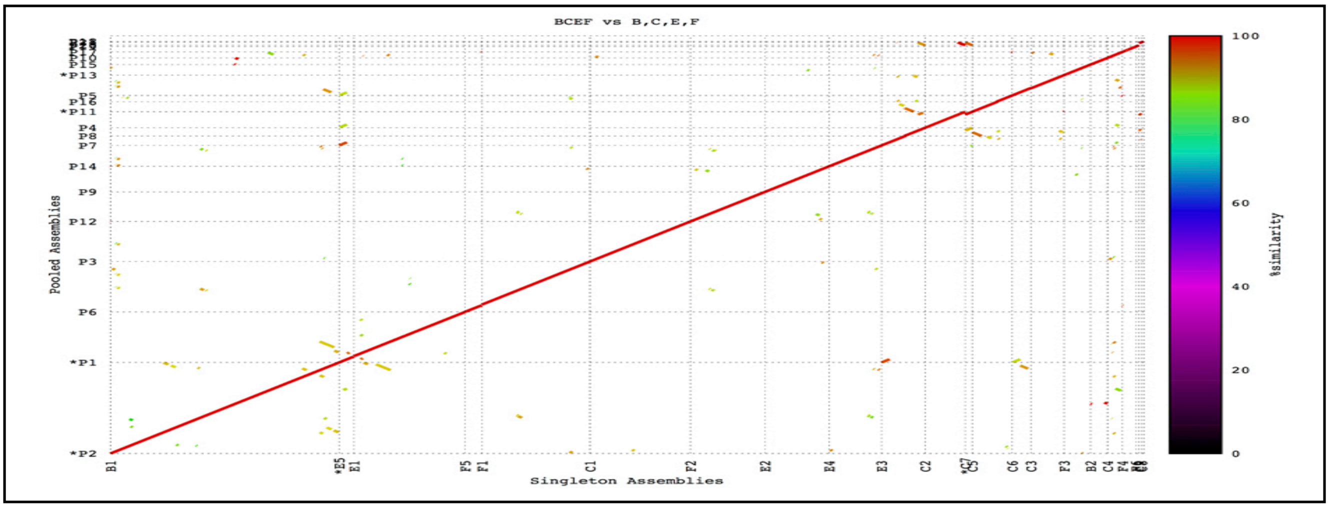

Figure 2-13: MUMmer plot of contigs of pool BCEF (y axis) versus contigs of BACs B, C, E, F (x axis).

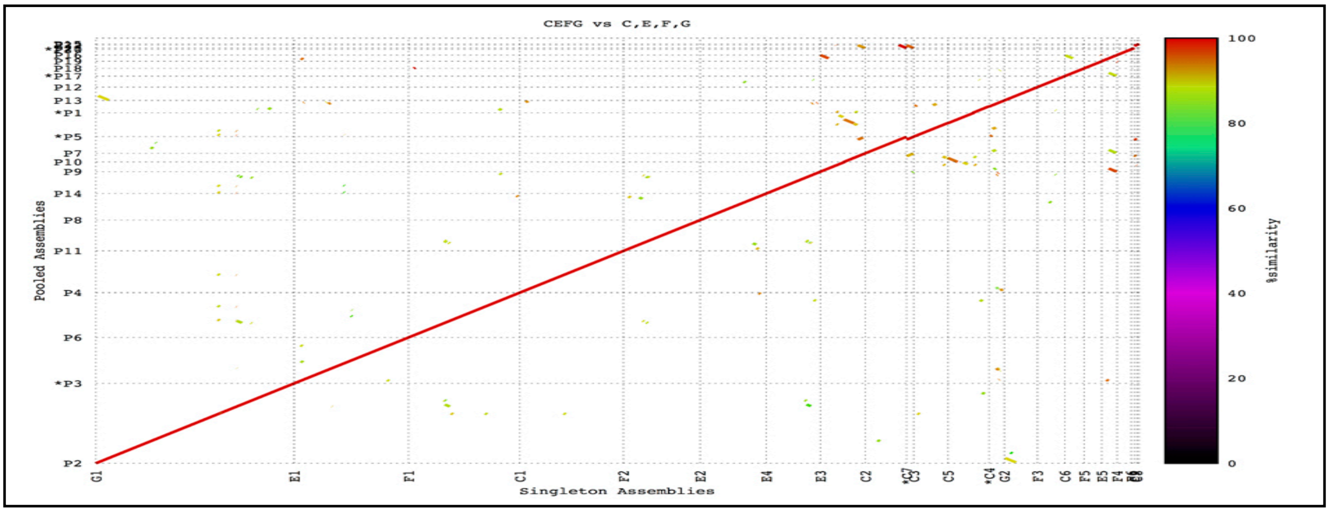

Figure 2-14: MUMmer plot of contigs of pool CEFG (y axis) versus contigs of BACs C, E, F, G (x axis).

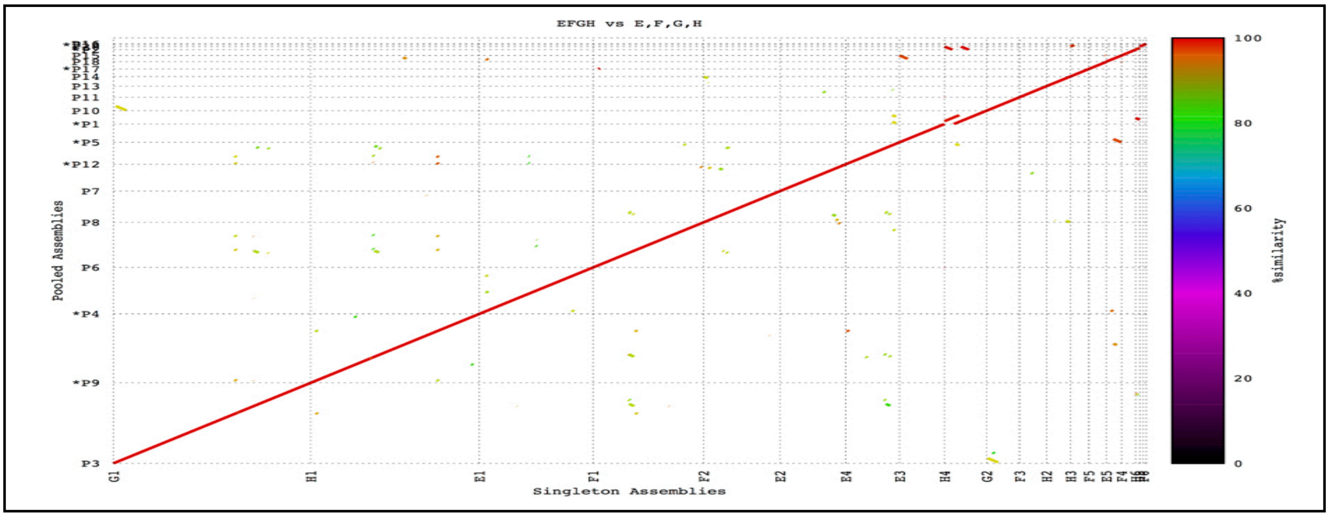

Figure 2-15: MUMmer plot of contigs of pool EFGH (y axis) versus contigs of BACs E, F, G, H (x axis).

Assemblies of single BACs and pooled BACs without the filtering of vector sequences showed the vector was always assembled in the middle of a contig (Figure 2-16 - Figure 2-19). BES mapping indicated that BES flanked the assembled vector. Fragments of the vector sequences also mapped to other contig ends at the same positions as BES (Figure 2-16 - Figure 2-19). The number of BES and vector fragment mapping positions at the end of contigs was proportional to the number of pooled BACs. For example, for a single BAC, vector fragments and the two BES (forward and reverse BES), mapped to two positions in the BAC assembly (Figure 2-16). For pooled assemblies, there were four locations for the AB pool which had two BACs A and B (Figure 2-17), six for the ABC pool (Figure 2-18), and eight for the ABCE pool (Figure 2-19). Inspection of the BES mapping also showed two BACs were assembled into one contig joined by the assembled vector (Figure 2-16 - Figure 2-19).

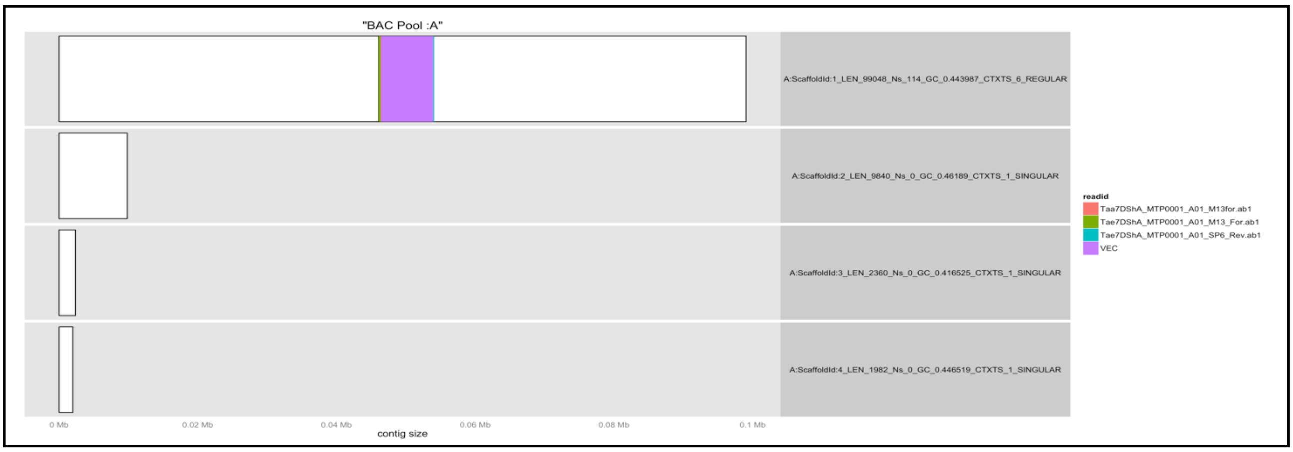

Figure 2-16: Single BAC “A” assembled with vector sequences unfiltered. The vector assembled in the middle of the contig (purple block).

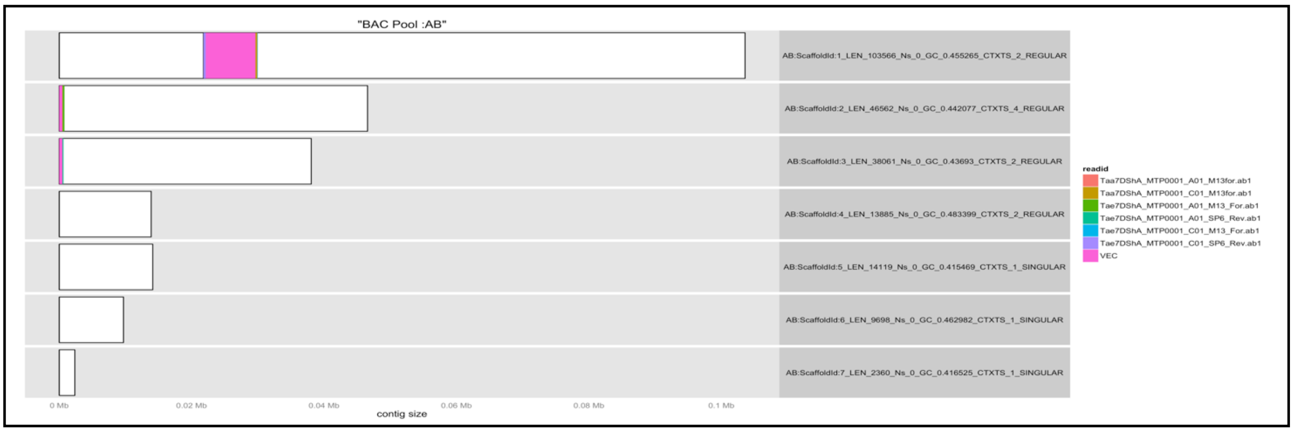

Figure 2-17: Two pooled BACs "AB" assembled with vector sequences unfiltered. The vector assembled in the middle of a contig (purple block).

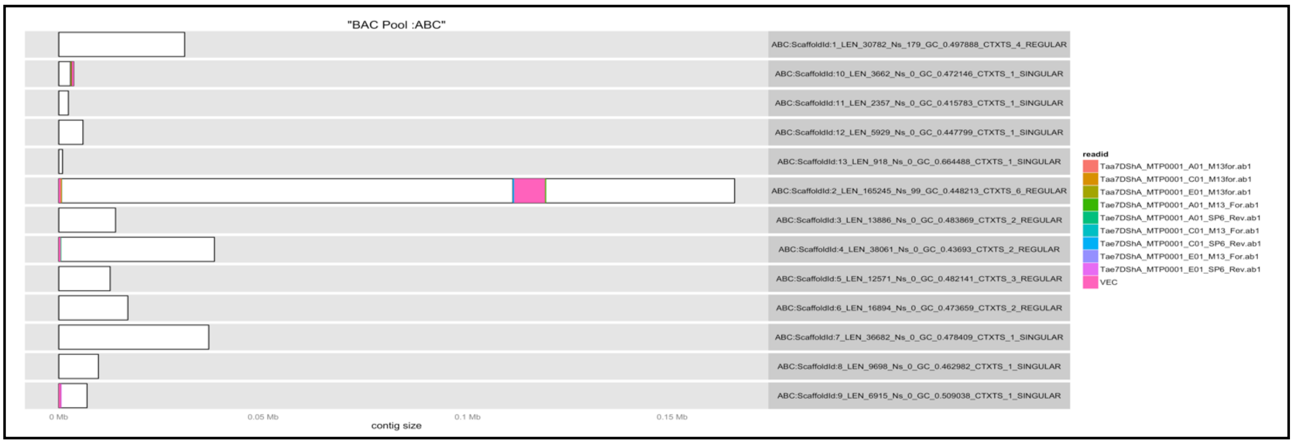

Figure 2-18: Three pooled BACs "ABC" assembled with the vector sequence unfiltered. The vector assembled in the middle of a contig (purple block).

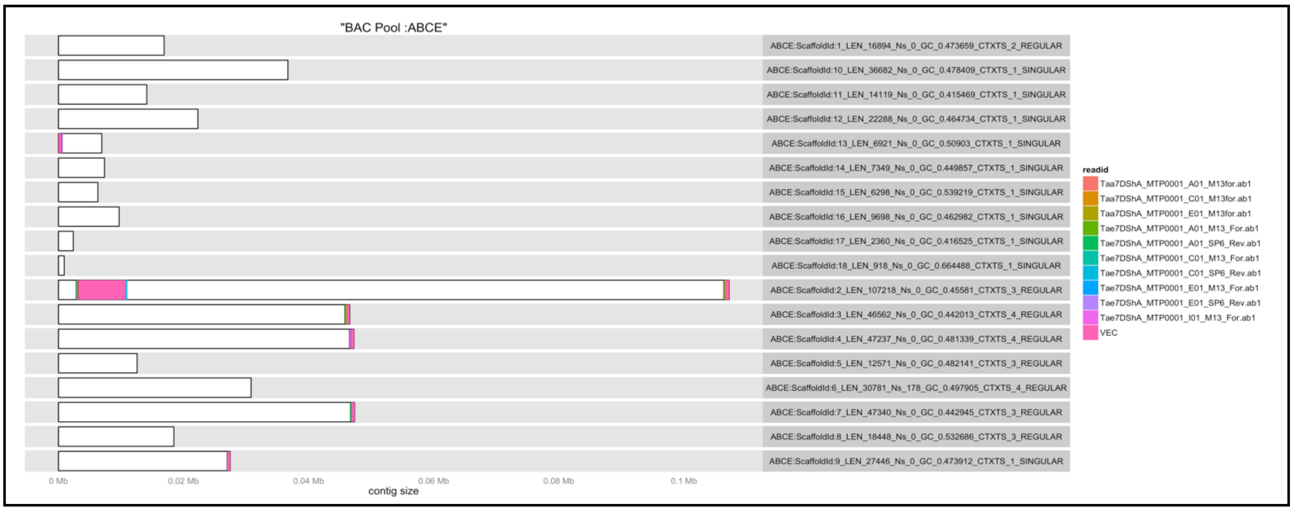

Figure 2-19: Four pooled BACs "ABCE" assembled with vector sequences. Vector assembled in the middle of a contig (purple block).

A pooling strategy based on four non overlapping MTP BACs was applied to the sequencing of chromosome 7DS resulting in the pooling and indexing of 4608 BACs into 96 pools on 12 plates with 384 cells each. A table of the pooled BACs in all 12 plates is attached in Appendix 2.

2.2.4 An optimised SASSY assembly method for pooled BACs

An evaluation of the effect of library preparation kits and read lengths on SASSY assemblies was performed. Three library preparation kits; TruSeq, Nextera and Nextera XT were used to prepare seven, six and four BAC pools respectively for sequencing on the Illumina HiSeq 2000. This resulted in 17 pools in total, with a pool having four BACs. Assemblies of the 17 pools at read lengths; 61 bp, 71 bp, 81 bp, 91 bp and 100 bp with SASSY were undertaken. All 17 assembled pools were compared based on the number of read pairs, assembly N50, longest assembled contig and number of contigs (Figure 2-20).

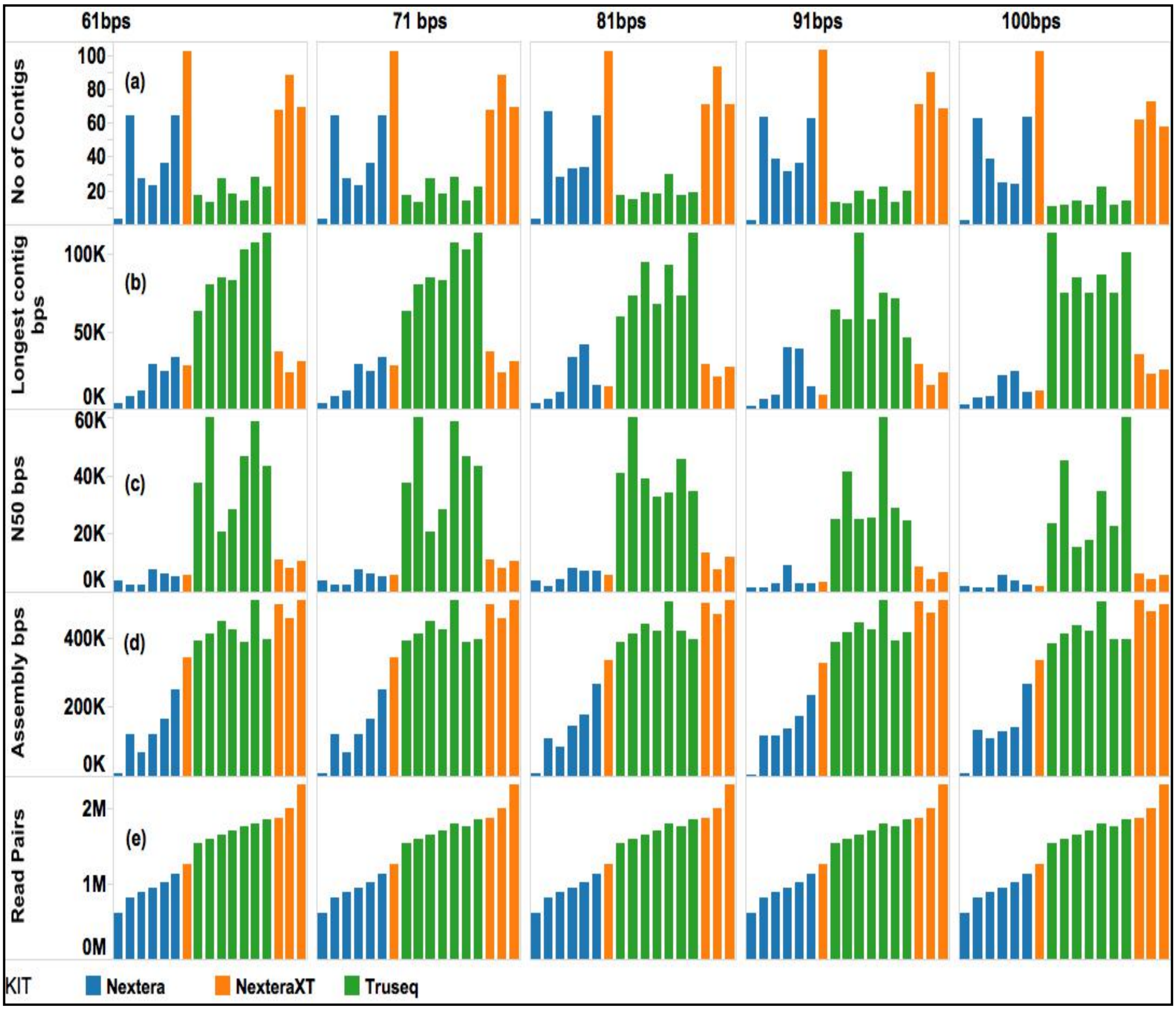

Figure 2-20: Illumina DNA library preparation kits (Nextera in blue, Nextera XT in orange, TruSeq in green) and their effects on assembly size, N50, longest contig, number of contigs for assemblies assembled at 61 bp, 71 bp, 81 bp, 91 bp and 100 bp.

TruSeq assemblies had the least number of contigs across all five assemblies at read lengths of 61 bps, 71 bps, 81 bps, 91 bps and 100 bps (Figure 2-20). TruSeq had the least average number of contigs (19 – 20) compared to Nextera (33 – 52) and Nextera XT (75 – 106) (Table 2-5). TruSeq assemblies similarly had the longest contigs across all five assemblies at read lengths of 61 bp, 71 bp, 81 bp, 91 bp and 100 bp (Figure 2-20 **b**). The average longest contigs for TruSeq assemblies ranged from 91 Kb - 100 Kb (Table 2-5) compared to 14 Kb - 26 Kb for Nextera and 26 Kb - 38 Kb for Nextera XT. In addition, the TruSeq assemblies had the highest N50's across all five assemblies at read lengths of 61 bp, 71 bp, 81 bp, 91 bp and 100 bp (Figure 2-20 **c**). The average N50 for TruSeq assemblies ranged from 34 Kb - 49 Kb compared to 4 Kb - 9 Kb for Nextera and 7Kb - 10 Kb for Nextera XT (Table 2-5). The TruSeq kit, however, did not produce the longest assemblies (Figure 2-20 **d**) or the highest number of read pairs after from sequencing (Figure 2-20 **e**). The Nextera XT kit had the highest total assembled bases (Table 2-5) (442 - 462 Kb) compared to TruSeq (420 - 430 Kb) and Nextera (115 - 134 Kb). The average sequence coverage per BAC was 339 x, 717 x and 648 x for Nextera, Nextera XT and TruSeq kits respectively.

Table 2-5: Average N50, longest contig, contig counts and assembly size

KIT	Read length	Longest contig (Kb)	N50 (Kb)	Contig counts	Total bases (Kb)
TruSeq	61	94	34	20	420
	71	91	42	20	423
	81	94	41	19	428
	91	100	48	19	430
	100	97	49	20	427
Nextera	61	25	9	33	115
	71	18	5	36	122
	81	21	5	38	132
	91	26	5	44	131
	100	14	4	52	134
Nextera XT	61	38	9	75	442
	71	30	9	82	450
	81	26	10	84	455
	91	28	9	93	456
	100	27	7	106	462

Comparisons of the 'longest contig length' statistic for TruSeq assemblies showed read lengths of 91 bp had the highest average (100 Kb) when compared to read lengths of 61 bp, 71 bp, 81 bp and 100 bp. Assemblies at read lengths of 61 bp, 71 bp, 81 bp and 100 bp had an average 'longest contig length' of 94 Kb, 91 Kb, 94 Kb and 97 Kb respectively (Table 2-6). Comparisons of the averages of the standard deviations of the 'longest contig length' showed assemblies at read lengths of 91 bp had the highest average standard deviation (31 Kb) when compared to assemblies at read lengths of 61 bp, 71 bp, 81 bp and 100 bp. These had average 'longest contig length' standard deviations of 19 Kb, 18 Kb, 22 Kb and 17 Kb (Table 2-6).TruSeq assemblies at read lengths of 100 bps had the highest average N50 (49 Kb) compared to assemblies at read lengths of 61 bp, 71 bp, 81 bp and 91 bp which had average N50's of 34 Kb, 42 Kb, 41 Kb and 48 Kb respectively (**Table 2-6**). The highest average assembly size was at a read length of 91 bps (430 Kb) compared to 420 Kb, 423 Kb, 428 Kb and 427 Kb observed for assemblies at read lengths of 61 bp, 71 bp, 91 bp and 100 bp respectively (**Table 2-6**).

Table 2-6: TruSeq assembly statistics.

	Read length				
	61	**71**	**81**	**91**	**100**
SD longest contig (Kb)	19	18	22	31	17
Avg longest contig (Kb)	94	91	94	100	97
Avg N50 (Kb)	34	42	41	48	49
Avg contig counts	20	20	19	19	20
Avg assembly size (Kb)	420	423	428	430	427

To further improve SASSY assemblies, three sets of SASSY assembly parameters were evaluated using assemblies of 96 pools of wheat BACs. Each of the 96 pools were assembled using SASSY naive and extension offset parameters '2,10', '4,12' and '8,16' at 100 bp. A comparison of assembly lengths across three assemblies for each pool showed there was variation in the total assembled length with variation increasing with assembly size (Figure 2-21) Analysis of variance (ANOVA) statistical tests showed the observed variation in assembly lengths was significant (*P value* < 0.000493, *F* ratio 7.82, df 2). Pool MTP sizes were obtained by adding up individual BAC MTP sizes in a pool. Pairwise statistical analysis of assembly lengths of all 96 pools assembled at offsets '2,10', '4,12' and '8,16' and their corresponding pool MTP sizes was performed. Paired t tests showed assemblies at offsets '2,10' were significantly larger than their respective pooled MTP sizes (one tailed t test, p value < 0.03) (Table 2-7). Assemblies at offsets '4,12' showed no significant difference compared to their MTP sizes (one tailed t tests, p value < 0.56 and 0.44) (Table 2-7), while assemblies at offsets '8,16' were significantly smaller compared to their MTP sizes (one tailed t test, p value < 0.03) (Table 2-7).

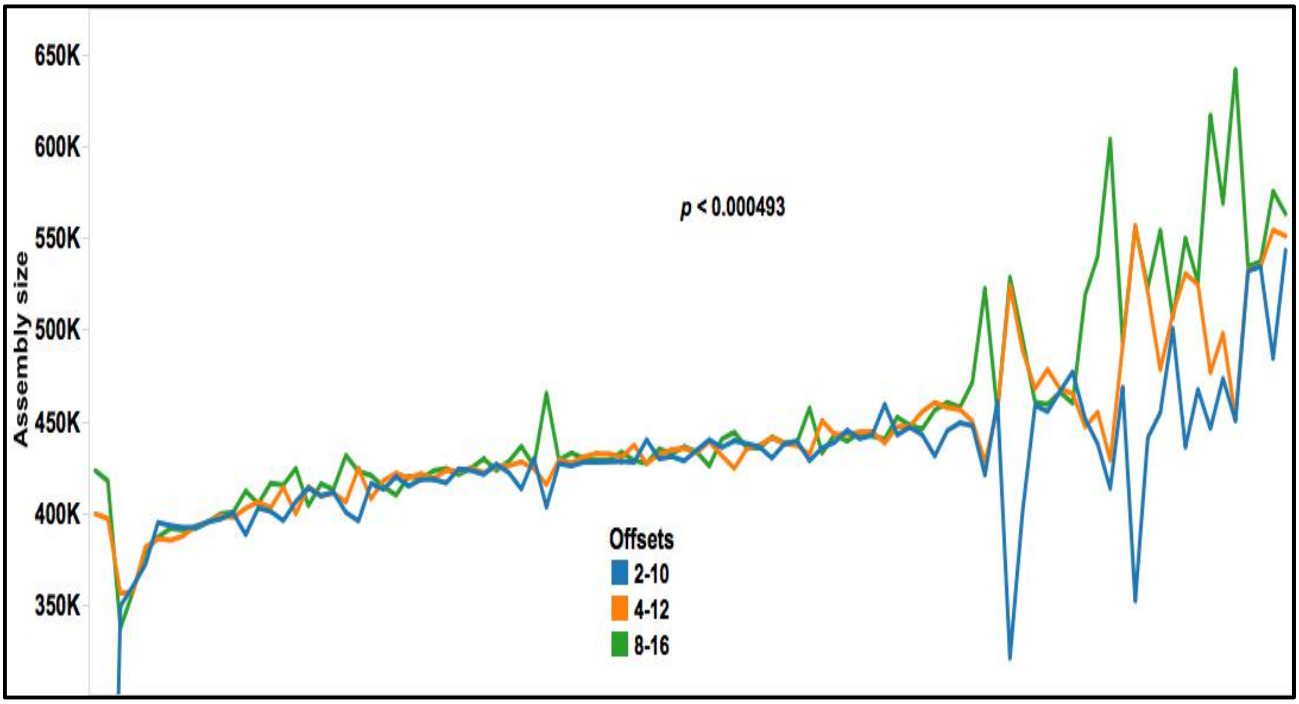

Figure 2-21: Variation in assembly lengths with different SASSY offset parameters. Y- axis represents assembly size. X-axis represents 96 pools. Assembly lengths were sorted in increasing size from left to right.

The average MTP size of all 96 BAC pools was 440,111 bp. The average assembly sizes of all the 96 pools assembled with SASSY offsets ‘2,10’, ‘4,12’ and ‘8,16’ were 421,146 bp, 440,964 bp and 452, 810 bp respectively (**Table 2-7**). The difference between the assembly size average and the MTP size average for assemblies at offsets ‘2,10’, ‘4,12’ and ‘8,16’ was 18,965 bp, -813 bp and -12,699 bp respectively (**Table 2-7**). Assemblies at offsets ‘4,12’ had the least assembly size difference (813 bp) from the MTP size estimates.

Table 2-7: Paired t tests of 96 pools assembled at offsets ‘2,10’, ‘4,12’ and ‘8,16’ with their MTP sizes.

Offsets 2,10			**Offsets 4,12**			**Offsets 8,16**	
MTP mean size	**440111**		**MTP mean size**	**440111**		**MTP mean size**	**440111**
Mean assembly size	421146		Mean assembly size	440924.00		Mean assembly size	452810
Mean difference	**18965**		**Mean difference**	**-813**		**Mean difference**	**-12699**
Std error	9590.93		Std error	5706.83		Std error	6621.10
Upper 95%	38005.60		Upper 95%	10516.30		Upper 95%	445.97
Lower 95%	-75.17		Lower 95%	-12143.00		Lower 95%	-25843
Correlation	-0.28		Correlation	0.25		Correlation	0.25
t-ratio	1.98		t-ratio	-0.14		t-ratio	-1.92
DF	95		DF	95		DF	95
Prob > |t|	0.05		**Prob > |t|**	**0.89**		Prob > |t|	0.06
Prob > t	**0.03**		**Prob > t**	0.56		Prob > t	0.97
Prob < t	0.97		**Prob < t**	0.44		**Prob < t**	**0.03**

The mean assembly length closest to that of the MTP mean of 440,111 bps was of the ‘4,12’ offsets at 440,924 bps compared to 421,146 bps for ‘2,10’ offsets and 452,810 bps for the ‘8,16’ offsets (**Table 2-7**).

2.2.5 A pipeline for high-throughput assembly of BACs

The assembly of 4,608 wheat chromosome arm 7DS BACs in 1,152 BAC pools of four BACs each was performed based on results from the selection of a suitable assembler (section 2.2.1), the determination of optimal coverage (section 2.2.2), the BAC pooling strategy (section 2.2.3) and the optimisation of assembly parameters (section 2.2.4). The assembly process was automated using Perl scripts. The resulting pipeline, the SASSY pipeline, runs on a Linux machine via command line. The pipeline setup depends on a configuration file that has absolute paths to required software packages, reads for assembly, BES files and settings for desired assembly parameters. The pipeline setup, required modules and software packages are described in detail in Appendix 1. Running the pipeline on average required approximately 15 minutes per pool when using five CPUs with 25 Gb memory. Step one of the pipeline involved the filtering of *E. coli* and vector reads (**Figure 2-1**) and required on average two minutes. Step two of the pipeline involved assembly of filtered reads with SASSY. The assembly took approximately 12 minutes. Step three and four involved mapping of BES to assembled contigs using BLAST, this took approximately one minute. A summary of assembly statistics for 1,120 pools from 12 plates is shown in **Table 2-8** below.

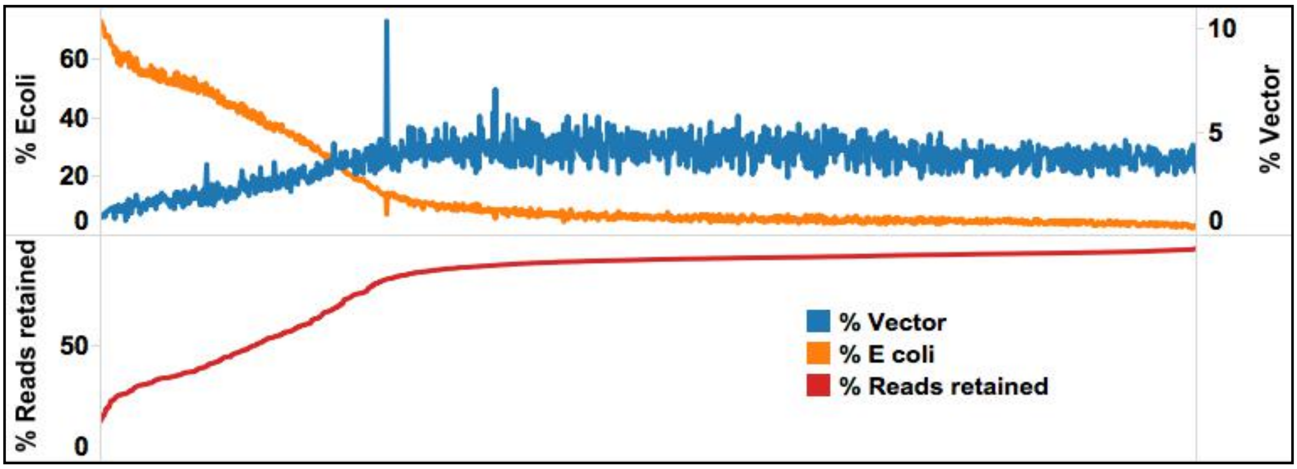

Figure 2-22: Percentage of *E. coli* reads, and vector reads (X-axis top row) in sequenced reads and % of reads retained after *E. coli* and vector filtering (X-axis bottom row) for 1,120 BACS. X axis represents 1, 120 BACs ordered in increasing % of reads retained from left to right.

For the 1,120 pools assembled, the average percentage of *E. coli* and vector sequences was ~15 % and ~4 % respectively (**Table 2-8**). The average clonality of the sequenced reads was < 1%. The average length of the longest contigs and N50 was 102 Kb and 65 Kb respectively. As expected, pools with a high *E. coli* percentage had a lower percentage of reads retained after *E. coli* and vector filtering (Figure 2-22). The assembly resulted in an average of 18 contigs per pool (4 contigs per BAC). A total of 418 pools (37%) had sequence coverage below the earlier determined minimum sequencing depth per BAC of 450 x. For pools with less than 450 x coverage, the average percentage of *E. coli* and vector sequences in sequenced reads was 27 % and 3 % respectively.

Pools with < 450 x had a lower average for the longest contig compared to pools with coverage > 450 x (77kb vs. 116 kb) and the average N50 was also lower (43 Kb vs. 77 Kb). Pools with coverage > 450 x had pool assembly lengths in agreement with their respective MTP estimates while pools with < 450 x coverage were shorter than their estimated MTP (Figure 2-23).

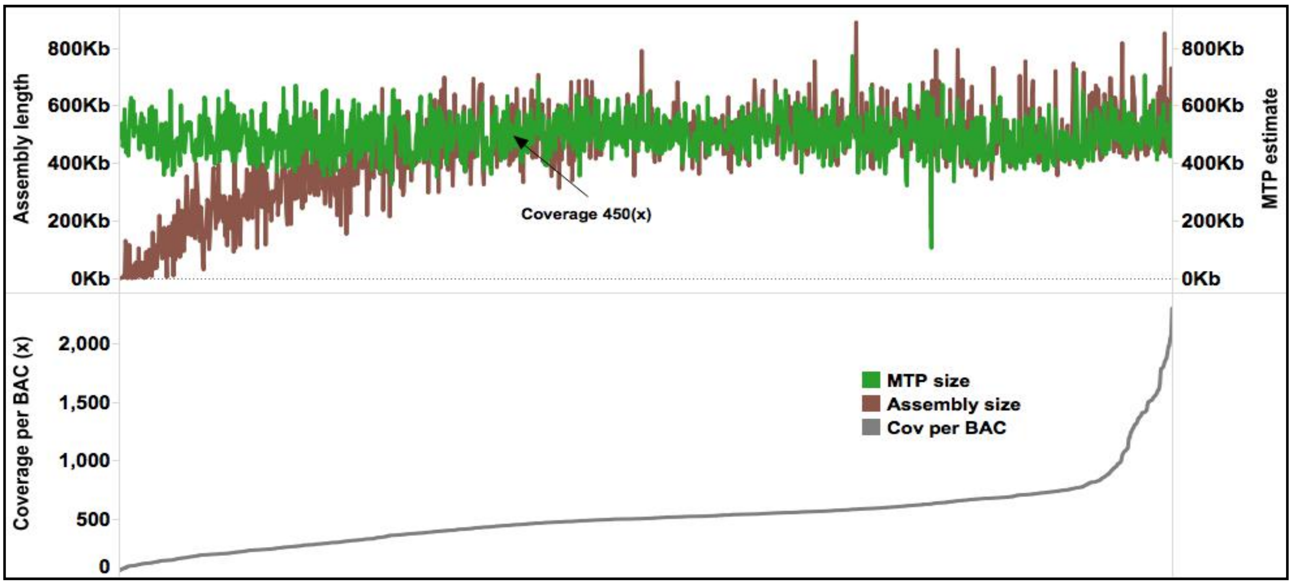

Figure 2-23: Comparison of assembly size, MTP size and coverage per BAC for 1,120 BAC pool assemblies. X-axis represents 1,120 pools sorted with lowest coverage to the left. Top Y-axis represents assembly and coverage per pool and bottom Y-axis coverage per BAC.

The percentage of unique reads in sequenced reads decreased with increase in coverage (Figure 2-24).

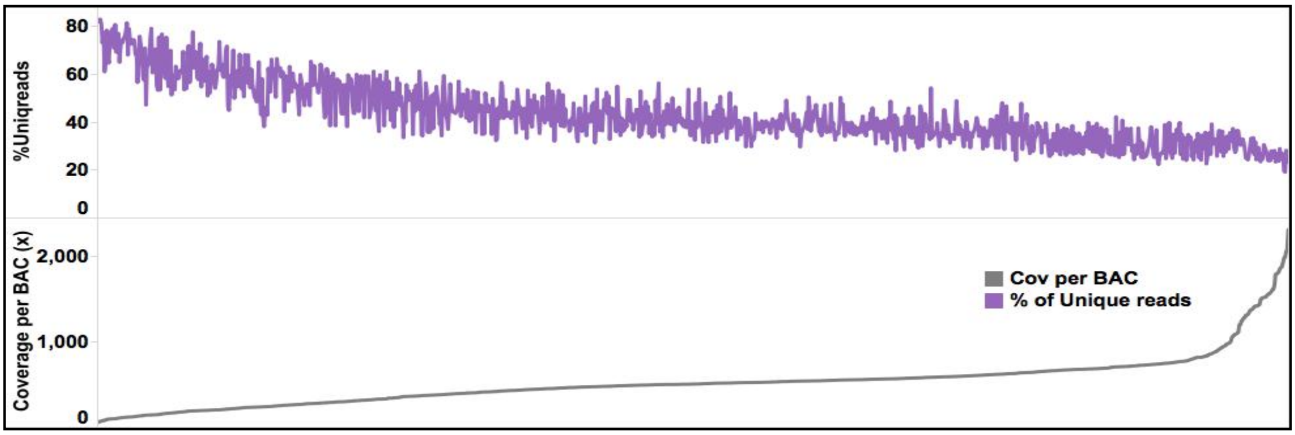

Figure 2-24: Comparison of sequence coverage with percentage of unique read pairs in pools. X-axis represents 1,120 pools sorted with lowest coverage to the left. Top Y-axis represents % of unique reads, and bottom Y-axis represents coverage per BAC.

As expected, an increase in coverage, resulted in a drop in the number of contigs in assembled pools (Figure 2-25)

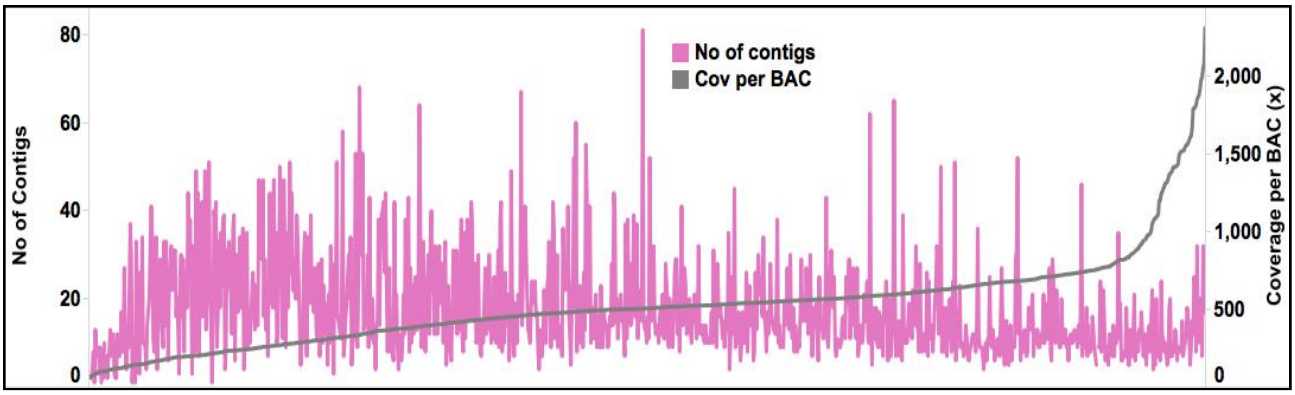

Figure 2-25: Comparison of coverage and number of contigs in pool assemblies. X-axis represents 1,120 pools sorted with lowest coverage to the left. Left Y-axis represents No of contigs and right Y-axis represents coverage per BAC.

Table 2-8 below shows the average assembly statistics per plate. Four plates 1, 4, 5, and 12 had the average coverage per BAC lower than < 450 x.

Table 2-8: Average assembly statistics per plate.

Plate	Avg. *E coli* %	Avg. vector %	Clonality %	Avg. coverage per BAC	Avg. longest contig (Kb)	Avg. N50 (Kb)	Avg. no of contigs per BAC	Avg. assembly length (Kb)	Avg. MTP (Kb)
1	28	3	0	408	93	65	4	355	442
2	5	5	1	691	106	81	3	441	440
3	32	3	0	906	102	74	4	419	472
4	39	3	0	304	90	59	4	307	467
5	35	3	0	315	71	46	4	295	459
6	10	4	0	550	110	65	5	482	497
7	7	4	0	524	117	73	4	520	536
8	6	4	0	535	114	72	4	519	494
9	4	4	0	556	114	67	5	530	515
10	5	4	0	651	116	68	6	526	529
11	5	4	0	525	106	59	7	511	560
12	6	4	0	329	89	50	5	390	551

A total of 418 pools (37%) had sequence coverage below our target sequencing depth per BAC of 450 x. Assemblies had an average N50 of 65 Kb, with an average of 4.4 contigs per BAC (median of 3.5). The number of contigs per BAC ranged between 1-20. The average percentage of Ns in the assemblies was 0.1%.

2.2.6 Mate pair (MP) scaffolded wheat chromosome arm 7DS BAC assemblies

To accurately scaffold BAC pool assemblies using MP libraries, filtering sequencing adapters from the MP reads is critical. An evaluation of the adapter filtering criteria was done using plate 1 MP, 6 -

10 Kb library. The plate 1 MP library had 70,386,715 read pairs sequenced, 70,227,606 (99.77%) of these were retained after trimming of external Illumina sequencing adapters. The read length distribution of a sample of 10,000 read pairs from the external adapter filtered reads (Figure 2-26) showed ~96 % of the reads had read lengths between 141 bp - 150 bp, ~3 % had between 131 bp - 140 bp and ~1 % had read lengths < 130 bp.

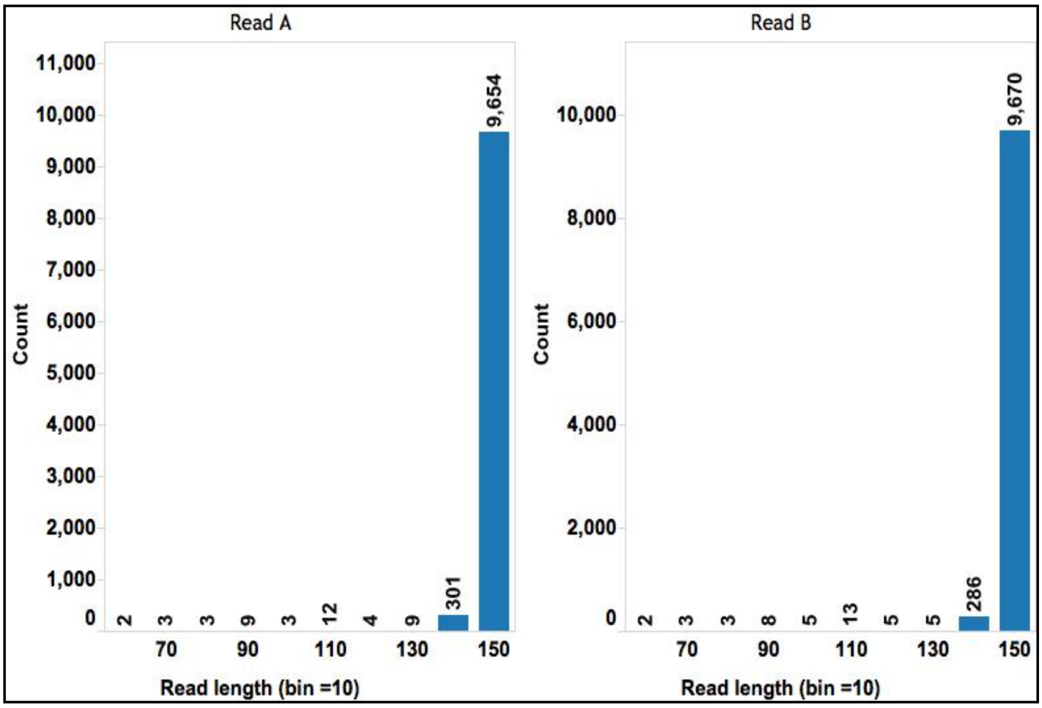

Figure 2-26: A histogram of a sample of 10,000 external adapter filtered MP reads (read A and B). ~96% of the reads had lengths between 141 - 151 bp, bin size 10.

After external adapter filtering the remaining 70,227,606 MP reads were then processed with the NxTrim software (O'Connell *et al.*, 2014). Reads less than 35 bp long post processing were discarded. 94/70,227,606 MP reads had 2 copies of adapters and were thus discarded. 45,334/70,227,512 (0.06%) MP reads were discarded because the combined length of the trimmed read pairs was less than the read length (150 bp). A total of 70,182,178 read pairs were left for downstream processing. From these 70,182,178 read pairs; NxTrim generated four virtual libraries designated as paired end (PE), mate pair (MP) and unknown. The virtual libraries are named as such as they are computationally. The generation is based on the presence and position of junction adapters while maximizing the number of paired reads recovered after trimming. If trimming off the junction adapter resulted in paired read fragments (designated as either **w**, **x** and **y** in Figure 2-27) having a RF orientation (<-- -->) (Figure 2-27 **a** & **b1**), the read pairs were designated as MP. If trimming off the junction adapter resulted in paired fragments having a FR orientation (--> <--) (Figure 2-27 **b2**), the read pairs were designated as PE. Where no junction adapters were found in the read pairs (Figure 2-27 **c**), the reads were not trimmed and were designated as unknown. Where the junction adapter was trimmed off and the remaining read fragment (either **w**, **x** or **y** in Figure 2-27) was less than 35 bp, the read was discarded and the remaining mate was designated as single end read.

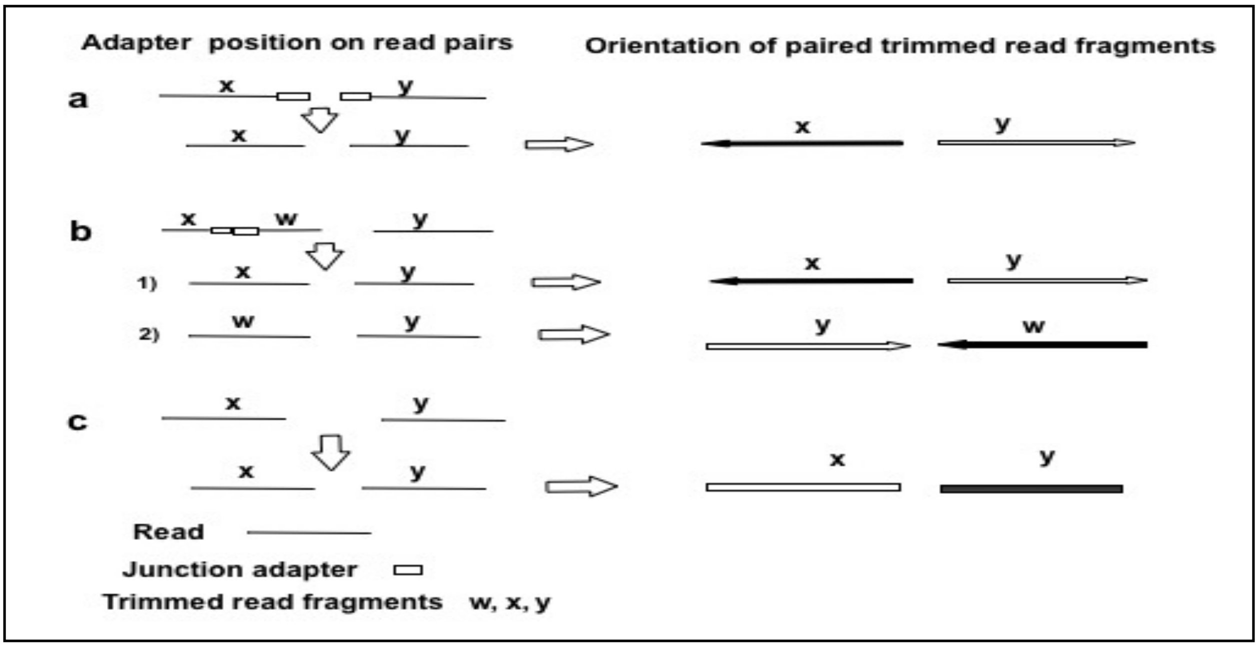
Adapter position on read pairs
Orientation of paired trimmed read fragments
a
x
y
x
y
x
y
b
x
w
y
1)
x
y
x
y
2)
w
y
y
w
c
x
y
x
y
x
y
Read
Junction adapter
Trimmed read fragments w, x, y

Figure 2-27: Orientations of junction adapter trimmed read pairs.

The resulting virtual libraries had the following percentages.

MP orientation: 35,212,151(50.17%).

PE orientation: 15,920,912 (22.69%).

“Unknown” orientation: 18,345,052 (26.14%).

Single end reads (orphaned reads): 14,380,568 (20.49%).

Read length distribution of the three-paired virtual libraries from a sample of 10,000 read pairs was determined for each read A and B (Figure 2-28 below).

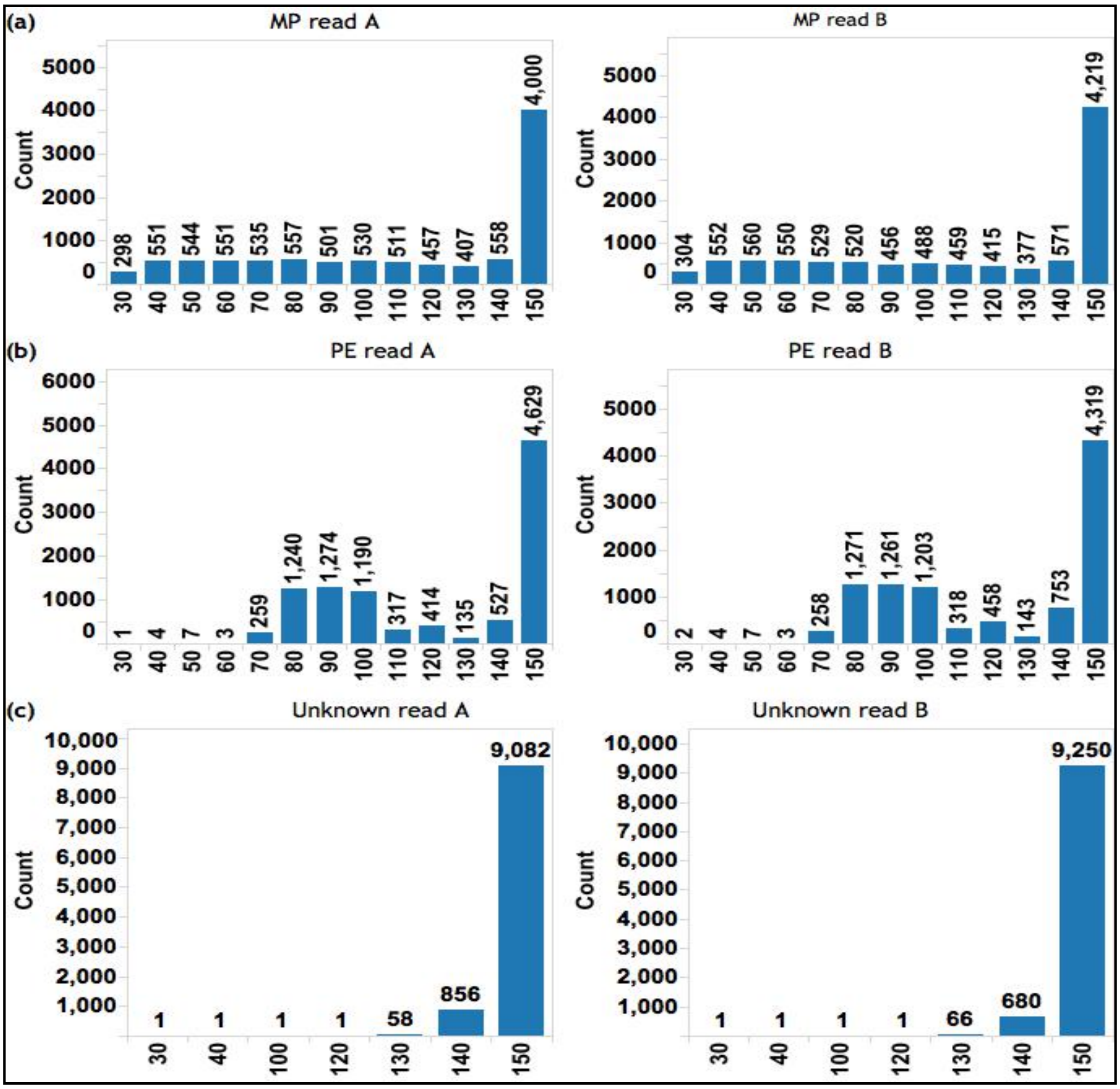

Figure 2-28: Junction adapter filtered read pairs read length distribution. For the three virtual libraries, PE, MP and Unknown, largest proportion of reads were of length 141-150 bps.

For the MP virtual library (Figure 2-28 **a**) ~40 % read pairs were between 141 bp -150 bp, ~20 % between 100 bp - 140 bp and ~40 % < 99 bp. For the PE virtual library (Figure 2-28 **b**) ~45 % were between 141 bp - 150 bp, ~25 % between 100 bp -140 bp and ~30 % < 99 bp. For the unknown (Figure 2-28 **c**) ~90 % were between 141 bp -150 bp, ~9 % between 100 bp -140 bp and ~1% < 99 bp. An important consideration when pre-processing mate pair libraries is the read length of the adapter trimmed reads. If the read lengths are too short, then read mapping specificity will be impacted given the repetitive nature of the wheat genome. Both adapter-trimming steps (Figure 2-26, Figure 2-28) resulted in ~96 % of the final reads having read lengths of between 140 bp - 150 bp. A plot of the insert sizes of the three virtual libraries when a sample of 10,000 read pairs from each of the adapter filtered reads were mapped to plate 1 contigs using BLAST (Altschul *et al.*, 1990) showed that the PE virtual library was a mate pair shadow library due to the small insert size of 200 bp (**Figure**

2-29 a) that is characteristic of mate pair shadow libraries. The "unknown" and MP virtual libraries both conformed to the expected library insert size of 6 Kb – 10 Kb (**Figure 2-29 a**, & **b**).

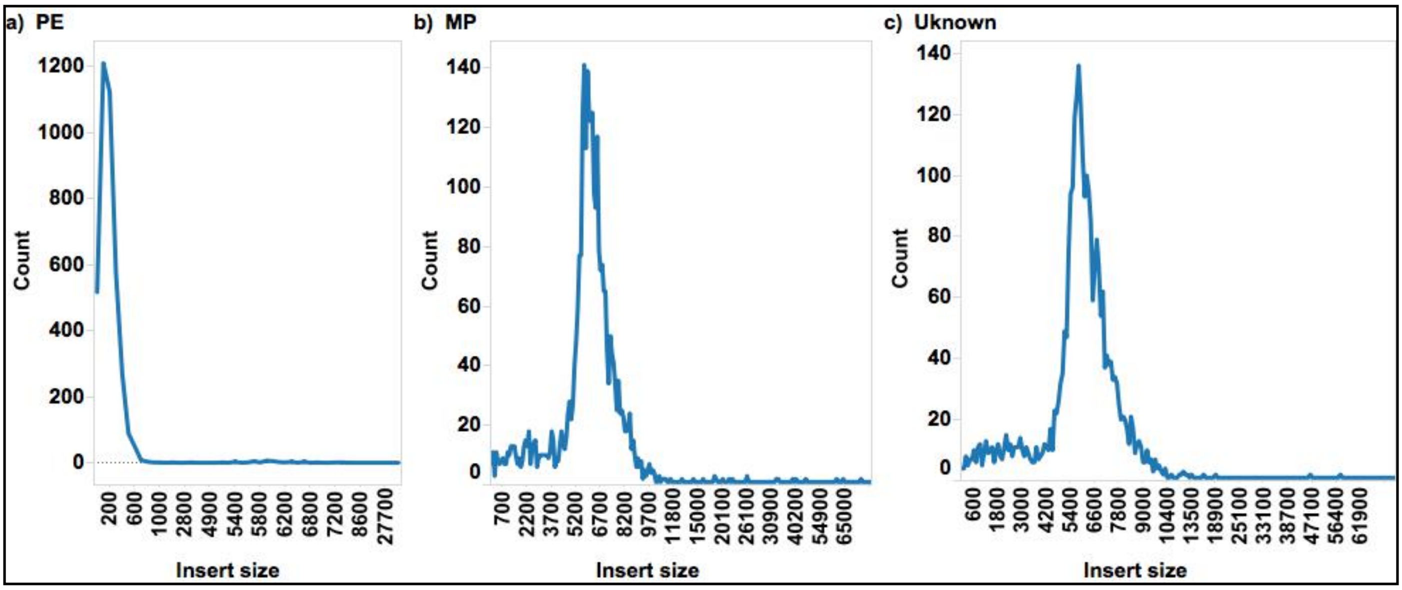

Figure 2-29:Insert sizes of the three virtual libraries, PE (a), MP (b), and Unknown (c).

Following the above approach, pre-processing of all 13 MP sequenced libraries with the established MP adapter filtering criteria resulted in percentages below (**Table 2-9**). After filtering external adapters, ~98 % - 100 % of the sequenced mate pairs are retained (EA%). 50 - 60% of the mate pairs (MP%) are retained after the removal of junction adapters and 22 -29% of the mate pairs had no junction adapters (Unknown%). Orphaned reads were ~20 - 33% (SE%) (**Table 2-9**).

Table 2-9: Mate pair adapter filtered read pairs. After filtering external adapters, ~98 - 100% of the sequenced mate pairs are retained (EA%). After the filtering of junction adapters, 50 - 60% of the mate pairs (MP%) are retained and 22 - 29% of the mate pairs had no junction adapters (Unknown%). Orphaned reads were ~ 20% - 33 % (SE%).

Plate	**Read pairs**	**EA %**	**MP %**	**Unknown %**	**PE %**	**SE %**	**Large insert pairs %**
1(4-6kb)	85,742,017	100	61	22	16	33	83
1(6-8kb)	70,386,715	100	50	26	23	20	76
2	15,461,980	100	59	25	15	33	83
3	14,725,143	98	58	25	16	32	83
4	12,560,129	100	55	29	15	31	84
5	15,675,565	100	59	25	16	32	83
6	11,566,541	99	56	29	15	31	84
7	14,778,328	99	59	24	16	32	82
8	13,192,646	100	58	26	15	32	84
9	14,045,620	100	59	24	16	32	83
10	13,262,494	99	58	25	16	32	83
11	13,570,093	100	59	24	16	32	83
12	14,355,862	100	60	23	16	33	82

For all the 12 plates of assembled BAC pools, MP adapter filtered reads were mapped using BLAST to their respective pool assemblies, each read mapped independently. MP reads were selected based on 100% sequence identity match, across the full length of the adapter filtered read with matches to the same pool as the corresponding paired mate. On average, the number of paired MP reads linking different contigs in a BAC were 2,036 with a median of 1,677, maximum of 24,888 and minimum of 0. A histogram of counts of links per BAC for all pools (**Figure 2-30**) showed that there was a peak corresponding to 97 pools that had between 0 - 10 MP reads linking contigs per BAC. Of the 97 pools, 52 pools (53%) had sequence coverage < 450 x. Of the remaining 45 pools, 11 pools had one contig per BAC, 14 pools two contigs per BAC, 11 pools three contigs per BAC and 9 pools had between 4 and 8 contigs per BAC. Thus, pools with fully assembled BACs and pools with low sequence coverage had fewer than 10 MP reads per BAC joining contigs. To limit over scaffolding, a minimum of 10 MP reads was thus required to join two contigs. This was in addition to the MP reads having correct orientation and insert size.

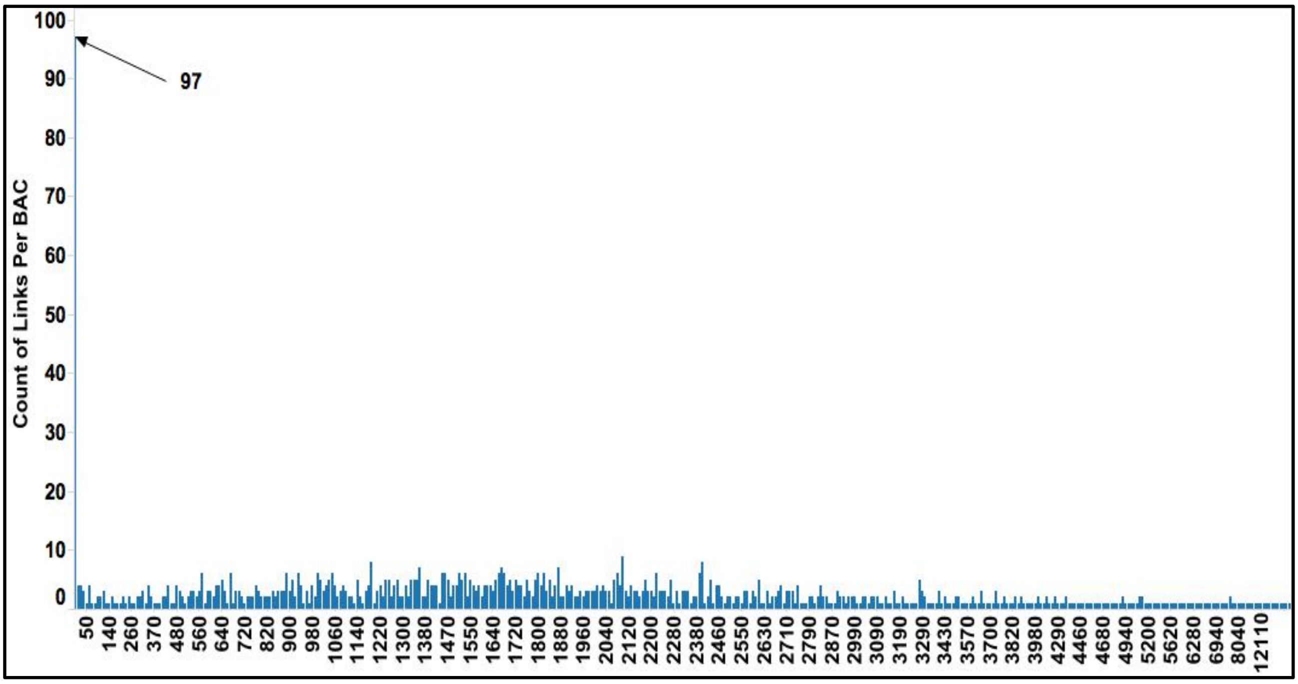

Figure 2-30: Distribution of MP read pairs linking contigs in BAC pools (bin=10).

For all the 12 plates of assembled BAC pools, scaffolding with large insert size libraries (**Table 2-9**) was performed using the SSPACE scaffolder. This was automated with Perl scripts for all the pools. Stringent scaffolding was ensured by setting the minimum number of mate pair links to scaffold two contigs together at 10 and the insert size set to 4 - 12 Kb based on the insert size plots (**Figure 2-29**

b & **c**). Scaffolding reduced contigs per BAC by an average of between 1.2 and 4.5 per plate. The N50 increased by an average of between 18 Kb and 76 Kb per plate. Final assemblies per plate now had an average of between 1.5 to 2.4 scaffolds per BAC and average N50 of between 77 Kb and 144 Kb per plate (**Table 2-10**).

Table 2-10: Assembly statistics and scaffolding per plate.

	Avg. coverage per BAC	**Avg. no of contigs per BAC**	**Avg. no of scaffolds per BAC**	**Avg. reduction in contigs**	**Avg. assembly N50 (Kb)**	**Avg. scaffold N50 (Kb)**	**Avg. increase in N50 (Kb)**
Plate 1	408	3.6	1.8	2	66	96	30
Plate 2	691	2.7	1.5	1	81	107	26
Plate 3	227	3.7	2.1	2	74	103	29
Plate 4	292	3.7	2.4	1	58	77	19
Plate 5	298	4.1	1.5	3	45	85	40
Plate 6	550	4.6	1.8	3	65	121	56
Plate 7	524	3.8	1.5	2	73	130	57
Plate 8	535	4.0	1.6	2	72	135	62
Plate 9	556	5.2	2.3	3	70	131	64
Plate 10	651	5.5	2.0	3	68	144	77
Plate 11	525	6.8	2.3	4	59	133	74
Plate 12	329	5.4	1.8	4	50	114	64

Post scaffolding, assemblies now had an improved average number of contigs per BAC of 1.5 (median 1.5) from 4.4 (median 3.5). The range in the number of contigs per BAC also reduced from 1 - 20 contigs per BAC to 1 - 16.5. The average N50 increased two fold from 65 Kb to 112 Kb while the number of ambiguous nucleotides incorporated (Ns) increased slightly from 0.1 % before scaffolding to 2.9 %.

2.2.7 A wheat chromosome arm 7DS super scaffold

Assembled and MP scaffolded BAC pools were further analysed to generate a super scaffold (pseudo molecule). The 7DS super scaffold was created using BAC overlaps within a MTP. The wheat chromosome 7DS MTP had 931 MTP contigs each containing between 1 to 33 BACs. From a total of 8,992 contigs, 4,340 contigs (48%) were merged into 876 super scaffolds (Table 2-11). 4,652 contigs could not be merged due to lack of overlaps. The final super scaffolds had a maximum contig size of 2 Mb and an N50 of 704 Kb. No BES mappings and overlaps were found for 31 MTP contigs. These MTP contigs had between 1 - 2 BACs each.

A comparison of how many contigs could be merged without considering the MTP and BES was done. From a total of 8,992 contigs, 3,571 contigs (39%) were merged into 1,328 super scaffolds

(Table 2-11), and 5,421 scaffolds could not be merged. The final super scaffolds had a maximum contig size of 463 Kb and N50 of 175 Kb.

Table 2-11: Statistics from generation of pseudo molecule with and without the use of MTP and BES.

	With MTP/BES	Without MTP/BES
Total scaffolds analysed:	8,992	8,992
Total scaffolds placed	4,340	3,571
Total scaffolds unplaced	4,652	5,421
Super scaffolds	876	1328
Total size (Mb)	391	193
Minimum contig size (Kb)	5	4
Maximum contig size: (Kb)	2000	145
Mean contig size (Kb)	447	145
N50 (kb)	704	175
Number of Ns	10,122,966 (2.5%)	3,906,617(2%)

2.3 Discussion

The results presented in this section follow the order of the materials and methods sections presented in section 2.2 above.

2.3.1 Selection of an assembler for wheat BAC assemblies

Comparisons of the three assemblers, SASSY, Velvet and Ray suggested that we had the best assembly contiguity when assembling wheat BACs using the SASSY assembler. Assemblies of seven individual wheat BACs with the SASSY assembler had the longest scaffolds, largest N50 with the least number of contigs compared to Velvet and Ray assemblers. Assembly sizes of SASSY and Velvet assemblies were within the 10% error rate of the MTP size estimation (Guo *et al.*, 1992), while the Ray assemblies were 1.5 x larger than the MTP average size (Table 2-3). Only the SASSY assembler was able to assemble two BACs into a single contig.

2.3.2 Optimal BAC sequencing depth and coverage

The optimal sequencing depth and coverage suitable for assembly with SASSY was determined using sugarcane BACs because sugarcane has a high repeat content similar to wheat and there were sugarcane BACs sequenced to > 10,000 x coverage available which enabled the assembly of subsets of reads from individual BACs at various coverage estimates. Although a similar experiment using wheat BACs would have been appropriate, the costs required to sequence a sufficient number of wheat BACs to > 10,000 x sequence coverage was prohibitive. Optimal coverage results using sugarcane BAC assemblies showed that > 450 x coverage was required for optimal BAC assembly. The variation in total assembly length observed for datasets greater than 900 x is likely to be due to the increase in number of erroneous reads confounding the assembly process. Sequencing errors contribute to the count of unique k-mers in reads, and as more reads are introduced to the assembler, the effect of sequencing errors increases. Given BAC cloning and sequencing library preparation protocols do not effectively filter out *E. coli* prior to sequencing, an accurate estimate of the sequencing depth to target should factor for the loss of coverage from *E. coli* filtration.

Assembly sizes levelling off at 450 - 900 x coverage suggest that no more reads could be incorporated into the assembly. This suggests that our assembly method can be applied to other crop genomes with similar or lower complexity to that of the sugarcane genome. For low coverage 0 -170 x coverage, no assemblies were obtained. This does suggest that the SASSY assembler is not suitable for the assembly of regions with low sequence coverage. Physical coverage and depth of coverage are both critical factors for optimal assemblies when using the SASSY assembler.

2.3.3 BAC pooling strategy

When BACs are sequenced to > 450 x, complete assembly of BACs can be achieved as evidenced with sugarcane pool assemblies in **Figure 2-5**. The assembly of seven single BACs, which had > 450 x sequencing coverage post *E coli* and vector sequence filtering, confirms this in wheat BAC pools, where the completeness of assemblies was evident as the assembly sizes agreed with those of the MTP estimates. Furthermore, assembly sizes from sequential additions of BACs in a pool was reflected in the cumulative sum of the MTP size estimates of the individual BACs in the pool. It appears that the number of BACs that can be pooled affects assembly contiguity, with larger pools resulting in more fragmented assemblies.

The number of BACs, which can be sequenced on a single lane of Illumina HiSeq is limited by the coverage required (450 x - 900 x) per BAC given a mean BAC length of 120 Kb, and data throughput from the Illumina HiSeq 2000 of around 40 Gb per lane. This suggests that pooling 384 BACs within a single lane, with accurate quantification and normalisation during BAC pooling should produce approximately 850 x coverage per BAC. Considering that BAC DNA will contain *E. coli* and vector sequences, the actual sequence coverage is likely to be less than 850 x but above 450 x, which is within our estimated range of 450 x to 900 x.

Evaluation of pools of four BACs using BES, showed that the structure of individual BACs was not compromised when assembling pooled BACs (**Figure 2-8**), and where a BAC assembled into a single contig, the BES were also correctly paired. Single BAC assemblies compared to the respective pooled assemblies confirmed that similar sequences were assembled in both cases (Figure 2-12 - 11). The presence of vector sequences in an assembly reflected the circular structure of a BAC clone where the assembly of a single BAC resulted in a single contig (Figure 2-16). Where a pool of BACs was assembled together the same did not apply as the resulting assemblies resulted in the assembled vector sequence joining two different BACs at the BAC ends. This is evident from Figure 2-18. The joining of different BACs together by the vector sequence in addition to vector fragments mapping to BAC ends necessitated the filtering of vector sequences. Assembling pools of non overlapping BACs sharing vector sequences would have resulted in mis-assemblies. Similarly, large repeats shared between two pooled BACs can result in chimeric assemblies. Using BES pairing, no such chimeric assemblies were observed in the BAC assemblies.

The assembly of pools of sequentially added BACs (A, AB, ABC to ABCEFGH) (**Figure 2-7**) showed an increase in assembly sizes equal to those of the added MTP sizes. No such trend in regard to the N50 and largest contig sizes was observed. Given the assembled BACs in this case (A, B, C, E, F, G and H) were non-overlapping and had vector and *E coli* sequences filtered out prior to assembly, such a trend if were present would have indicated pooled assemblies were chimeric as two or more BACs would have to co-assemble and thus progressively increase the N50 and longest contig values.

2.3.4 Optimised assemblies of wheat chromosome arm 7DS BACs

A comparison of the effect of Illumina TruSeq, Nextera and Nextera XT DNA library preparation kits on assembly outcomes was done. The results showed the TruSeq sequenced libraries provided the best assemblies despite having on average a marginally lower coverage of 648 x per BAC compared to Nextera XT libraries at 717 x per BAC. This in addition to the fact that both kits had sequence coverage above the determined minimum coverage of 450 x suggests library preparation has a significant effect on assembly outcomes. Nextera and Nextera XT protocols differ from TruSeq protocols mainly in the method used for cleavage of DNA fragments. Nextera and Nextera XT protocols use enzymes to cleave DNA fragments (Adey *et al.*, 2010) while TruSeq requires mechanical shearing. Illumina recommends the use of precision ultrasonicators for high throughput DNA shearing and size selection.

Although the use of enzymes for DNA fragmentation is quicker, their use has been shown to result in GC biases resulting in uneven (non random) coverage of targeted genomic regions (Marine *et al.*, 2011). In addition, although enzymatic protocols offer further advantages of being able to use low quantities of DNA as the starting material, such low quantities have been shown to skew the distribution of genomes represented in metagenomic samples (Marine *et al.*, 2011). The Nextera and Nextera XT protocols resulted in uneven coverage across pooled BACs when compared to TruSeq. Mechanical shearing is more random and results in more even fragment size distribution and coverage across pooled BACs. Unlike other assemblers, the SASSY assembly algorithm uses paired read information at the start of the assembly process during graph building and relies on the stability and even distribution of fragment sizes.

Assembling at longer read lengths contributed to longer and more contiguous assemblies. Even though assemblies at 91 bp had a higher average longest contig statistic by 2Kb (99 Kb vs 97 Kb) and larger average assembly sizes by 2 Kb (429 Kb vs 427 Kb) than assemblies at 100 bp, the standard deviation of the longest contig statistic for the 100 bp assemblies was half that of the 91 bp (16 Kb vs 30 Kb) (Table 2-6). This indicates that assemblies at 100 bp were more contiguous than those at 91 bp. This was also evident by the 100 bps assemblies having a higher average N50 compared to the 91 bp assemblies (Table 2-6). Assembling with longer read lengths lead to longer assemblies with higher N50's and fewer contigs. Where possible longer reads should be used during assembly, unless there is a need to use shorter (trimmed) reads as a result of contaminating sequences such as sequencing adapters or ambiguous nucleotides (Ns) that degrade the contiguity of the assembly.

Optimisation of SASSY assembly parameters using MTP sizes was explored. SASSY offset parameters set the maximum allowable length of the non-overlapping part of two overlapped reads during assembly. The naïve offset is enforced during the initial contig growing while the extension offset is enforced during contig extension, where previously assembled contigs with naïve offsets are extended. Accurately assembling highly repetitive genomes based on overlaps between short reads is challenging and optimising assemblies is thus critical. ANOVA tests showed there was significant variation in assembly lengths as a result of changing SASSY assembly parameters (Figure 2-21. The determination of parameters resulting in assembly lengths that differed the least from MTP sizes showed that assemblies at offsets '4,12' were optimal. This was done using paired t tests. Paired t tests showed there was no significant difference in assembly lengths between assemblies at offsets '4,12' and their MTP sizes. Offsets '2,10' produced significantly longer assemblies while assemblies at offsets '8,16' produced significantly shorter assemblies compared to MTP sizes. An assembly of all wheat MTP BACs was thus undertaken using pooled BACs with sequencing libraries prepared with the TruSeq kit and assembly at 100 bp, with SASSY assembly offset parameters of '4,12'.

2.3.5 Automation of wheat chromosome arm 7DS BAC-by-BAC assemblies

Assembly of 4,608 MTP wheat chromosome 7DS BACs was effectively automated through the SASSY pipeline. With a scripting solution, the pipeline is highly configurable and easy to set up and run once all dependencies are installed. Bottlenecks in efficiency were as a result of the lack of multi-threaded support when running the SASSY assembler, as it does not use multiple CPU's. The nature of the SASSY assembler also hinders the assembly of large genomes as this leads to an exponential increase in memory requirements. This limits the use of the assembler in assembling large target sequences. Multiple instances of the pipeline were run to increase the throughput of the pipeline.

During assembly, it was observed that pools with coverage < 450 x were more fragmented than those with coverage > 450 x. In addition, assemblies with coverage > 450 x had the pool assembly lengths in agreement with their respective MTP pool size estimates while pools with < 450 x coverage appeared to be shorter than their estimated pool MTP (Figure 2-23). Pools with low sequence coverage after filtering of *E coli* reads had a higher percentage of *E coli* sequences. This suggests that some of the BAC libraries had high contamination with *E coli* which diluted the amount of BAC sequence to less than 450 x. The percentage of unique reads in pools decreased with increase in coverage (Figure 2-24). This suggests that sequencing at high coverage does translate to more physical coverage of the sequenced genomic sample. Pools with low sequence coverage were flagged and new libraries were generated for re-sequencing.

2.3.6 Mate Pair scaffolding of wheat chromosome arm 7DS BACs

MP sequencing provides longer insert size information that spans repeats and low complexity regions of an assembly. MP protocols can currently produce read pairs that span between 3 to 40 Kb. The generation of MP libraries is more laborious and costly compared to paired end libraries. This is because the process requires additional molecular steps for cloning and circularization of DNA fragment ends. Furthermore, bias introduced during library preparation and sequencing of MP reads results in loss of coverage as effective MP reads with the ideal long insert sizes are reduced. Improvements in the accuracy and cost of MP sequencing have the potential of greatly improving assemblies.

Ideally for sequencing of complex repetitive genomes, long insert sizes such as those produced by MP libraries should be the standard and therefore complement paired end libraries. Sequencing to high coverage large insert libraries and assembling these with SASSY would probably result in assembly across repeats that are less than the MP insert size. The MP adapter pre-processing steps presented reduced the loss of paired reads. Processed MP reads remained long (141 bp - 150 bp) and this may have contributed to increased specificity of read mapping and scaffolding accuracy. The scaffolding efficiency was on average 1.5 contigs per BAC. Insert size plots of MP adapter filtered read pairs onto single contigs prior to scaffolding agreed with the MP library's insert sizes, validating the contiguity of the BAC assemblies.

With an average of 2,036 MP reads per BAC available for scaffolding, a strict scaffolding stringency was set to avoid over scaffolding. Pools with 0 - 10 MP reads linking contigs (links) per BAC were the majority with a count of 97 (**Figure 2-30**). These pools had either a sequence coverage < 450 x or were fully assembled BACs. This suggests that the low numbers of MP read pairs (links) available for scaffolding these pools was as a result of lack of fully assembled regions for the MP reads to map or there were fewer contigs to scaffold due to a more contiguous assembly. Setting the minimum number of links required to scaffold to 10 would thus avoid over scaffolding. During scaffolding, further stringency was enforced in that MP reads had to satisfy an insert size of between 4 Kb - 12 Kb and the expected MP mapping orientations. This further reduced the effect of the MP shadow library and chimeric reads that are expected in MP sequenced reads. These conditions in addition to scaffolding pools only with MP read pairs that mapped to the same pool reduced the effect of repeats during scaffolding.

2.3.7 The generation of wheat chromosome arm 7DS super scaffold molecule

The generation of the 7DS super scaffold entailed identification of BACs within a pool using BES, identification of sequence overlaps within overlapping BACs and merging of overlapping sequences to generate a consensus super scaffold. Manual curation in genome finishing, especially in wheat is important due to the high repeat content of the wheat genome that confounds automated analysis. Generating the super scaffold molecule by manual curation using the MTP, BAC overlaps and BES enabled the merging of 10% more contigs (4,340/8,992) compared to generating the same molecule without the use of the MTP and BES which resulted in merging 3,571/8,992 scaffolds. In addition, there was a corresponding 4 x increase in N50.

No suitable tools were available that could accurately generate a pseudo molecule using a MTP ordered list of BACs and BAC sequence overlaps with corresponding BES. Although time consuming, manual curation was undertaken. The development of a tool to automate this process would have taken considerable time and effort. As advances in sequencing technologies and assembly algorithms continue to be made, there will be better and more accurate algorithms that are suitable for the assembly of complex polyploid genomes.

2.4 Summary

Sequencing of BAC pools to > 450 x coverage per BAC and assembling pools with the SASSY assembler provides accurate and highly contiguous assemblies. The selection of sequencing library preparation kits and assembly parameters affect the outcome of an assembly and optimal kits and assembly parameters should be determined in addition to sequence coverage. Mate pair reads and BES can also be used to not only improve assembly contiguity, but also evaluate the quality and accuracy of assemblies before and after scaffolding.

3 Validation of bread wheat chromosome arm 7DS BAC-by-BAC assemblies

A robust validation process should ideally limit bias. Bias in assembly validations may be introduced by the use of a previous assembly or from using assemblies of related species that may have errors in addition to having differences compared to the newly assembled genome. Bias in assembly validation may also be introduced through the reliance on a single sequencing platform or dataset. *De novo* approaches for assembly validation present an opportunity to determine how true an assembly is representative of the actual genome.

The validation of bread wheat chromosome arm 7DS BAC assemblies involved the use of four 7DS specific genomic datasets generated using different platforms. These were; BAC end sequences (BES), minimum tilling path (MTP), Illumina long insert mate pair libraries (MP) and a 7DS optical map. The above resources were generated using the same genomic DNA as that of the BAC assemblies, specifically, isolated chromosome arm 7DS DNA. In addition to using the same genomic DNA, the four datasets above were generated using independent methods and platforms. Sanger sequenced BESs were used to identify contig BAC ends of assembled BACs, to assign contigs in pooled assemblies to their respective BACs and to validate the order of overlapping BACs in a MTP. The MTP was used to validate the order of overlapping BACs and size of assembled BACs and BAC pools. MP datasets were used to validate the long-range accuracy and contiguity of assembled contigs and to further scaffold contigs within an assembled pool. The 7DS optical map enabled the assessment of contig accuracy using *BspQ1* restriction profiles.

3.1 Materials and methods

3.1.1 Assembly contiguity of 7DS BAC assemblies using mate pair libraries

Mate pair datasets from pooled BACs, as described in materials and methods section 2.1.5 were used in this analysis. The MP datasets were used to assess the accuracy of the scaffolding approach. Contigs and scaffolds from 96 BAC pools from plate one in addition to the *E coli* str K-12 reference genome were used in this validation. The *E coli* str K-12 was used as a control as this was the *E coli* strain used for the maintaining the BAC clone libraries.

To assess the scaffolding accuracy, MP datasets corresponding to plate one's pooled BACs (see sections 2.1.5 and 2.2.5) were used. The MP reads were mapped to plate one's BAC assemblies, plate one scaffolds and the *E coli* genome reference as a control. 70,182,178 plate one MP reads were mapped to plate one contigs and scaffolds using BLAST (Altschul *et al.*, 1990). Read mates (A and B) were mapped independent of each other and hits later paired during analysis. To mitigate the effect of repeats, only reads mapping with 100% sequence identity, to one locus and a match length of 150 bp were considered. A Perl script, Estimate_chimeras.pl was used to filter the BLAST mapping results, pair hits and identify the mapping orientation of mate pairs on plate one contigs, plate one scaffolds and on the *E coli* reference.

Further evaluation of assembly accuracy, contiguity and the identification of mis-assemblies was done through plotting of PE and MP read insert sizes vs their position on contigs. The mapping orientations of the PE and MP reads on contigs was determined and visualized using Perl and R scripts.

3.1.2 Minimum tilling path (MTP) conformity of 7DS BAC assemblies

A MTP presents an opportunity to evaluate the sequence identity and order of BAC overlaps in a BAC-by-BAC assembly. Most importantly, a MTP provides a physical map onto which contigs can be anchored post assembly. The wheat chromosome arm 7DS MTP was developed by Hana Simkova's lab (Šimková *et al.*, 2011) using a 7DS specific BAC library of fingerprinted clones with BAC contigs assembled using FPC software (Soderlund *et al.*, 1997; W., Nelson and Soderlund, 2009). The MTP consisted of 931 contigs each with between 2 - 40 overlapping BACs. The MTP covered 95% of the 7DS chromosome arm and comprised 4,608 BAC clones.

For evaluation, four randomly selected MTP contigs 190, 162, 3945 and 307 representing a total of 25 overlapping BACs were used to ascertain if the corresponding assembled BACs conformed to the MTP. For each MTP contig, assemblies from each of the BACs constituting that MTP were compared pairwise. The contigs were compared using the BLAST algorithm to determine contig overlaps. The respective BES's were also used to determine the BAC ends. For each hit, the longest matches were selected for comparison. For identified overlaps between contigs, their order and sequence identity across overlapping regions was assessed. For each of the pools containing BACs constituting a MTP contig, the following was also determined. The total assembly size of the pool, the *E coli* content of sequenced reads in that pool, the total MTP size of all BACs in that pool, the difference between the MTP and the total assembly size and the coverage per BAC estimated for that pool. The number of overlapping BACs in all the MTP contigs was also determined.

3.1.3 Assembly accuracy of 7DS BAC assemblies using the optical map

Optical maps assist in the detection of structural variation, scaffolding of genome assemblies and the accurate estimation of gaps and indels. We employed a bread wheat chromosome 7DS optical map from BioNano genomics (www.bionanogenomics.com) to access the accuracy of the BAC assemblies. The optical map was developed using high molecular weight DNA from isolated chromosome arm 7DS. This offered us an unbiased opportunity to critically evaluate the 7DS BAC assemblies. The high molecular weight DNA provided long-range DNA molecules suitable for optical imaging and the generation of an optical map based on *BspQ1* restriction sites. The number and size distribution of long-range DNA molecules used in the generation of the optical map were determined using an algorithm developed by BioNano genomics and implemented within the IrysView software (http://www.bionanogenomics.com/products/irysview/). The optical map data was loaded into IrysView and molecule size and distributions plotted within the IrysView interface. Given the high repeat content of the wheat genome, the number and distribution of repeat units (repeating series of two or more equally spaced *BspQ1* restriction sites) in the long range DNA molecules was similarly determined using the IrysView software interface. Statistics of assembled *BspQ1* restriction profiles (consensus maps) were calculated.

The chromosome arm 7DS optical map was used to evaluate MTP contigs of BAC assemblies. Restriction profiles of BAC assemblies from two MTP contigs 307 and 190 were compared to the optical map. Given the optical map does not include sequence information, we further assessed the restriction profiles of known overlapping BACs that were sequenced individually and not in pools of non overlapping MTP BACs. MTP contig 190 previously assessed for sequence identity in section 3.1.2 was compared with the optical map information. Contigs from BAC pools containing overlapping BACs belonging the MTP were loaded into the IrysView software and *in silico* digested to generate restriction site maps. The generated restriction site maps were then anchored to the 7DS optical map. To further evaluate the applicability of the optical map in identifying mis-assemblies, five overlapping 7DS BACs were sequenced to > 600 x coverage, with an insert size of 300 bp on the Illumina HiSeq 2000. The BACs were individually sequenced and assembled with the SASSY assembler and overlaps between contigs identified using BLAST (Altschul *et al.*, 1990). All contigs from the assemblies were then loaded into the IrysView software and *in silico* digested to produce restriction site maps. The restriction maps were then compared to the 7DS optical map using the IrysView software.

To test the anchoring of BACs onto the optical map, 68 fully assembled BACs, all with correctly paired BES, were selected. The 68 BAC contigs were loaded into the IrysView software and *in silico* digested to generate restriction maps. The resulting map was compared to the 7DS optical map using the IrysView software. The sizes and restriction profiles of contigs that could be anchored and those that couldn't were compared.

3.1.4 Anchoring of 7DS BAC assemblies to the optical map

An evaluation of the anchoring of all the 7DS BAC pool assemblies to the optical map was performed. All 7DS BACs un-scaffolded pool assemblies, mate pair scaffolded assemblies and pseudo molecules were anchored to the optical map. Assemblies were loaded into the IrysView software and *in silico* digested with a BioNano supplied "Knickers tool" on a remote server to produce restriction site maps. The restriction maps were then compared to the 7DS optical map. A comparison of the mapping percentages was done. The number of anchored contigs in the pre scaffolded assemblies, mate pair scaffolded assemblies and the pseudo molecule were compared.

3.2 Results

For ease of reference, the results presented in this section follow the same order as the material and methods sections 3.1.1 to 3.1.3.

3.2.1 Assembly contiguity of 7DS BAC assemblies using mate pair libraries

The mapping of MP plate one reads to both plate one pre scaffolded contigs, and plate one scaffolded contigs (scaffolds) was done to evaluate the accuracy of scaffolding. The accuracy of scaffolding was evaluated by analysing the mapping orientations and insert size of mapped MP reads. Orientations of the MP mapped reads are designated as follows. Forward forward mapping pairs (--> -->) as (FF), forward reverse mapping pairs as (--> <--) as (FR), reverse forward mapping pairs as (<-- -->) as (RF) and reverse reverse mapping pairs (<-- <--) as (RR). The expected MP library mapping orientations is (RF), and mapping orientations (RR, FF) are as a result of chimeric reads. The MP reads were also mapped to the *E coli* genome reference as a control. The *E coli* reference was used as a control because the sequenced MP reads contained *E coli* sequences and the *E* coli reference is well established. From a random sample of 4.5 M read pairs, only reads matching with 100% identity, to one loci and a match length of 150 bp were considered. Mapping counts, orientations and percentages are shown in table Table 3-1 below. The insert size distributions of the above orientations are shown in Figure 3-1 to

Figure 3-3 Table 3-1 below. To show the distribution of low mapping percentages (1% - 3%) in Table 3-1 below, figures in Figure 3-1 to Figure 3-3 show both a normal distribution (a) and log distribution (b) of MP reads mappings on *E coli*, plate one contigs and plate one scaffolds respectively.

Table 3-1: Mapping percentage, orientation and average insert size of MP reads mapped to plate 1 contigs, scaffolds and *E coli* as a control.

Orientation	Reference	% of pairs	Median insert size	Counts of pairs
RF	*E coli*	97	5,959	306,600
	Plate 1 Contigs	98	5,905	3,472,460
	Plate 1 Scaffolds	96	5,903	3,745,863
FR	*E coli*	1	864	4,568
	Plate 1 Contigs	1	1,860	39,172
	Plate 1 Scaffolds	3	83,786	101,017
FF/RR	*E coli*	1	2,615	4,504
	Plate 1 Contigs	1	2,063	30,643
	Plate 1 Scaffolds	1	2,387	34,899

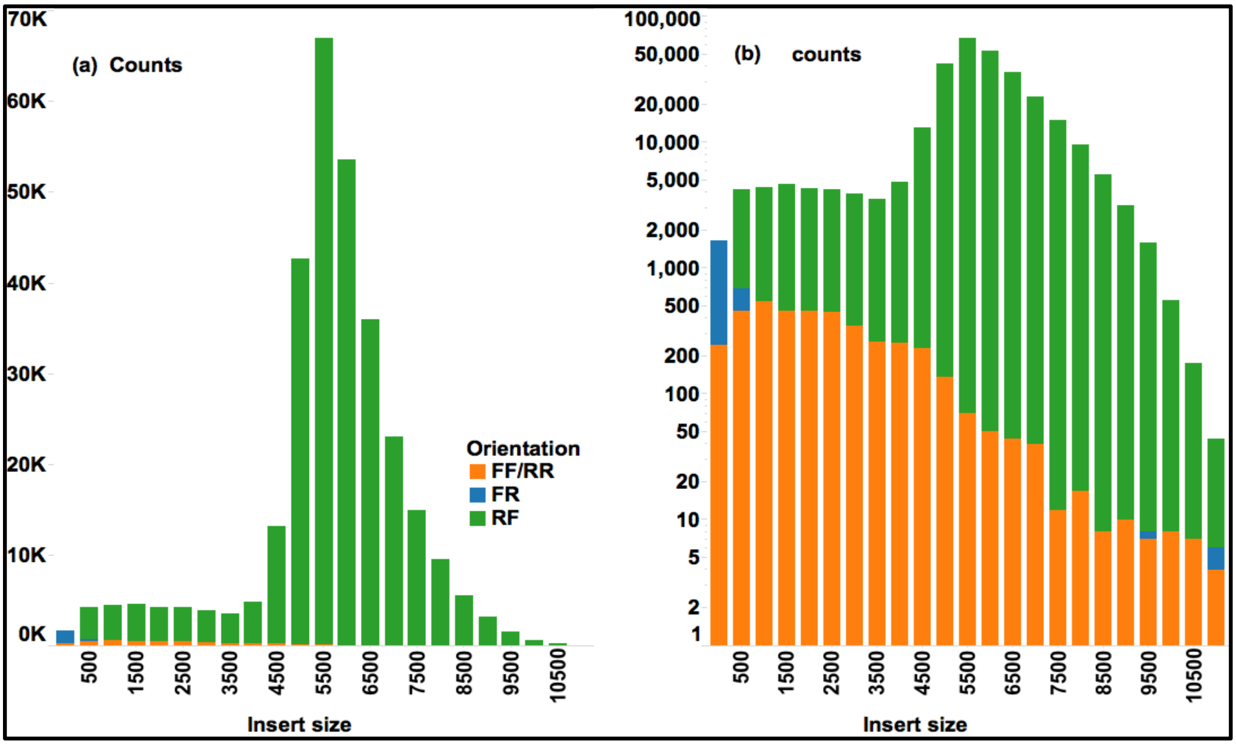

Figure 3-1: MP read mapping orientations and inserts size distributions on *E coli*. (a) Counts vs. insert sizes. (b) Log counts vs. insert size. Shadow library in blue (FR) and chimeric orange (FF and or RR) distributions.

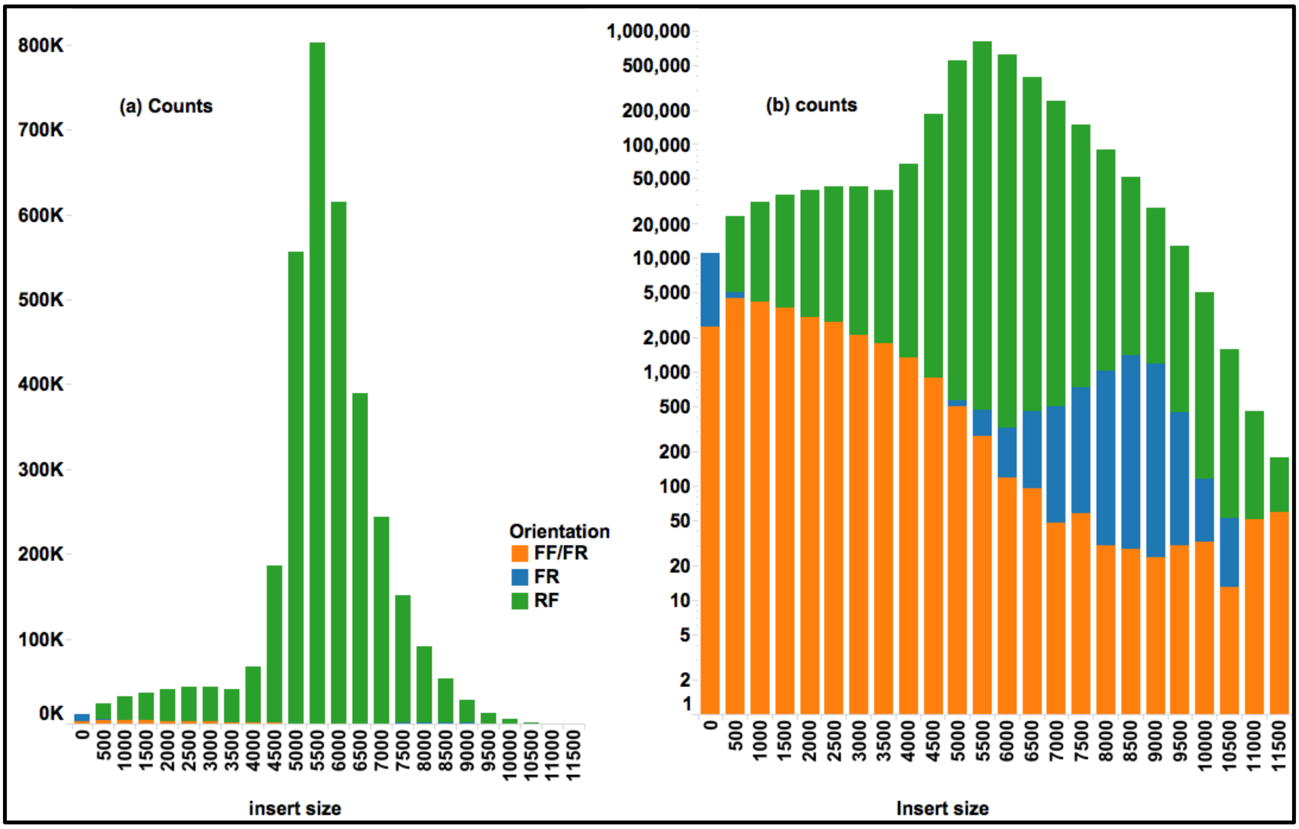

Figure 3-2: MP read mapping orientations and insert size distributions on Plate 1 contigs. (a) Counts vs. insert sizes. (b) Log counts vs. insert size. Shadow library in blue (FR) and chimeric orange (FF and or RR) distributions.

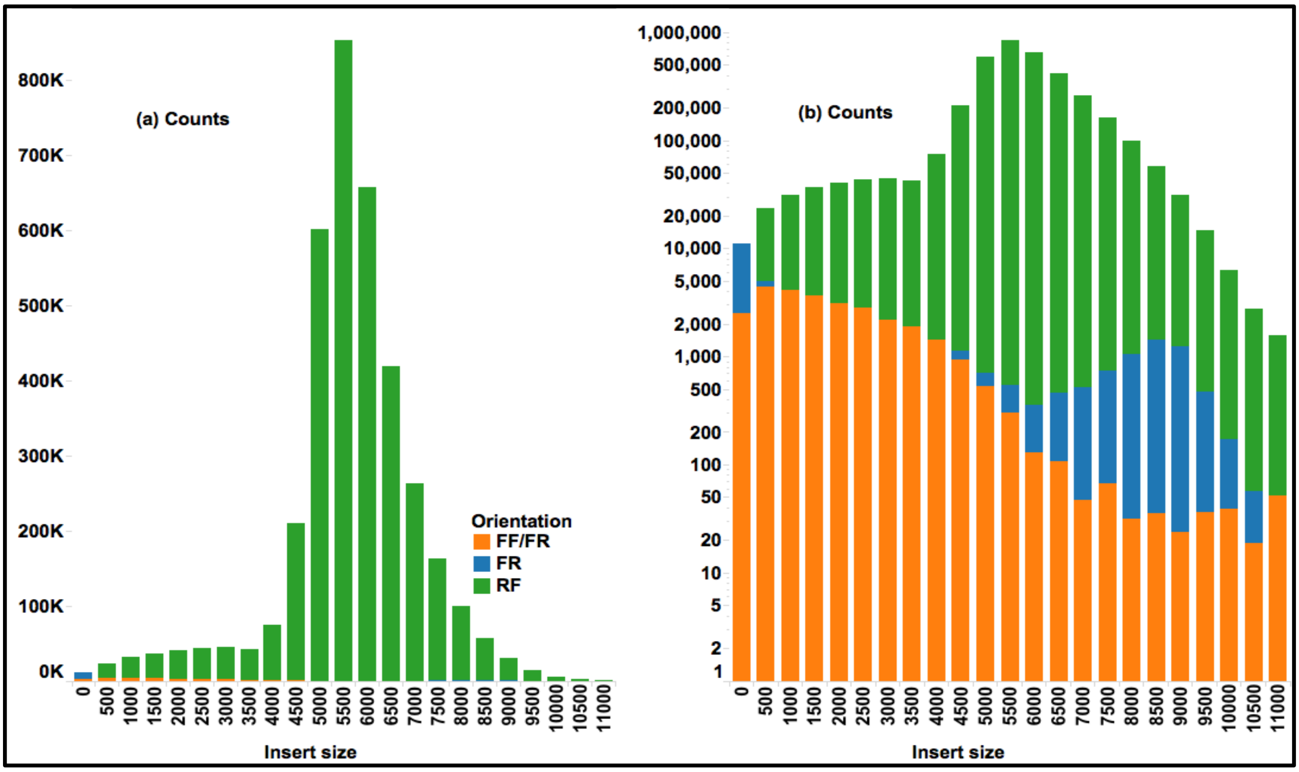

Figure 3-3: MP read mapping orientations and insert size distributions on Plate 1 scaffolds. (a) Counts vs. insert sizes. (b) Log counts vs. insert size. Shadow library in blue (FR) and chimeric orange (FF and or RR) distributions.

Correct MP mapping orientations (RF) on *E coli* and plate one contigs were similar at 97% and 98 % with a slight reduction for the scaffolded contigs at 96% (Table 3-1). The median insert size for the correctly orientated mate pairs (RF) on *E coli*, plate one contigs and plate one scaffolds were also similar at 5,959 bp, 5,905 bp and 5,903 bp respectively (Table 3-1). The Insert size distributions of the correctly oriented pairs (green bars in Figure 3-1 to Figure 3-3) were also similar on *E coli* (Figure 3-1), plate one contigs (Figure 3-2), and plate one scaffolds (Figure 3-3). Plots of the insert size versus position of MP and PE reads on contigs showed MP and PE reads were evenly distributed with the expected insert sizes and orientation on contigs *Sc-4_L_126160* and *Sc-4_L_107350* from pools 7DS-2-I23-K23-M23-O23 and 7DS-2-A10-C10-E10-G10 respectively (Figure 3-5 -Figure 3-5).

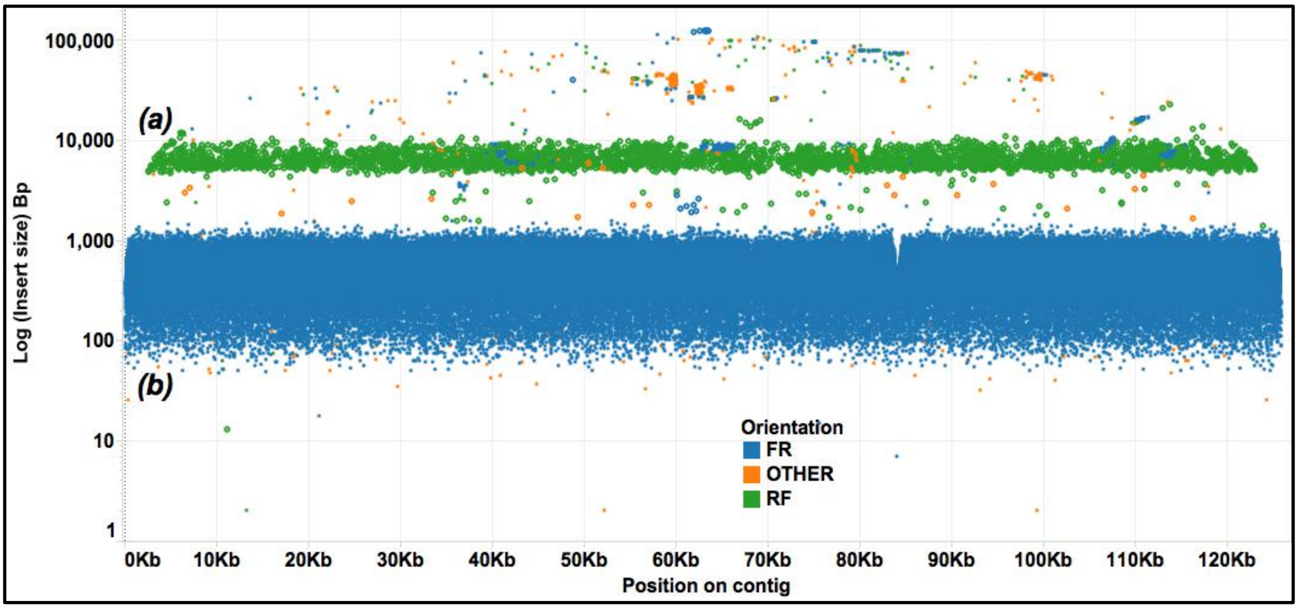

Figure 3-4: Plot of insert size vs position of MP (a) and PE (b) read pairs with the corresponding correct orientations of RF (green) and FR (blue) respectively on contig *Sc-4_L_126160* from pool 7DS-2-I23-K23-M23-O23.

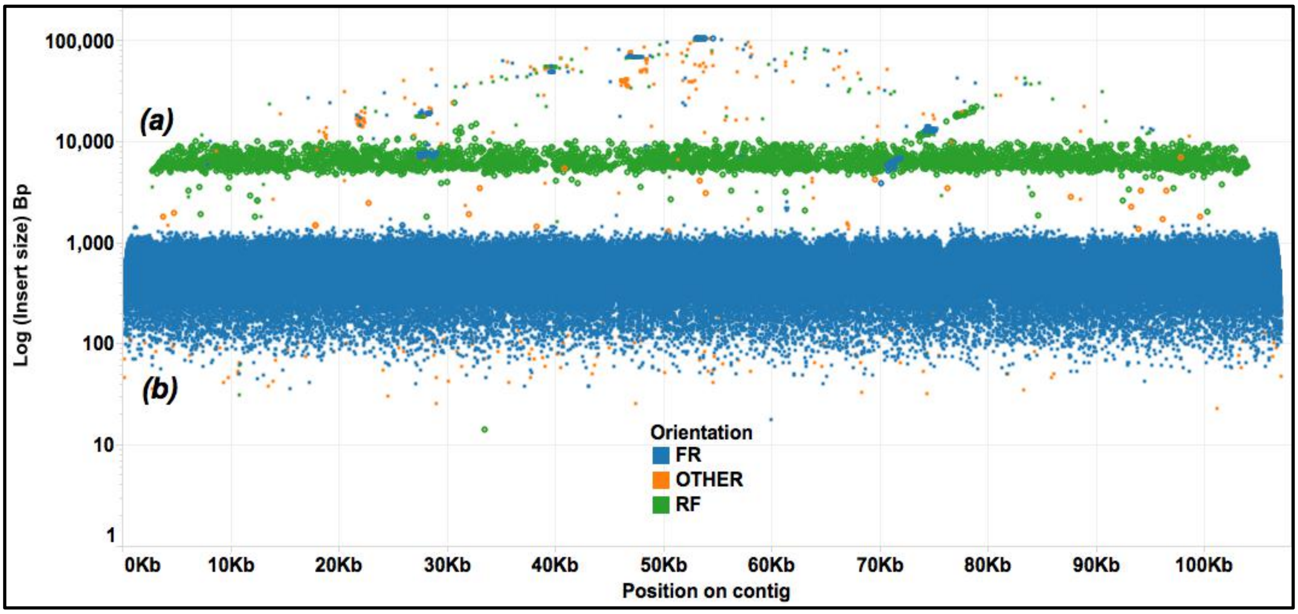

Figure 3-5: Plot of insert size vs position of MP (a) and PE (b) read pairs with the corresponding correct orientations of RF (green) and FR (blue) respectively on contig *Sc-4_L_107350* from pool 7DS-2-A10-C10-E10-G10.

The sequencing of MP reads involved sequencing of pools of BAC pools. To select MP reads for scaffolding, MP reads were mapped to assemblies of pools from which the MP reads were sequenced. Only paired MP reads mapping to the same pool were retained for scaffolding that pool. Overlaps within adjacent BACs in a MTP were expected. MP reads that mapped across pools were ignored to increase specificity and accuracy of scaffolding, as such those mapping across MTP overlapping

BACS were also ignored. This is evident in the MP and PE plot of contig *Sc-1_L_164859* from pool 7DS-2-A8-C8-E8-G8 (Figure 3-6) which had MP reads (Figure 3-6 ***(a)***) that mapped to overlapping region (Figure 3-6 ***z***) excluded, but had PE reads mapping to the same region. PE reads were sequenced per pool and mapped to the same pool. There was even coverage of PE reads with the correct orientation and PE insert size on the overlap region (Figure 3-6 ***z***).

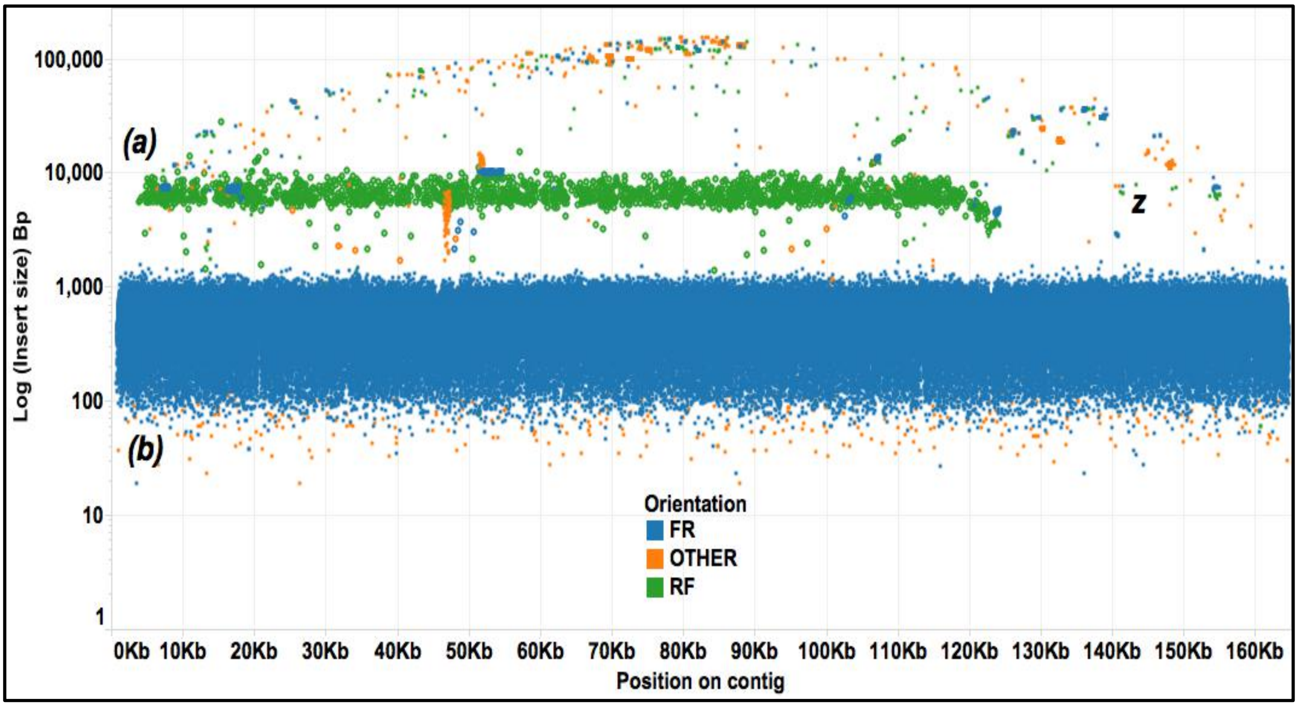

Figure 3-6: Plot of insert size vs position of MP (a) and PE (b) read pairs with the corresponding correct orientations of RF (green) and FR (blue) respectively on contig Sc-1_L_164859 from pool 7DS-2-A8-C8-E8-G8. Region z is a MTP overlap.

Complex repetitive regions with repeats that were greater than the PE insert size (500 bp) were confirmed to be correctly assembled based on MP read pairs spanning such regions with the correct insert size and mapping orientation. A plot of the insert size versus position of MP and PE reads on contig *Sc-1_L_128543* from pool 7DS-2-A10-C10-E10-G10 (Figure 3-7) showed there were gaps in coverage (***x*** and ***y*** in Figure 3-7) of PE reads (Figure 3-7 ***(b)***), but these gaps were spanned with the longer insert size MP reads (Figure 3-7 ***(a)***). Repeat regions were identified by an increase in both the insert size and erroneous orientation of either MP or PE reads in that region. On contig *Sc-1_L_128543,* this was evident by wrongly orientated MP reads with a higher insert size (orange peak above region ***x*** in Figure 3-7) and wrongly orientated MP reads with a higher insert size (blue above region ***y*** in Figure 3-7).

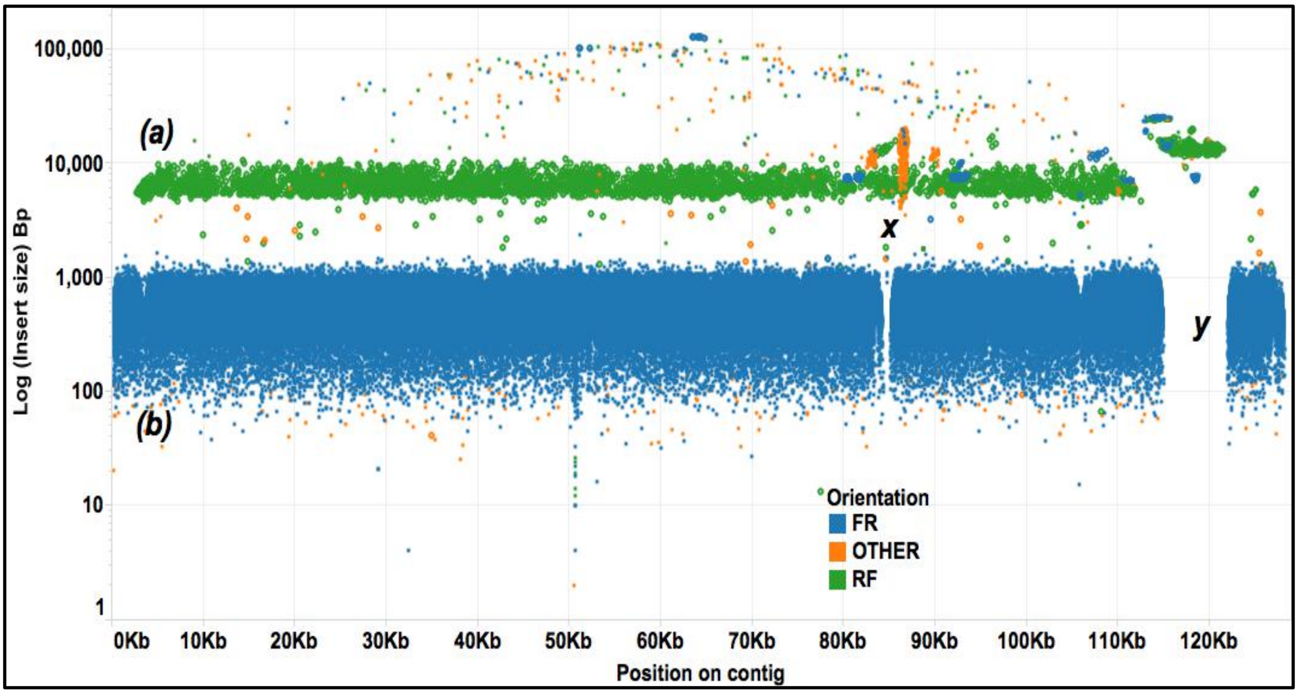

Figure 3-7: Plot of insert size vs position of MP (a) and PE (b) read pairs with the corresponding correct orientations of RF (green) and FR (blue) respectively on contig *Sc-1_L_128543* from pool 7DS-2-A10-C10-E10-G10.

The shadow library is an observed characteristic of MP sequenced reads. The shadow library results from a small fraction of un-circularized fragments during mate pair library construction remaining in the sample. Such fragments have a low insert size distribution < 600 bp, and have a paired end (FR) orientation. The percentages of the shadow library mappings on *E coli*, plate one contigs and plate one scaffolds were 1%, 1% and 3% with median insert sizes of 864 bp, 1,860 bp and 83,786 bp respectively Table 3-1, Figure 3-1 to Figure 3-3. MP reads mapping with an FF and RR orientation on *E coli,* plate one contigs and plate 1 scaffolds were all equal at 1% with median insert sizes of 2,615, 2,063 bp and 2,387 bp respectively. The percentage of FF and RR reads remain constant despite an increase in mapping counts of 4,504 pairs, 30,643 pairs and 34,899 pairs on *E coli,* plate one contigs and plate one scaffolds respectively.

Based on *E coli* mappings as a control, these results suggest that the MP dataset has approximately 2% of the reads consisting of a shadow library and chimeric reads. This is in agreement with the estimated chimeric percentage of MP reads reported by Illumina using their Nextera MP protocols (www.illumina.com).

3.2.2 Minimum tilling path (MTP) conformity of 7DS BAC assemblies

The assessment of the MTP conformity of BAC assemblies revealed the following. For MTP contig 190 (Figure 3-8 **a**), overlaps of six BACs E9, J12, D1, I22, P20 and G14 were compared. Each of these six BACs had been sequenced in pools containing three other non-overlapping BACs based on the earlier determined pooling strategy (see 2.3.3). These pools are listed in Table 3-2 with their corresponding coverage per BAC, MTP sizes, assembly sizes, difference between the MTP estimate and assembly sizes and the percentage of *E coli* reads in the pools.

Contigs from BACs E9, J12, D1, I22, P20 and G14 overlapped as expected (Figure 3-8 **b**), based on the MTP (Figure 3-8 **a**). Overlap regions numbered ***1 - 10*** in Figure 3-8, were 99% - 100% identical. Two BACs assembled as a single contig (BACs E9 and G14) (Figure 3-8 **b**). Overlap numbered ***1*** between BACs E9 and J12 (Figure 3-8 **b**) was 56,828 bp long with 100% sequence identity. Overlaps numbered ***2*** and ***3*** between BACs J12 and D1 (Figure 3-8 **b**) were 23,262 and 5,104 bp long respectively and both had 100% sequence identity. Overlaps numbered ***4, 5, 6*** and ***7*** between BACs D1 and I22 (Figure 3-8 **b**) were 23,301 bp, 8,279 bp, 2,394 bp and 9,811 bp long respectively and all the four overlaps had 100% sequence identity. Although the overlaps between BACs D1 and I22 were 100% identical, they were fragmented. Overlap numbered ***8*** between BACs I22 and P20 (Figure 3-8 **b**) was 5,340 bp long with 100% sequence identity. Overlaps numbered ***9*** and ***10*** between BACs P20 and G14 (Figure 3-8 **b**) were 21,529 and 21,836 bp long respectively and both overlaps had 100% sequence identity. BACs J12 and P20's pool had sequence coverage per BAC of 300 x and 379 x. Even though this coverage was below the earlier estimated coverage per BAC of > 450 x, these BACs assembled correctly as evidenced by the position and overlaps with other BACs in the MTP and BES mappings. BACs P20 and J12's individual MTP sizes estimates were 90 Kb and 96 Kb respectively, less than 120 Kb, suggesting that coverage was sufficient.

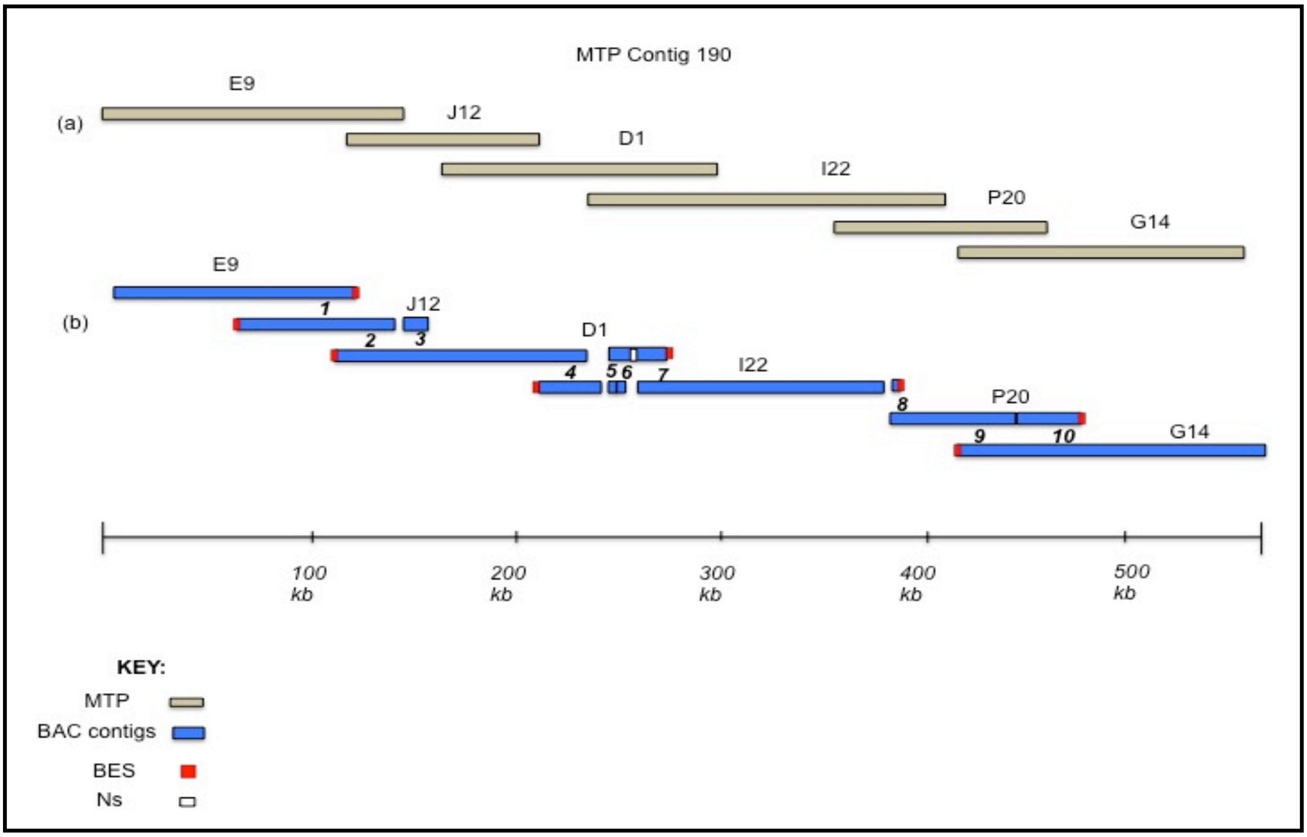

Figure 3-8: Overlaps from BACs constituting MTP contig 190. Image drawn to scale. Overlaps are denoted from 1-10.

Table 3-2: Pool assembly statistics of MTP 190 BACs.

BAC	Pool name	Cov per BAC	Assembly size (Kb)	MTP- size (Kb)	MTP - assembly size (Kb)	% E coli
D1	Ta_7DS-7-B1-D1-F1-H1	579	482	562	80	7
E9	7DS-10-A9-C9-E9-G9	955	573	491	-82	5
G14	Ta_7DS_7DS-6-A14-C14-E14-G14	535	477	497	20	7
I22	Ta_7DS_7DS-6-I22-K22-M22-O22	594	500	501	1	11
J12	7DS-MTP-P5-J12-L12-N12-P12	300	351	375	24	37
P20	7DS-MTP-P5-J20-L20-N20-P20	379	439	406	-33	37

For MTP contig 162 (Figure 3-9 a), five overlapping BACs P3, O24, J12, L18 and D6 were compared. Pools containing the five BACs are listed in Table 3-3 with their corresponding coverage per BAC, MTP sizes, assembly sizes, difference between the MTP estimates and assembly sizes and the *E coli* percentage in the respective sequenced BAC pools. Contigs from BACs P3, O24, J12, L18 and D6 overlapped as expected (Figure 3-9 b), based on the MTP Figure 3-9 a). Unlike MTP contig 190, overlap regions between BAC contigs in MTP contig 162, numbered 1 - 15 (Figure 3-9 b), had between 83% to 100% sequence identity. Overlaps 3, 5 and 9 (Figure 3-9 b) had 99%, 83% and 99 % sequence identity respectively. Other overlaps had 100% sequence identity. Overlap numbered 1 between BACs P3 and 024 (Figure 3-9 b) was 19,334 bp long with 100% sequence identity. Overlaps numbered 2, 3, 4, 5, 6 and 7 between BACs 024 and J12 were 11, 312 bp, 8,132 bp, 6,713 bp, 12,044 bp, 5,596 bp and 1,998 bp long respectively. Among these overlaps, overlaps numbered 3 and 4 had 99% and 83 % sequence identity respectively. Overlaps numbered 2, 5, 6 and 7 all had 100% sequence identity. BAC O24 assembled as a single contig while BAC J12 had gaps where sequences were expected in the overlap region (Figure 3-9 b). Both BACs J12 and L18 had fragmented assembly compared to the adjacent BACs. Overlaps numbered 8, 9 and 10 all between BACs J12 and L1 were 10,637 bp, 6,971 bp and 5,023 bp long and had 100%, 99% and 100 % sequence identity respectively (Figure 3-9 b). Overlaps numbered 11, 12, 13, 14 and 15 between BACs L1 and D6 were 4,380 bp, 5,000 bp, 10, 340 bp, 4, 238 bp and 8, 268 bp long respectively with 100% sequence identity. Overlap numbered 16 between BAC L1 and D6 was 11, 698 bp long with 85% sequence identity.

BAC J12's pool had sequence coverage per BACs of 398 x, below the expected threshold of > 450 x but the assembled pool size was 574 Kb, 83 Kb larger than the MTP estimate of 657 Kb (Table 3-3). The *E coli* content was also low as expected at 5%. This suggests BAC J12 did not assemble fully because of coverage when compared to the overlapping BAC O24 that had pool coverage of 525 x per BAC (Table 3-3). BAC L18's pool coverage was marginally larger than the required > 450 x at 462 x. Both BACs L18 and J12 were much larger than 120 Kb at 151 Kb and 164 Kb respectively. Their combined MTP pool estimates were also larger than 480 Kb at 589 Kb and 574 Kb respectively (Table 3-3). BAC D6 overlapped with BAC L18. BAC D6's individual MTP size was also smaller, at 116 Kb and its pool was smaller than 480 Kb at 409 Kb MTP estimate (Table 3-3). Thus, 442 x coverage was sufficient for assembly of BAC D6's pool, although below 450 x estimate based on a BAC size of 120 Kb.

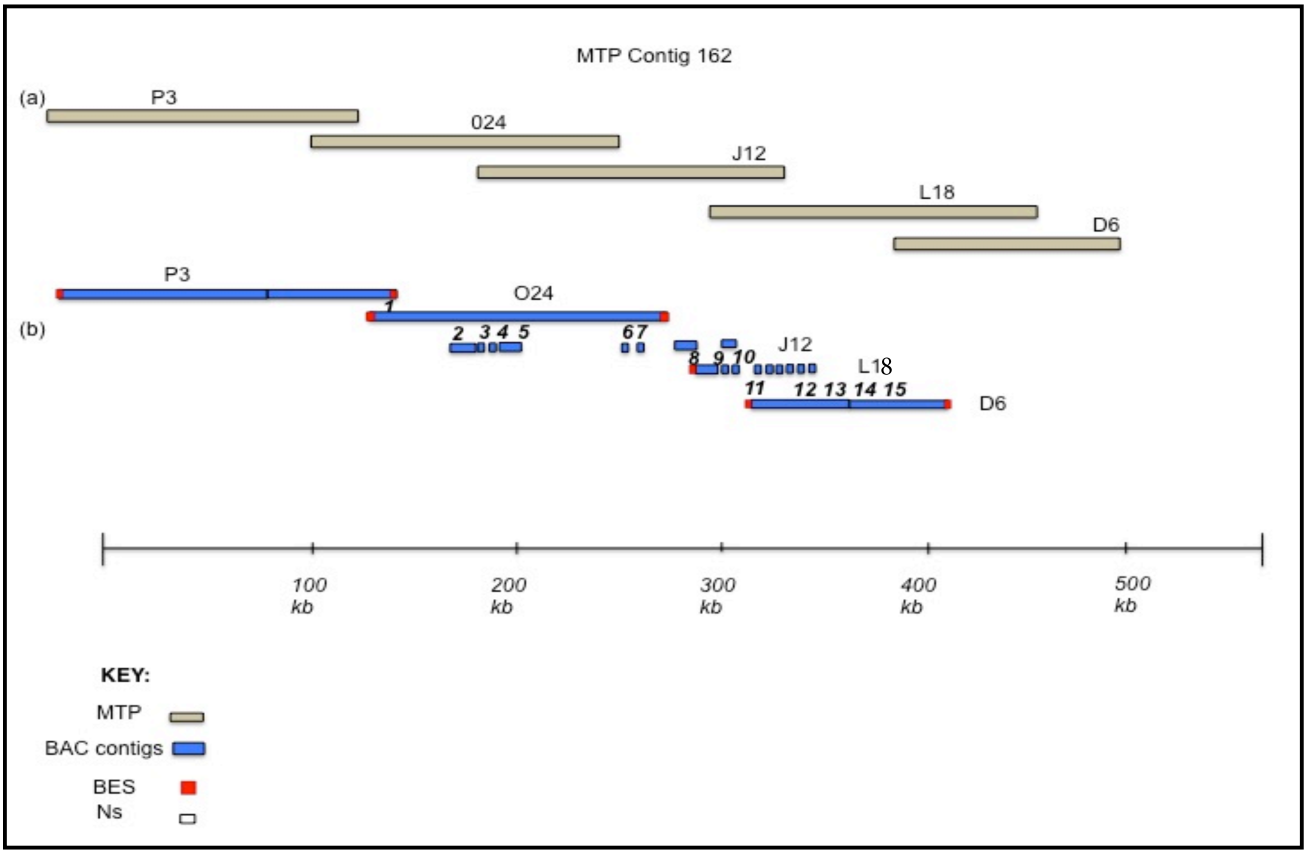

Figure 3-9:Overlaps from BACs constituting MTP contig 162. Image drawn to scale. Overlaps are denoted from 1-15.

Table 3-3: Pool assembly statistics of MTP 162 BACs.

BAC	Pool name	Cov per BAC	MTP - assembly size (Kb)	MTP-size (Kb)	MTP - assembly size (Kb)	% *E coli*
D6	7DS-MTP-P4-B6-D6-F6-H6	442	-27	409	-27	14
J12	Ta_7DS-12-J12-L12-N12-P12	398	-83	574	-83	5
L18	Ta_7DS-11-J18-L18-N18-P18	462	0	589	0	5
O24	Ta_7DS-7-I24-K24-M24-O24	525	-15	513	-15	13
P3	Ta_7DS-12-J3-L3-N3-P3	661	-112	581	-112	6

Seven overlapping BACs P22, G11, D3, L14, C14, P17 and E6 constituting MTP contig 307 were compared (Figure 3-10). Pools containing the seven BACs are listed in Table 3-4 . Table 3-4 shows the seven BACs pools with their corresponding coverage per BAC, MTP sizes, assembly sizes, difference between the MTP estimates and assembly sizes and the *E coli* percentage. BACs P22, G11, D3, L14, C14, P17 and E6 overlapped as expected (Figure 3-10 ***b***), based on the MTP (Figure 3-10 ***a***). Overlaps across BACs of MTP contig 307 were numbered ***1 - 8*** (Figure 3-10 ***b***) and had between 82% and 100 % sequence identity. The overlap between BAC P22 and G11 numbered ***1*** (Figure 3-10 ***b***) was 2,877 bp long and had 83% sequence identity. The overlap between BAC G11 and D3 numbered ***2*** (Figure 3-10 ***b***) had was 25,354 bp long and had 100% sequence identity. The overlap numbered ***3*** between BAC D3 and L14 (Figure 3-10 ***b***) was 54,278 bp long with 100% sequence identity. BAC L14 assembled with 1,021 Ns while BAC C14 had 2 contigs. This resulted in two overlaps numbered ***4*** and ***6*** with a 6,932 bp gap between BACs L14 and C14 (Figure 3-10 ***b***). The overlaps ***4*** and ***6*** (Figure 3-10 ***b***) was 9,880 bp and 15,123 bp long with 82% and 100% sequence identity respectively. The overlap between the two BACs L14 and C14 had a gap despite both BACs having > 450 x coverage (Table 3-4). The overlap numbered ***7*** between BAC C14 and P17 (Figure 3-10 ***b***) was 2,331 bp long with 100% sequence identity while the overlap numbered ***8*** between BACs P17 and E6 was 3,709 bp long with 100 % sequence identity (Figure 3-10 ***b***). The order of overlapping BACs was also confirmed based on individual BAC BES mappings (Figure 3-10 ***b***).

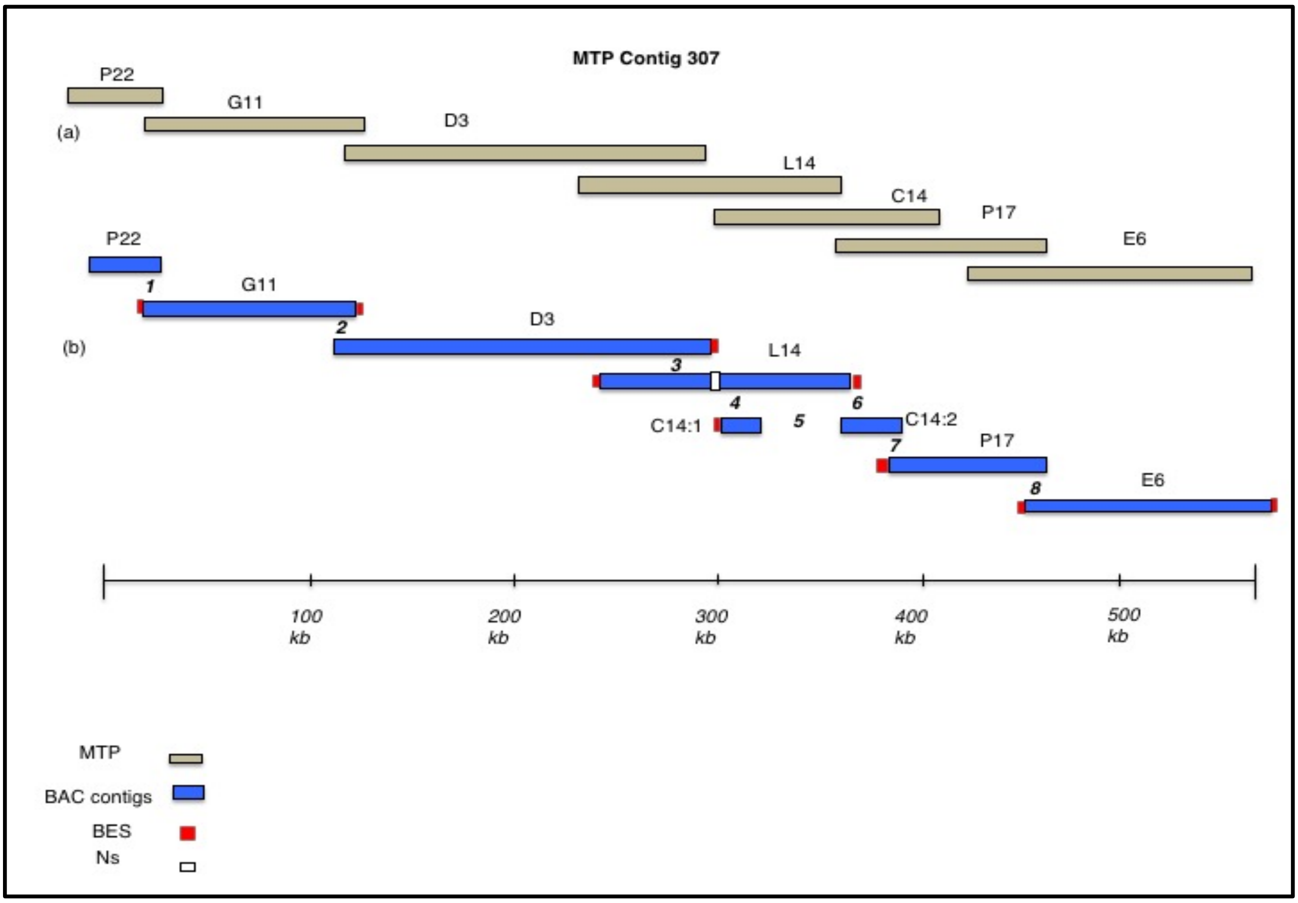

Figure 3-10: Overlaps from BACs constituting MTP contig 307. Image drawn to scale. Overlaps are denoted from 1-8

Table 3-4: Pool assembly statistics of MTP 307 BACs.

BAC	Pool name	Cov per BAC	MTP - assembly size (Kb)	Assembly size (Kb)	MTP- size (Kb)	% *E coli*
P22	7DS-2-J22-L22-N22-P22	718	25	415	440	4
G11	7DS-6-A11-C11-E11-G11	684	78	554	631	6
D3	7DS-12-B3-D3-F3-H3	488	-65	690	624	7
L14	7DS-6-J14-L14-N14-P14	575	-146	654	508	7
C14	7DS-6-A14-C14-E14-G14	535	19	477	497	7
P17	7DS-4-J17-L17-N17-P17	123	324	101	426	55
E6	7DS-2-A6-C6-E6-G6	610	128	421	541	5

MTP contig 3945 constituted seven overlapping BACs 022, E19, F12, B12, L4, I7 and E18 (Figure 3-11). Pools in which the seven BACs were sequenced are listed in Table 3-5 . This table shows the seven BACs pools with their corresponding coverage per BAC, MTP sizes, assembly sizes, difference between the MTP estimates and assembly sizes and the *E coli* percentage. Overlaps across BACs in MTP contig 3945 were numbered ***1* - *4*** (Figure 3-11 ***b***). Overlap numbered ***1*** between BAC 022 and E19 (Figure 3-11 ***b***) was 20,257 bp long with 100% sequence identity. The overlap between BAC E19 and F12 numbered ***2*** (Figure 3-11 b) was 1,765 bp long with 100% sequence identity. There was no overlap between BAC F12 and B12 although this overlap was expected based on the MTP (Figure 3-11 ***a***) and pools from which BACs F12 and B12 were sequenced had > 450 x sequence coverage per BAC (488 x and 575 x respectively) (Table 3-5). The overlap between BAC B12 and L4 was not observed as no contigs from BAC L4’s pool overlapped with BAC B12 or I7 (Figure 3-11 b, BAC L4 in green). BAC L4’s BES mapped to BAC B12 and I7 at locations where BAC L4 should have mapped. BAC L4’s pool had a sequence coverage per BAC of 253 x with an *E coli* % of 34. This suggests BAC L4 was absent in its pool assembly (Figure 3-11 ***b*** in green). However there was a small overlap between BAC B12 and I7 numbered ***3*** in (Figure 3-11 ***b***) which was 693 bp long with 100% sequence identity. BAC I7 overlapped with BAC E18 for 10,953 bp with 100% sequence identity.

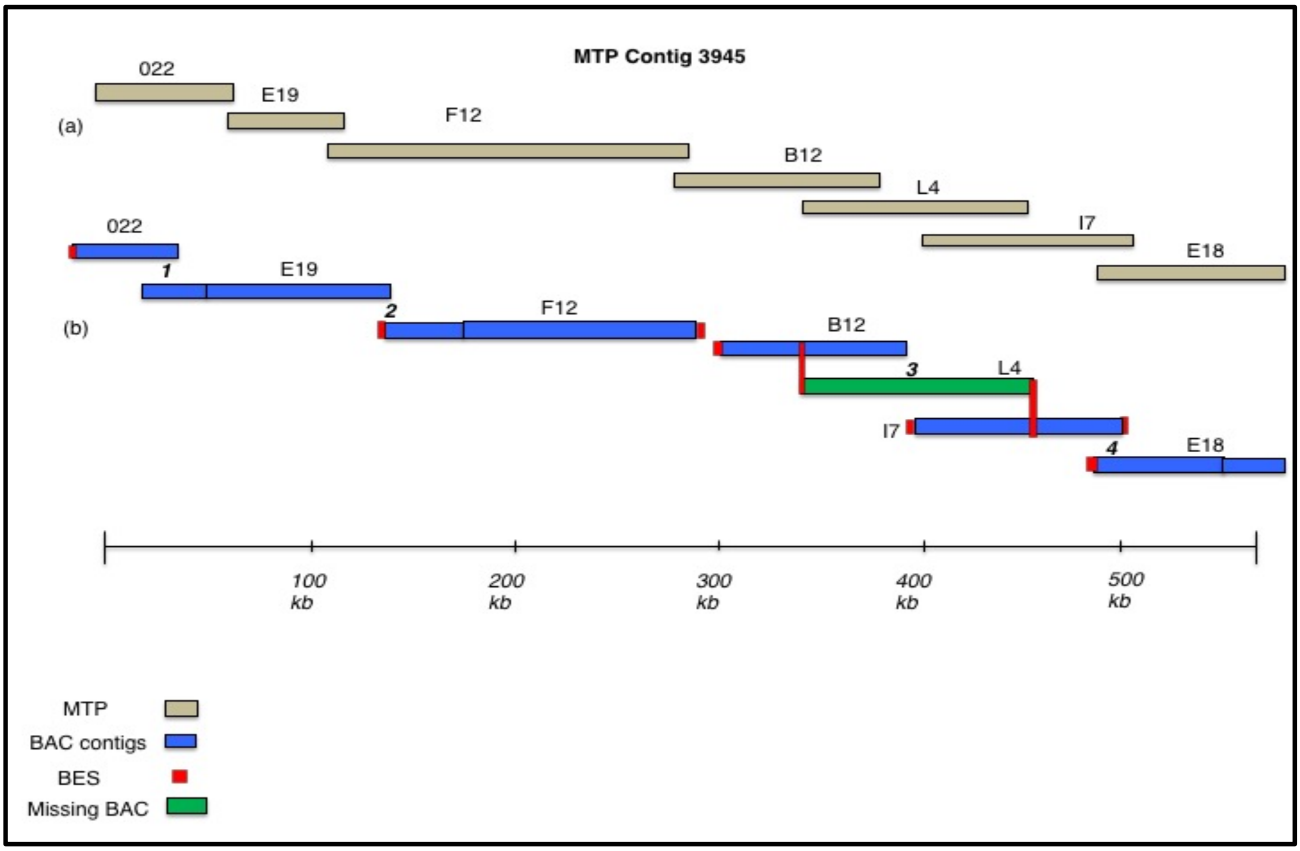

Figure 3-11: Overlaps from BACs constituting MTP contig 3945. Image drawn to scale. Overlaps are denoted from 1-4. BAC L4 (in green) was missing but its BES mapped to its expected position.

Table 3-5: Pool assembly statistics of MTP 3945 BACs.

BAC	Pool name	Cov per BAC	MTP - assembly size (Kb)	Assembly size (Kb)	MTP size (Kb)	% *E coli*
O22	7DS-9-I22-K22-M22-O22	601	-84	524	440	4
E19	7DS-9-A19-C19-E19-G19	684	78	554	631	6
F12	7DS-12-B12-D12-F12-H12	488	-65	690	624	7
B12	7DS-MTP-P5-B12-D12-F12-H12	575	-146	654	508	7
L4	7DS-MTP-5-J4-L4-N4-P4	253	182	265	448	34
I7	7DS-2-I7-K7-M7-O7	746	76	439	515	5
E18	7DS-2-A18-C18-E18-G18	686	15	441	457	6

Anchoring all BAC assemblies to the MTP showed that out of 4,680 BACs in a MTP consisting of 931 scaffolds (Table 3-6), 2,197 BACs had overlapping contigs. The average number of BACs in a MTP contig was 4.9, and the average number of overlapping BACs in a MTP contig was 2.3.

Table 3-6: MTP BAC overlaps.

No of MTP contigs	931
Total No of BACs	4,608
Total No of overlapping BACs	2,197
Avg. No of BACs/MTP contig	4.9
Avg. No of overlapping BACS/MTP contig	2.3

3.2.3 Assembly accuracy assessment of 7DS BACs using the optical map

The 7DS chromosome arm optical map generated by BioNano genomics (http://www.bionanogenomics.com) for Jaroslav Dolezel using *BspQ1* restriction sites had the following statistics. 75 Gb of sequence from 262,357 chromosome 7DS molecules (long elongated DNA molecules) analysed with the Irys instrument resulted in ~ 200 x depth of coverage of the 7DS chromosome arm. The molecules had an average length of 287 Kb and median length of 239 Kb (Table 3-7). No molecules were less than 80 Kb in length and the maximum molecule length was 2.4 Mb (Figure 3-12). The average density of *BspQ1* restriction sites that were identified and scanned was 10 per 100 Kb (Table 3-7).

Table 3-7: Molecule size statistics.

Statistic	
Total DNA (Gb)	75
Total molecules	262,357
Average length (Kb)	287
Median length (Kb)	239
Avg. restriction site density (per 100 Kb)	10

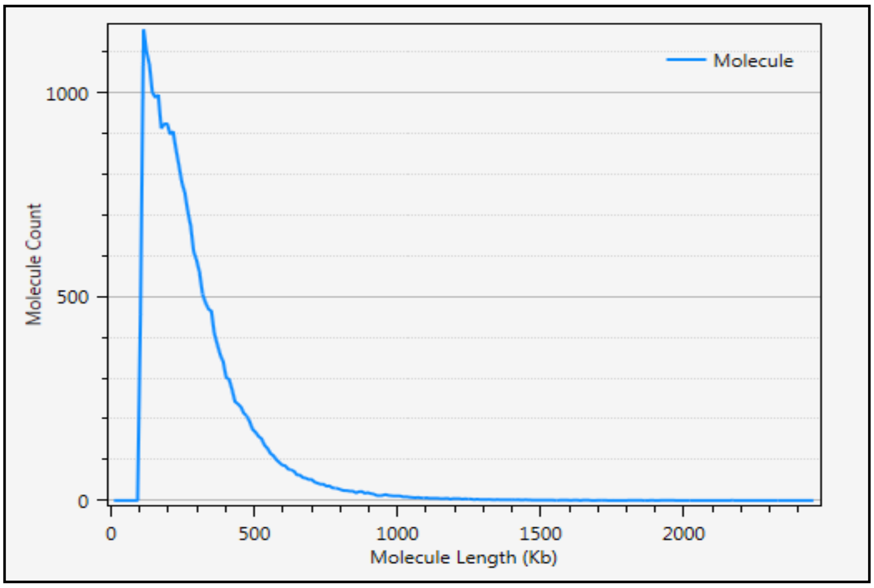

Figure 3-12: Molecules size distribution.

Using the IrysView software with default parameters, the distribution of repeat units (repeating series of two or more equally spaced restriction sites) showed from 262,357 chromosome 7DS molecules, 11% of the molecules had repeat units (Table 3-8). The total number of repeat units in all 262,357 molecules was 35,209 and the average percentage of repeat units per molecule was 13%. Repeat units contained 1030 Mb (1.3%) of all bases in molecules (Table 3-8).

Table 3-8: Repeat unit statistics.

Total No of molecules	262,357
% Molecules with repeat units	11.5
Total No of repeat units	35,209
Avg. % repeat units/molecule	13.4
Total bases (Gb)	75
Total bases in repeat units (Mb)	1030
% Bases in repeat units	1.3

From the IrysView interface, the average, median, maximum and minimum repeat unit size was 847 Kb, 7.3 Kb, 266,422 Kb and 2 Kb respectively. The Actual repeat unit sizes and counts were extracted from the IrysView software and plotted (Figure 3-13).

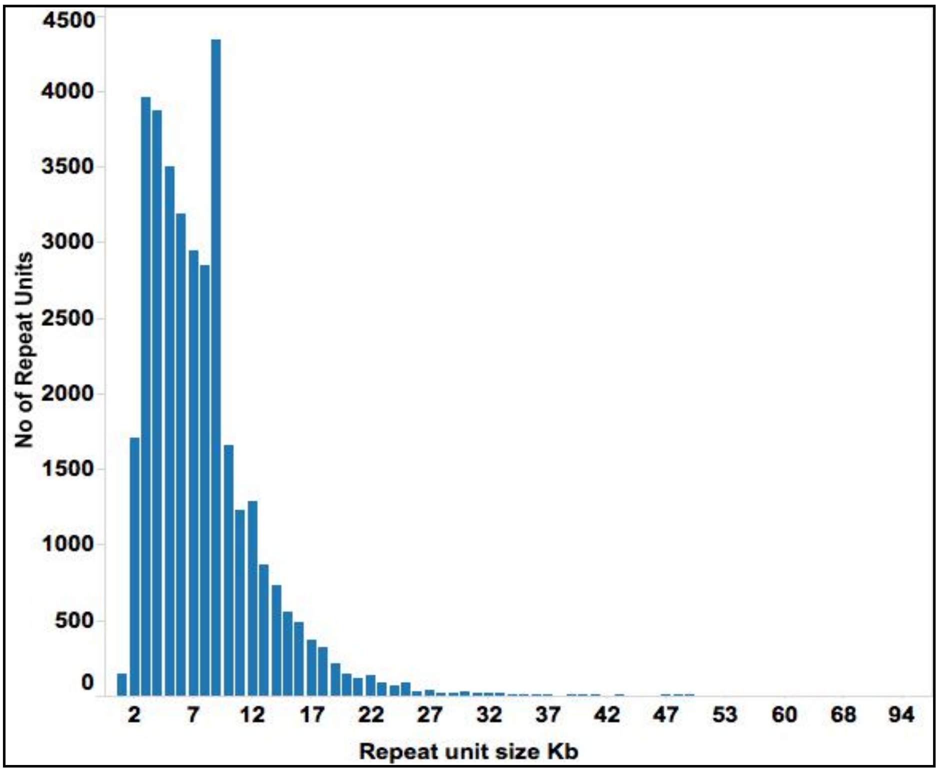

Figure 3-13: Counts of Repeat unit sizes.

Using the IrysView software interface, chromosome 7DS molecules and consensus map statistics resulting from the optical map assembly by BioNano showed assembly of *BspQ1* restriction site profiles in 262,357 molecules resulted in 371 consensus maps (cmaps) with an N50 of 1.29 Mb. The total length of the cmaps was 349 Mb, capturing 92% of bread wheat chromosome arm 7DS. The average cmap length was 0.9 MB (Table 3-9).

Table 3-9: Consensus map statistics.

Total consensus maps	371
Consensus map N50 Mb	1.3
Total length of consensus maps Mb	349
Avg length of consensus maps Mb	0.9

An evaluation of BACS constituting MTP contig 307 revealed the restriction profiles of the optical map matched with those of the assemblies (Figure 3-14). The order of six out of seven BACS constituting MTP contig 307 (P22, G11, D3, L14, C14, P17 and E16) (Figure 3-14 **a**) that were anchored to the the optical map (Figure 3-14 **b & c**) agreed with the MTP order. BAC P22 did not anchor to the optical map as it had only one *BspQ1* restriction site. BAC P17 had a restriction site missing compared to the optical map (Figure 3-14 **e**). This may have been as a result of an extra restriction site on the optical map or a collapse on the corresponding BAC P17 assembly. BACs G11 and C14 had extra *BspQ1* restriction sites compared to the optical map (Figure 3-14 **blue arrows**). Extra restriction sites on BAC assemblies did not result in a distortion of the overall restriction block profile.

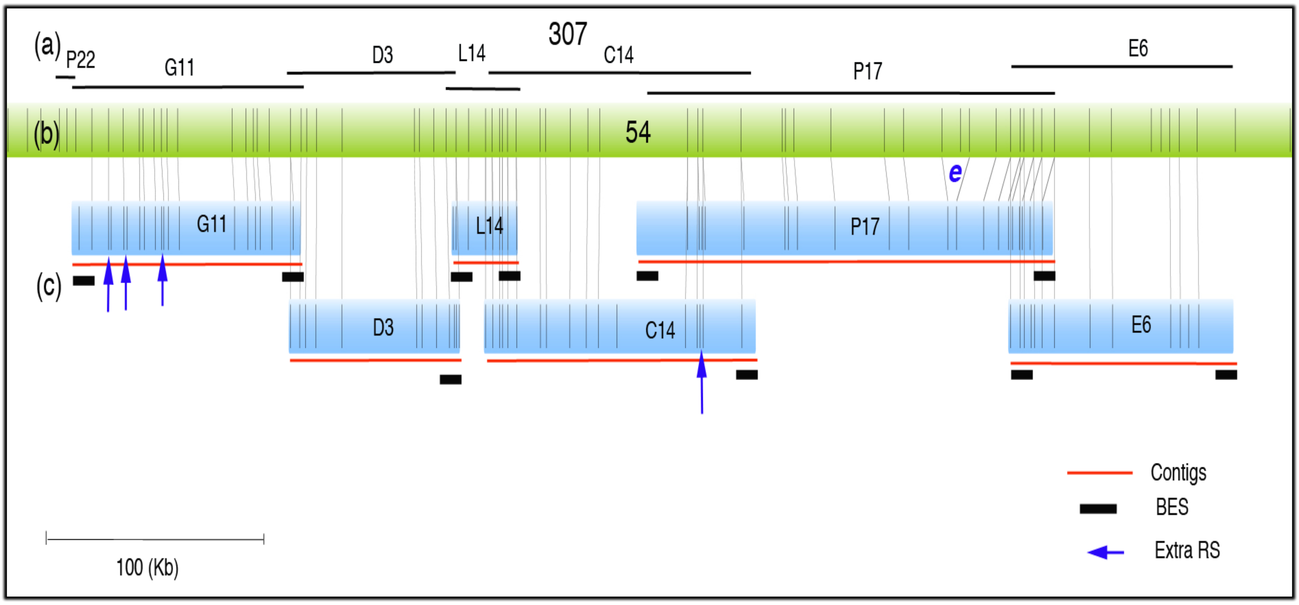

Figure 3-14: BACs constituting MTP contig 307 all anchored in the correct order with matching restriction profiles to optical map contig 54. Image from the IrysView software with annotations added for clarity.

An evaluation of BACs constituting the MTP contig 190 using the optical map revealed the restriction profile of the BAC assemblies and the order and overlaps were in agreement with those of the optical map (Figure 3-15). All contigs from BACs E9, J12, D1, I22, P20 and G14 (Figure 3-15 ***c***) constituting MTP contig 190 (Figure 3-15 ***a***) were anchored to the same optical map, cmap contig 28 (Figure 3-15 ***b***). The *BspQ1* restriction sites (vertical lines in Figure 3-15) on the optical cmap and the BAC contigs were in agreement for accurately assembled BACs. The overlap between BAC I22 and P20 that showed extended restriction sites to the right on BAC P20 (Figure 3-15) , indicating a possible mis-assembly of BAC P20.

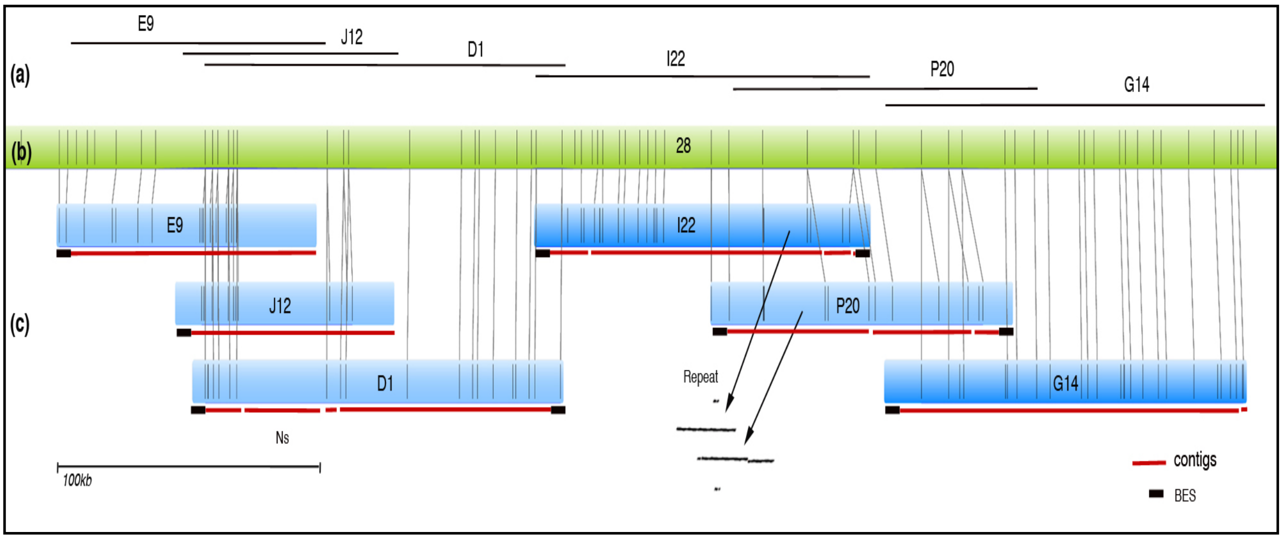

Figure 3-15: BACs constituting MTP contig 190 all anchored in the correct order with matching restriction profiles to optical map contig 28. Image from the IrysView software with annotations added for clarity.

The possible mis-assembly at overlaps between BAC I22 and P20 was supported by earlier results in 3.2.1 that showed there was a fragmented overlap between BAC I22 and P20. A closer inspection of the overlap between BACs I22 and P20 showed that the overlap spanned a repeat resulting in 32,113 bp overlap region with 99.84% (32,062/32,113 bp) sequence identity, 35 base pair mismatches and 16 indels (5 on BAC I22 and 11 on BAC P20). The assembly of BAC P20 resulted in the placement of 145 ambiguous nucleotides (Ns) in this region. A search in the Triticeae repeats database (TREP) (http://wheat.pw.usda.gov/ITMI/Repeats/) of 6,771 bp region preceding the Ns with sequence from BAC I22 revealed 90% identity (2114/2347) bp with 11 gaps to the LTR retrotrasposon RLG_Sabrina_C_AY494981-4. Thus, the placement of Ns by the assembler due to lack of sequence coverage over estimated the indel size on BAC P20. The over estimated size resulting in extended restriction site motifs in the region (Figure 3-15). Coincidentally, BAC P20's pool had relatively lower sequence coverage at 394 x per BAC compared to BAC I22's pool which had 592 (x) coverage per BAC (Table 3-2).

Assembled BACs often showed extra BspQ1 restriction sites compared to the optical map as shown on BACs E9, J12, D1, I22 and G14 (Figure 3-15). Further analysis of the occurrence of extra restriction sites was done by assembling individually sequenced (un pooled) BACs. Five overlapping BACs (E01, E09, E19, H04 and N11) were thus assembled and anchored to the optical map. Assemblies had a mean N50 60 Kb, mean assembly length of 149 Kb and mean contig counts of 6.6 (Table 3-10).

Table 3-10: Assembly statistics of five overlapping BACs individually sequenced and assembled.

BAC	Assembly Length Kb	Coverage	Longest contig Kb	N50 Kb	No of Contigs
E01	111	10,037	81	81	2
E09	194	9,126	86	27	16
E19	143	7,528	68	46	5
H04	134	8,409	62	60	6
N11	167	9,778	89	89	4

All contigs of BACs (E01, E09, E19, H04 and N11) anchored to the same cmap on the optical map, cmap contig 280 (Figure 3-16). This was consistent with the MTP (Figure 3-16 ***a***). Anchored contigs (Figure 3-16 ***c***) had several BspQ1 restriction sites identical with the optical map (Figure 3-16 ***b***). Contig 2 of BAC H04 (H04-2) overlapped with contig 9 of BAC E09 (E09-9) at 26,294 bp with 100% sequence. Contig H04-2 also overlapped with contig E09-8 at 32,723 bp with 99.9% sequence identity, with one mismatch. The overlap between H04 and EO9 totalled 59,017 bp, suggesting an under estimate in the MTP estimated overlap of 31,240 bp. Contig E09-8 overlapped with contig N11-24 at 45,143 bp with 100% sequence identity, 0 mismatches consistent with the MTP overlap estimate of 42,600 bp. Contig E01-27 overlapped with contig E19-29 at 44,388 bp with 99.99% sequence identity, 1 mismatch, consistent with MTP overlap estimate of 51,120 bp.

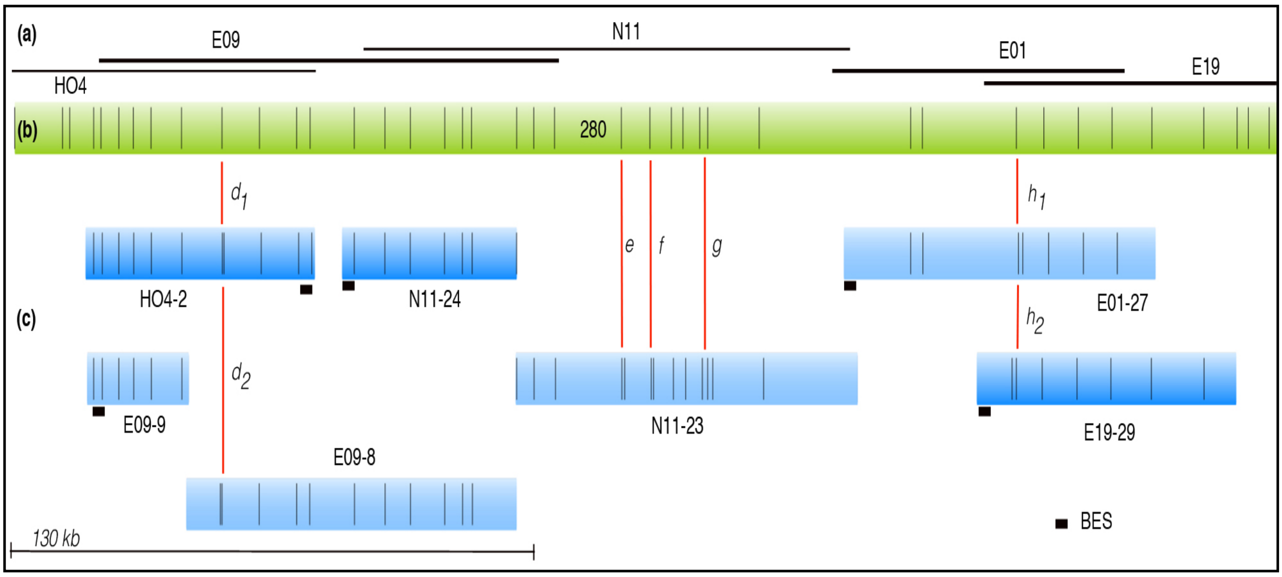

Figure 3-16: Five overlapping BACs, individually sequenced and anchored to optical map. Image from the IrysView software with annotations added for clarity

Several *BspQ1* restriction sites were observed to be present on contigs (E09-8 and H04-2), (E01-27 and E19-29) and contig N11-23, but absent on the optical map (Figure 3-16 ***c***). This is as a result of limited optical resolution of the map (Hastie *et al.*, 2013; Lam *et al.*, 2012), currently estimated at restriction sites being at least 1.5 kb apart. Overlap regions on the BAC contigs showed 99.9 % - 100% sequence identity with distances between restriction sites on the same contigs being equidistant. Positions with more restriction sites on contigs compared to the optical map. d_1 , d_2 , e, f, g, h_1 and h_2 (Figure 3-16 ***c***) were 422 bp, 422 bp, 730 bp, 608 bp, 1392 bp, 1319 bp and 1125 bp apart respectively on the BAC contigs nucleotide sequence. These confirmed the optical map resolution threshold of 1.5 Kb (Hastie *et al.*, 2013; Lam *et al.*, 2012). The IrysView software assembles restriction profiles on several molecules into a consensus map (cmap). Using the IrysView software, BAC H04-2 (Figure 3-16 ***c***) was mapped to the corresponding molecules that were assembled to generate cmap 280 Figure 3-17).

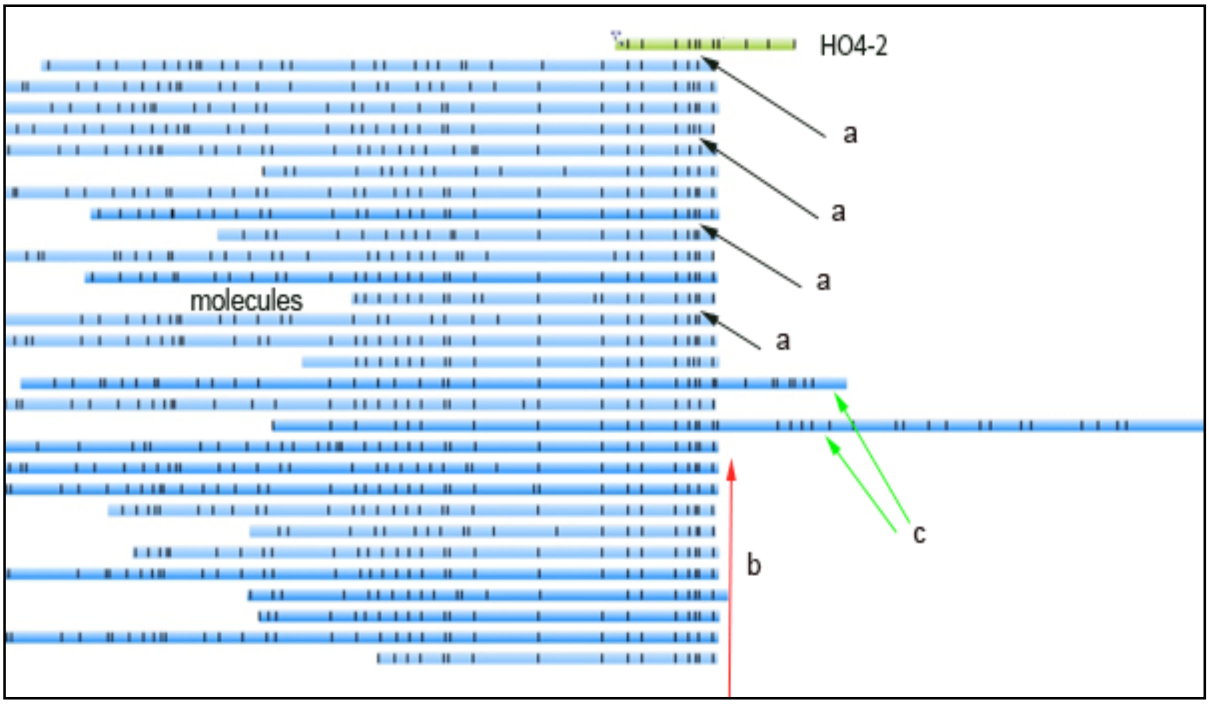

Figure 3-17: BAC H04 contig 2 mapped to optical map molecules.

Several single molecules that were assembled into cmap 280 further confirmed the extra restriction sites on BAC HO4-2 Figure 3-17 ***a***). This suggests that during assembly of the consensus map, restriction sites that are close to each other are discarded, even though these were captured as distinct restriction sites during the optical imaging stage. Furthermore, most molecules in the BAC HO4-2 region appear to have had a strand break Figure 3-17 ***b***), while other strands had no break Figure 3-17 ***c***). This suggests close restriction sites may result in strand breaks during partial digestion with the *BspQ1* restriction enzyme.

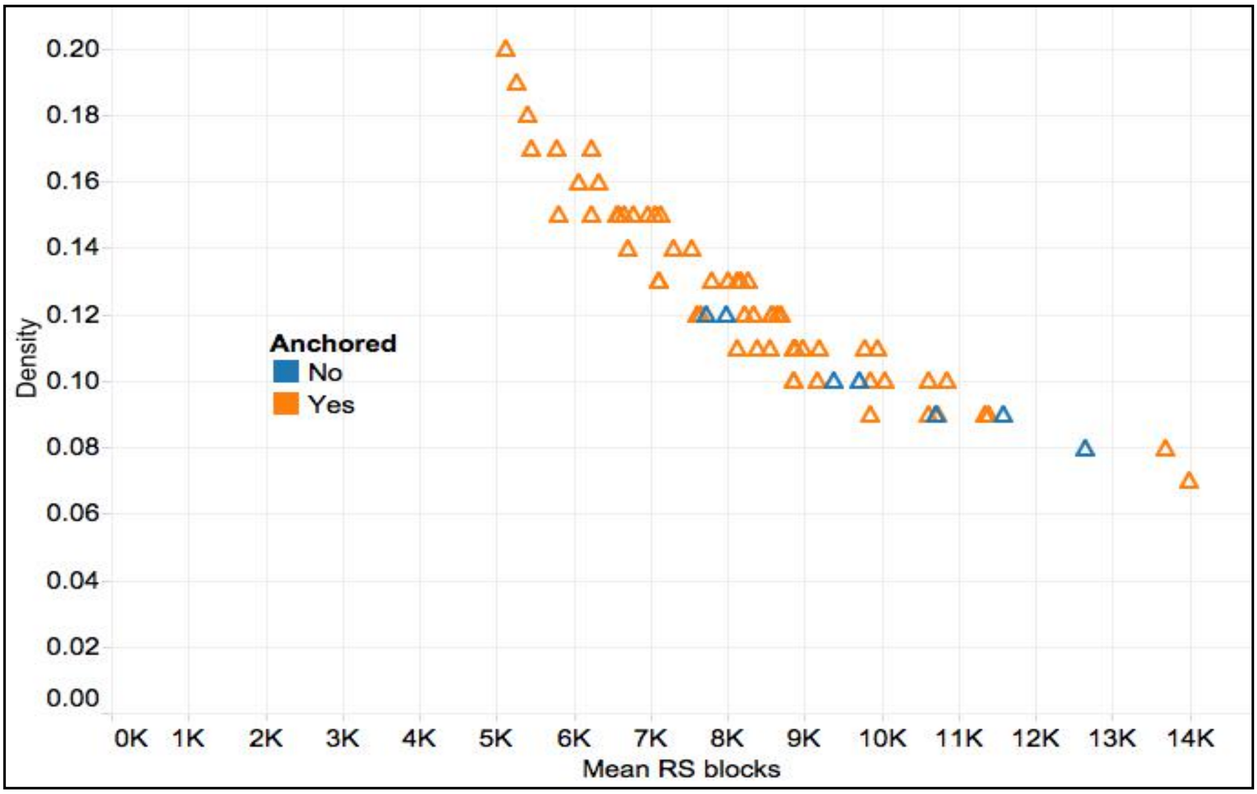

Figure 3-18: Restriction site density.

Using 65 fully assembled BACs that each assembled into a single contig and were confirmed by having BES correctly paired and mapped, only 58 BACs out of the 65 BACs could be anchored to the optical map (Figure 3-18). An evaluation of the density of *BspQ1* restriction sites and the sizes of the restriction blocks (distance between two adjacent restriction sites) on the anchored and un-anchored clones suggests that contigs with a high restriction site density and a lower average restriction block size are more likely to be anchored to the optical map. Un-anchored BACs had significantly larger average restriction blocks sizes of 9,964 bp (7,731 bp - 12,643 bp) compared to anchored clones that had an average restriction block size of 8,252 bp (5,133 bp - 13,997 bp) p value = 0.02821. All BACs with > 0.12 restriction sites per Kb (12 per 100 Kb) were anchored.

3.2.4 Anchoring of 7DS BAC assemblies to the optical map

Assemblies from un-scaffolded pools, mate pair scaffolded pools and super scaffolds were anchored to the optical map. Assembled contigs, mate pair scaffolded pools and the super scaffold sequences were evaluated to determine the percentage of sequences that could be anchored. Scaffolding and pseudo molecule generation increased the overall percentage of sequences that could be anchored (Table 3-11) as shown by the increase of anchored sequences on the optical map consensus map 1 prior to scaffolding (Figure 3-19) and post mate pair scaffolding (Figure 3-20).

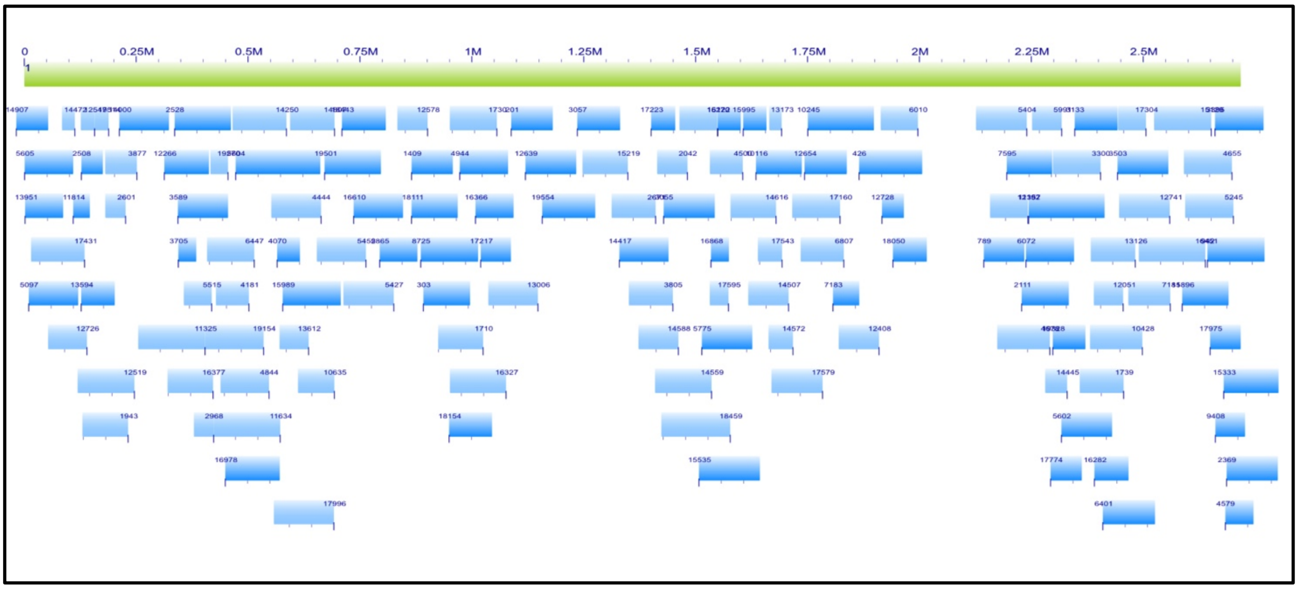

Figure 3-19: Pre-scaffolded assemblies (blue) anchored to the 7DS optical map's cmap 1 (green).

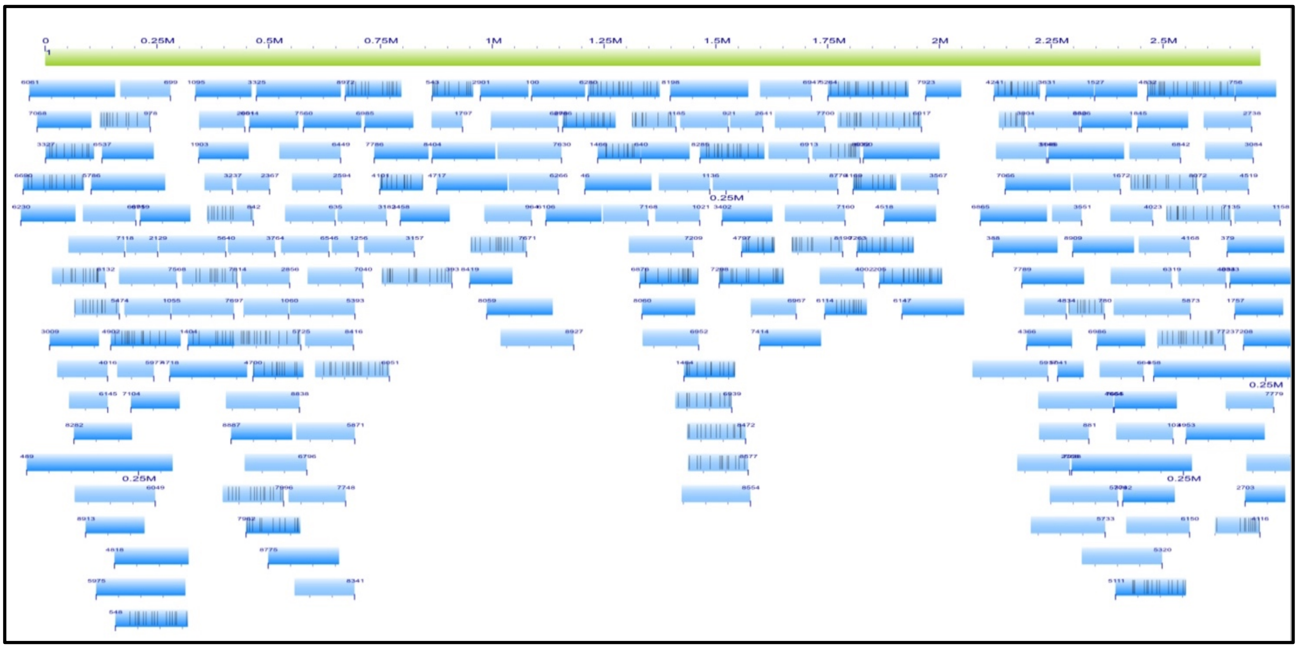

Figure 3-20: MP scaffolded assemblies (blue) anchored on the 7DS optical map's cmap 1 (green).

Table 3-11: Anchoring statistics of 7DS assemblies to optical map.

Query	**No of sequences**	**Average sequence size Kb**	**No sequences anchored**	**% Sequences Anchored**	**No of cmaps**	**Average size of RS blocks Kb**
Contigs	19,569	24	3,528	18.0	365	45.90
Mate pair Scaffolds	8,993	54	3,725	41.4	369	49.60
Super scaffolds	5,528	117	2,710	49.0	371	50.10

All the 7DS assembled BAC pools had 19,569 contigs, of which 3,528 (18%) could be anchored to 365 consensus maps of the optical map (Table 3-11). The average size of anchored restriction blocks for contigs was 45.9 Kb (Table 3-11). The 7DS mate pair scaffolded pools had 8,993 scaffolds, of which 3,725 (41%) were anchored to 369 consensus maps. The average size of anchored restriction blocks for scaffolds was 49.6 Kb (Table 3-11). The average size of sequences increased due to scaffolding. Un-scaffolded contigs, mate pair scaffolded scaffolds and super scaffolds had average lengths of 24 Kb, to 54 Kb and 117 Kb respectively (Table 3-11).

3.3 Discussion

For ease of reference, the discussion presented in this section follows the same order as the results sections 3.2.1 to 3.2.3.

3.3.1 Assembly contiguity of 7DS BAC assemblies using mate pair libraries

MP mapping orientation and insert size distributions offer powerful tools for estimating the integrity and contiguity of genome assemblies, especially when evaluating complex, highly repetitive genomes. BLAST (McGinnis and Madden, 2004) was used for mapping of the MP reads onto assemblies. Other short read mappers such as SOAP2, BWA and Bowtie were unsuitable for this analysis as they only allow for a fixed minimum number of mismatches and indels when mapping reads (H., Li and Homer, 2010). Allowing for mismatches or indels is suitable for variant discovery, but not for the identification of the exact locus of a read where hits with zero mismatches and no indels are appropriate when mapping reads sequenced from the same genome as the reference. Furthermore, the identification of a read's locus based on the length of the match vs. the length of the query cannot be easily implemented using provided parameters of current short read mappers.

Mapping of reads as pairs, though invaluable in structural variation analysis, has been shown to reduce accuracy in mapping when compared to mapping reads pairs separately (H., Li and Homer, 2010).

We combined the expected percentages of mapping orientations of MP reads in a Nextera MP sequenced dataset and insert size distributions to assess the accuracy of our *de novo* assembly and scaffolding approach. Assemblies of 96 pools from plate one had the expected insert size distributions and mapping orientations of MP datasets. There was an increase in median insert sizes across the three references (*E coli*, plate one contigs and plate one scaffolds) despite all three having the same shadow library mapping percentages of 1%. *E coli* had a median insert size of 864 bp, plate one contigs 1,860 bp and plate one scaffolds 83,786 bp. The increase in median insert size values is attributable to the increased counts of mapped read pairs due to a larger mapping area. *E coli* had 4,568, plate one contigs had 39,172 and plate one scaffolds had 101,017 read pairs mapped. This is because the rise in mappings counts resulted in an increase in the median insert size value and not the overall percentage of the shadow library. This suggests that an increase in mapping counts increase the deviation of insert sizes but not overall percentages. The increase in mapping counts was expected given there was a corresponding increase in size and contiguity across the three references. The *E coli* reference was 4 Mb compared to plate one contigs which were 35 Mb in total. Although the total size of plate one contigs remained the same compared to plate one scaffolds at 35 Mb, scaffolding did increase the contiguity of plate one scaffolds enabling more reads to map in pairs on single contigs.

The scaffolding approach resulted in an increase in the number of MP mapping identity from 3,542,275 to 3,881,779 mate pairs. This was an increase of 339,504 additional mate pairs, a percentage increase of 9.5%. An increase in the number of MP mapping shows improved scaffolding and contig joining as no sequence information from the MP dataset was incorporated into the assembly during scaffolding. The SASSY assembler did not merge contigs with short overlaps probably as a result of low sequence complexity. These were correctly merged during scaffolding resulting in an increase in the number of MP reads mapping. The increase (339, 504 mate pairs) however consisted of 1.25% reads mapping in a chimeric orientation, 18% reads mapping with a shadow paired end orientation and 80% of the reads mapping in the correct mate pair orientation. This is expected given highly repetitive content of wheat. For such regions, read pairs would have multiple mapping orientations.

The sequencing of MP reads involved sequencing of pools of BAC pools. Scaffolding involved the selection of MP reads that both mapped to the same pool and not across pools. Although MP reads mapping to regions of overlapping BACs were excluded (Figure 3-6), such regions were later merged through the evaluation of the MTP conformity and thus no contiguity was lost. The significance of long insert MP reads in scaffolding across repeat regions was also evident in (Figure 3-7) and repeat regions that could not be spanned by PE reads were spanned by MP reads. Plots of paired reads insert sizes, orientation and position on assemblies enables comprehensive and accurate identification of mis-assemblies.

3.3.2 Minimum tilling path (MTP) conformity of 7DS BAC assemblies

BAC assemblies conformed to the MTP with overlaps between ordered MTP BACs at 100% sequence identity where coverage was sufficient in pools from which the overlapping BACs were sequenced. Where either or both BACs had low coverage, as indicated in their estimated coverage in pools, their assemblies were fragmented. Where fragmented contigs were available, sequence identity was lower suggesting sequence complexity and low coverage confounded assembly. The estimated sequencing depth of > 450 x offered a rough estimate for the optimal sequencing depth to aim for. For example, in MTP contig 190, BACs J12 and P20's pools in (Figure 3-8) had sufficient coverage because they were relatively smaller although the sequence coverage was < 450 x (Table 3-2). A better estimation of coverage would have been to consider each BACs MTP size estimate. Although this would be more accurate, the sizes of individual BACs vary and setting a target for each BAC prior to sequencing would be impractical.

For MTP contig 162 (Figure 3-9) overlapping BACs L18 and J12 had fragmented assemblies which overlapped with adjacent BACs. Given that both BACs L18 and J12's pools had sequence coverage per BAC of 398 x and 462 x respectively while the overlapping BACs to BACs L18 and J12 (BACs D6 and 024) had 442 x and 525 x coverage per BAC respectively (Table 3-3), suggest sequence coverage and complexity resulted in fragmented assemblies of the genomic region covered by BACs L18 and J12.

MTP contig 3945 (Figure 3-11 ***b***) had no overlap between BACs F12 and B12 although this overlap was predicted to be present in the MTP. Pools from which BACs F12 and B12 were sequenced had 488 x and 575 x sequence coverage per BAC respectively, which was above our minimum coverage estimate of 450 x coverage per BAC. BAC F12's BES did not map to any of the contigs from BAC B12's pool and vice versa. This suggests that the overlap between BACs F12 and B12 was absent and the wrong BAC was picked into the pool. MTP physical maps have been shown to have errors in which predicted overlaps were not present (Shearer *et al.*, 2014). MTP contig 3945 also had BAC L4 having no overlaps (Figure 3-11 ***b,*** BAC L4 in green). But unlike BACs F12 and B12, BAC L4's BES mapped to the adjacent overlapping BACs (B12 and I7). BAC L4's pool had a sequence coverage per BAC of 253 x with an *E coli* percentage of 34. This suggest that BAC L4 was correctly placed on the MTP but was absent in BAC L4's pool due to the high *E coli* content. A high *E coli* percentage is indicative that a BAC clone may have lost its clonal insert resulting in a higher proportion of *E coli* reads.

The wheat genome is estimated to contain approximately 80% repetitive DNA sequences (Wicker *et al.*, 2011) which mostly comprise of transposable elements (TEs) (Gulick *et al.*, 2009; Choulet *et al.*, 2010). TEs confound accurate sequencing, assembly, gene assignment and read mapping. Repeats have varied lengths and may span several fragments. Sequenced reads from such fragments are identical and thus pose a challenge when reconstructing the true size of the repeat during assembly. Repeats can also be shorter than read lengths and occur in large numbers. This further confounds assembly. Repeats also form nested complex structures resulting from insertion of repeats into pre existing repeats. Simple tandem repeats would pose the greatest difficulty during assembly and would result in fragmented assemblies even with high sequence coverage.

Assembly contiguity and accuracy depends in part on sufficient sequence coverage of the BAC to be assembled. Using the MTP overlaps, it is evident where coverage was sufficient in one BAC and lower in the adjacent BAC, the same genomic region was assembled correctly in the assembly with sufficient coverage. As a result, when rebuilding the MTP using BAC overlaps, the effect of the poorly assembled BACs was minimized by adjacent BACs. An evaluation of the overall conformity of the BAC assemblies to the MTP showed approximately 50% of the BACs had contigs with overlaps. Sequence coverage may have reduced the number of overlaps across BACs as 37% of all the pools had sequence coverage below 450 x (see section 2.2.5).

3.3.3 Anchoring of 7DS BAC assemblies to optical map

The BioNano genomics optical map enabled the identification of mis-assembled regions. This is a formidable tool for genome finishing, especially for complex genome assemblies. The resolution of complex nested repeat regions in an assembly have been a challenge that current sequencing platforms and assembly algorithms have not been able to satisfactorily tackle. The use of optical maps offers long-range information for contig placements coupled with low-density markers. The BAC-by-BAC assemblies agreed with the restriction profiles of the optical map. This validated the structure of the contigs at the restriction profile level. The optical map construction entails optical imaging of long stretches of DNA molecules. The molecules are partially nicked and labelled with a fluorescent tag, and elongated in Nanochannels for imaging. Restriction site density analysis results suggest that for optimal anchoring of assemblies to the optical map, a high density of restriction sites (> 12 sites per 100 Kb) is better. For example BAC P22 from MTP contig 307 could not be anchored to the optical map (Figure 3-14), although it had overlaps with an anchored BAC G11 which was anchored. BAC P22 was not anchored as it had a single *BspQ1* restriction site. The presence or absence restriction sites on assemblies compared to the optical map cannot accurately determined as to whether the error occurs on the optical map or is a mis-assembly especially where no overlaps are shared across BACs. This can be resolved by an increase in the density of restriction sites through the use of multiple restriction enzymes.

The depth of coverage from molecules analysed by the Irys instrument (~200x) and the length of molecules (average 287 Kb) are sufficient for the anchoring of BAC clones, which have an average size or 130 Kb. The median repeat unit size (7.3 Kb) compared to the average BAC clone size of 130 Kb suggests that the effects of equally spaced repeating units on anchoring BACs to the optical map would be minimal. It would be interesting to determine the effect of such repeating units on the assembly of restriction profiles into consensus maps and the possible resulting errors.

Anchoring all 7DS assemblies to the optical map suggest that longer contigs are more likely to be anchored. Of the un-scaffolded sequences, 18 % with an average size of 24 Kb could be anchored to 365 consensus maps, while 41 % of the mate pair scaffolded sequences with an average size of 49 Kb, twice that of un-scaffolded sequences could be anchored to 369 consensus maps. Anchoring of mate pair scaffolded sequences compared to un-scaffolded sequences increased by 23% and 5 more consensus maps had sequences anchored to them. The super scaffold sequences had a higher average size (117 Kb) and 49% anchored to all 371 consensus maps. The length of a sequence and the number of restriction sites affect anchoring to the optical map. While sequence contiguity of assemblies is currently limited by read lengths produced by sequencing platforms, increasing the density of restriction sites when generating optical maps could improve anchoring. A higher density of restriction sites posses a challenge in that an increased density of restriction sites may result in more frequent molecule strand breaks where two restriction sites are close to each other. The use of two restriction enzymes with two labels has been demonstrated (Lam *et al.*, 2012) where the same sample was labelled with different enzymes independently at lower restriction site densities. Probable molecule strand breaks are reduced while increasing the resolution of restriction sites. Redundancy of restriction sites is also reduced, as the additional enzyme will have a different restriction profile. This may improve the resolution of the map in that, where restriction sites less than 1.5 Kb apart confound the resolution and accurate assembly of molecules into a consensus map; the extra resolution provided by a second enzyme would mitigate this. The possible effect of repeating units on assembled consensus maps will also be reduced.

The applicability of the optical map for high throughput assembly correction and scaffolding is limited due to the graphical user interface provided by the IrysView software. It is envisioned that as development continues, more functionality will become available, enabling the incorporation of sequence information directly into the optical map, and high throughput assembly scaffolding and correction. The scaffolding approach had the greatest impact in the number of sequences that could be anchored to the optical map from 18 % prior to scaffolding to 49 %. The increase in sequence contiguity resulted in more restriction sites occurring on a contig and thus enabling the anchoring of that contig onto the optical map.

3.4 Summary

The use of datasets from multiple independent platforms using the same genomic DNA source offers a powerful approach for the *de novo* validation, optimization and improvement of genome assemblies. For chromosome arm 7DS, these multiple independent platforms included the use of long insert mate pair libraries (MP), minimum tilling path (MTP) of BAC clones and an optical map all from isolated chromosome arm 7DS genomic DNA. All the above platforms did validate the paired end assemblies. Although the assembly of complex genomes still poses a challenge, the assembly approach presented enables the identification of assembly, MTP and BAC pooling errors resulting in better genome assemblies. With rapid advancements in sequencing technologies and assembly algorithms, the sequencing and assembly of complex genomes will no longer be such a challenge.

4 Genome annotation of bread wheat chromosome arm 7DS

4.1 Introduction

Comprehensive genome annotation is a challenging task especially for *de novo* sequenced genomes where previous existing gene models from model organisms are absent. The absence of gene models or model organisms makes training of annotation tools for gene finding difficult. This is further confounded by fragmented and incomplete assemblies that result from assembling short reads that were characteristic of current NGS sequencing technologies. As more genomes are sequenced and annotated, genomic database repositories contain more annotations and as such, a previously annotated genome may need to be re-annotated based on the availability of newer gene models and or the availability of a complete or better genome assembly. Due to the wide availability of gene annotation tools, the evaluation and comparison of various annotation methods is difficult as there are no agreed standards on genome annotation procedures. Genome annotation involves structural annotation, where gene features such as UTR's, intron and exon boundaries are detected and functional annotation where the biological role of the detected genes are inferred (Anon, 2012).

Gene annotation in wheat presents greater challenges in addition to those discussed above as the wheat genome primarily consists of repetitive sequences and is polyploid. Repeats confound gene annotation, they are not always well conserved, are numerous and form nested structures within other repeats. Therefore, repeats need to be identified and annotated prior to gene finding, a process termed as repeat masking. Repeat sequences in wheat and *Triticeae* in general have been annotated and curated in public databases such as the *Triticeae* Repeat Sequence Database (TREP) (Wicker *et al.*, 2002) and Repbase (Jurka *et al.*, 2005; W., Bao *et al.*, 2015). As of the writing of this thesis, the TREP database had the most comprehensive collection of manually curated *Triticeae* specific repeat elements with 477 compared to Repbase with 62. Several tools are available for the identification and annotation of repeats based on homology and *de novo* approaches. These include MITE-Hunter (Han and Wessler, 2010), The Genome Parsing Suite (GPS) (McClure *et al.*, 2005), RepeatScout (Price *et al.*, 2005) and WindowMasker (Morgulis *et al.*, 2005). Identified regions are marked with either N's or in lowercase nucleotide characters using tools such as RepeatMasker (Tempel, 2012) and CENSOR (Huda and Jordan, 2009).

After repeat masking, gene prediction is undertaken using both *ab initio* approaches, homology to known proteins and alignments to ESTs and RNA-seq datasets. *Ab initio* gene prediction relies on mathematical models derived from a training set of genes. Genic properties such as distributions and sizes of introns and exons, codon usage, GC content and intron-exon boundaries are used to identify potential genes. Such gene predictors include GeneMark-ES {ErVLukashin:2002ut}, Augustus (Stanke and Waack, 2003) and SNAP (Korf, 2004). The use of manually curated protein sequences from UniProtKB/SwissProt (Bairoch, 2008; Bairoch *et al.*, 2004; Schneider *et al.*, 2004), ESTs and RNA-seq datasets improves the accuracy and sensitivity of gene prediction. Such resources from closely related organisms if available can also be used. Algorithms capable of identifying alternate splicing events such as Exonerate (Slater and Birney, 2005) and Splign (Kapustin *et al.*, 2008) are also used to identify alternate splicing and exon boundaries. Mapping of RNA-seq data requires either assembly of the RNA-seq data prior to mapping to the genome or mapping to the genome followed by assembly of alignments. Assembly of RNA-seq data prior to mapping to the genome can be accomplished using tools such as Trinity (Grabherr *et al.*, 2011), ABySS (Simpson *et al.*, 2009) and SOAPdenovo (R., Li *et al.*, 2010) while assembly of RNA-seq alignments can be accomplished using tools such as TopHat and Cufflinks (Trapnell *et al.*, 2012) and GSNAP (T., D., Wu and Nacu, 2010).

Gene annotation in wheat is based on a combination of multiple approaches. These include the use of *ab initio* gene prediction algorithms, with supporting experimental evidence provided by expressed sequence Tag datasets (EST's), RNA-seq data and similarity to known proteins. Several databases contain highly curated *Triticeae* gene annotations including Swissprot /ExPASy (Schneider *et al.*, 2004), Ensembl plants (Bolser *et al.*, 2014), GrainGenes (O'Sullivan, 2007) and the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov). Given the combination of diverse gene prediction programs, databases and datasets, efficient and comprehensive annotation becomes a daunting task necessitating automation. Automation involves the use of multiple tools, with an aim of generating high confidence gene annotations evidenced by protein, EST and RNA-seq alignments in a high throughput manner. Annotation algorithms are thus combined into annotation pipelines such as TriAnnot (Leroy *et al.*, 2012), MAKER (Campbell *et al.*, 2013) and DAWGPAWS (Estill and Bennetzen, 2008).

Annotation pipelines enable easy configuration, inclusion, and assessment of multiple methods of annotation while enabling comparison of annotation methods across the wider wheat genome research community. Incorrect annotations not only confuse downstream analysis, but are also propagated forward when annotations are used in other genome annotation efforts. A common metric developed by Sequence Ontology Project for the assessment of the quality of a genome annotation and also for comparison of genome annotations is the annotation edit distance (AED) (Eilbeck *et al.*, 2009). The AED score is calculated based on the overlaps between a predicted gene model and aligned supporting evidence where an AED of 0 represents perfect evidence and an AED score of 1 represents no supporting evidence for the predicted gene model. Thus to effectively annotate and evaluate the quality of a genome annotation, sufficient supporting experimental evidence (ESTs, RNA-seq and annotated protein sequences) should he used in the annotation process.

Functional annotation of novel DNA sequences and predicted gene models based on experimental evidence has lagged behind the massive data produced by NGS projects and the corresponding increase in assembled genomes. High throughput functional annotation of genomes currently relies on sequence homology at the nucleotide or protein level of newly sequenced genes to genes with experimentally confirmed functions. Gene ontology (GO) terms are then used to transfer the putative functional annotations. GO (http://www.geneontology.org) is a species neutral structured vocabulary of terms with identifiers that are used to provide a formal and standardized representation of biological knowledge (Thomas *et al.*, 2007). GO terms ensure where putatively orthologous sequences have been detected between a new gene and a previous curated gene, the functional information transferred across is uniform.

For example, recognizing that descriptive terms "isoform", "splice variant" and "alternate splicing" refer to the same biological process "protein translation". This is possible through the storage of terms such as "splice variant" and "alternate splicing" as nodes in a directed acyclic graph (DAG) with the relationship between the terms such as "is a" represented as edges. In the example presented, "splice variant" (node) would be linked with an edge "is a" to node "protein translation". At the highest level of the GO hierarchy, terms are categorised into three main groups; biological process (P), molecular function (F) and cellular component (C). Given that GO annotations on curated genomic sequences are produced computationally or manually, evidence codes are also associated with GO terms when these are assigned to a gene (Škunca *et al.*, 2012). Evidence codes are grouped into three forms; electronic evidence, curated but non experimental evidence and experimental evidence. The use of evidence codes offers further insights into the reliability of a transferred GO term during annotation.

Several studies have shown the reliability of inferring biological functions from orthologues sequences across species (Koonin *et al.*, 1996; Altenhoff and Dessimoz, 2009; Nehrt *et al.*, 2011). Putative orthologues are commonly identified by sequence similarity using BLAST algorithms (McGinnis and Madden, 2004), distant tree based methods such as OrthoMCL (L., Li *et al.*, 2003) or similarity searches based on protein domain signatures using pipelines such as InterProScan (Mulder and Apweiler, 2007; Jones *et al.*, 2014). GO annotations are then transferred from the well curated target sequence to the unknown query sequence. While there are a variety of disparate databases that store well annotated nucleotide sequences, such as NCBI, protein sequences such as SwissProt (Schneider *et al.*, 2004), protein families based on conserved domains Pfam (Finn *et al.*, 2014) and metabolic pathways such as KEGG (Kanehisa *et al.*, 2015), there is limited interoperability between these databases. There is also no consistency in the use of biological terms. To exploit such vast resources to functionally annotate a sequence, the use of pipelines that offer interoperability between these database resources while incorporating GO mapping is a necessity.

InterProScan (Mulder and Apweiler, 2007) offers such functionality for combined searching of publicly available protein sequence, domain and metabolic pathway databases. InterProScan also offers an InterPro2GO mapping service (Burge *et al.*, 2012) that contains manually curated GO mappings of interProScan ids to GO terms. Tools such as Blast2GO (Conesa and Götz, 2008) provide a high throughput combined interface on which both nucleotide database collections such as NCBI (http://www.ncbi.nlm.nih.gov) and protein databases through InterProScan can be queried. Blast2GO returns annotations with GO terms. Other tools offering similar functionality include OntoBlast (Zehetner, 2003), AutoFact and Goblet (Groth *et al.*, 2004) but these are web based tools and thus not suitable for high throughput annotation of large genomes. Blast2GO has been used successfully in plant genome projects (Conesa and Götz, 2008), for example, for functional characterization of specific plant gene families in maize (Jiong Ma *et al.*, 2006) and soybean (R., T., Nelson and Shoemaker, 2006). Blast2GO was thus used for the functional annotation and comparison of the 7DS BAC and 7DS WCS assemblies.

4.2 Materials and methods

4.2.1 Chromosome arm 7DS annotation

Bread wheat chromosome arm 7DS pseudo molecules were annotated using the MAKER genome annotation pipeline (Campbell *et al.*, 2013). Annotation process followed general guidelines provided by the IWGSC (http:// www.wheatgenome.org/tool.html). The guidelines proposed the annotation of repeat elements followed by repeat masking prior to gene prediction due to the high repeat content of wheat. The MAKER pipeline was configured to use the *Triticeae* Repeat Sequence Database (TREP) (Wicker *et al.*, 2002) for repeat masking and Augustus (Stanke *et al.*, 2006) for *ab initio* gene prediction. Augustus was selected as the *ab initio* gene predictor due to the availability of Augustus wheat gene models. The TREP database was used for repeat masking as it has the most comprehensive and curated collection of *Triticeae* repeats.

To support annotation, additional empirical evidence for structural annotation was provided by ESTs, protein and RNA-seq datasets. EST datasets from several *Triticeae* species were downloaded from the NCBI database (http://www.ncbi.nlm.nih.gov) and the GrainGenes database (http://wheat.pw.usda.gov). These included 6,596 *T. aestivum* ESTs downloaded from the GrainGenes database, 1,287,297 *T. aestivum* ESTs from NCBI, 21,793 *T. turgidum* ESTs from NCBI, 11,190 *T. monococcum* ESTs from NCBI and 9, 343 *T. dicoccoides* ESTs from NCBI. All 139,165 *T. aestivum* swissprot proteins downloaded from NCBI. *T. aestivum* RNA-seq datasets were downloaded from (https://urgi.versailles.inra.fr/files/RNASeqWheat/). The RNA-seq datasets totalled 1,720 Million reads which were generated from wheat Chinese spring cultivar's five organs (leaves, root, stem, spike and grain) and from three developmental stages each (Choulet *et al.*, 2010; Cossegal *et al.*, 2008). The RNA-seq samples were sequenced in replicates on the Illumina HiSeq 2000. The comprehensive sampling of RNA transcripts ensured differentially expressed genes in the different organs and developmental stages were captured and used to improve gene annotation.

RNA-seq datasets were mapped to the whole wheat genome assembly to avoid false mappings as a result of wheat homoeologous arms. PE reads from the IWGSC isolated chromosome arm assemblies were downloaded from NCBI short read archive and re-assembled by Juan Montenegro (Dave Edward's group) using the Velvet assembler. All wheat short and long chromosome arm assemblies of chromosome 1A, 1B, 1D, 2A, 2B, 2D, 3A, 3D, 4A, 4B, 4D, 5A, 5B, 5D, 6A, 6B, 6D, 7A, 7B and 7DL were used as a combined reference for mapping. A BAC assembly of chromosome 3B (Choulet *et al.*, 2014) was also included in the combined reference. The whole chromosome shotgun (WCS) assembly of chromosome 7DS and the BAC assemblies of 7DS were each independently added to the combined wheat genome reference to generate two genome references, one with the 7DS BAC assembly and the other with the 7DS whole chromosome shotgun (WCS) assembly. RNA-seq reads were mapped to the two references independently using Tophat (Trapnell *et al.*, 2012) and transcripts assembled with Cufflinks (Trapnell *et al.*, 2012). Resulting gff files of RNA-seq mappings were then used in combination with *Triticum* ESTs and swissprot proteins described above to annotate the 7DS BAC and 7DS WCS assemblies using the MAKER annotation pipeline. A comparison of gene calls between the 7DS BAC and 7DS WCS assemblies was done to determine the effect of assembly approaches on genome annotation.

Isoform lengths of 7DS BAC and 7DS WCS assemblies were compared to determine if there was a significant difference in the mean lengths. Isoform lengths were retrieved from Cufflinks mappings of RNA-seq datasets from five organs (leaves, root, stem, spike and grain). The samples were mapped onto the two combined whole wheat genome references one with the 7DS WCS assembly and the other with the 7DS BAC assembly. Only mappings specific to the 7DS WCS and the 7DS BAC assemblies were compared. Due to expected variation within RNA-seq samples and the variation in scaffold numbers between the 7DS WCS and 7DS BAC assemblies, statistical significance test was performed using Turkeys HSD to account for statistical bias that may be introduced as a result of such variations.

A comparison of genes shared between the 7DS BAC assembly and the 7DS WCS assemblies were determined by sequence similarity using reciprocal BLAST hits (RBH). Only RBH that covered > 75 % of query length of both gene pairs were accepted. This ensured genes paired as a result of RBH had a large proportion of their sequences was matched.

Functional annotation was done using Blast2GO (Conesa and Götz, 2008). 7DS BAC and WCS genes were first mapped to the latest NCBI nr database (released on Jun 10, 2015) using default BLASTX parameters (blast+ version 2.2.29). BLAST mappings were saved in xml format. 7DS BAC and WCS genes were also mapped to protein domain databases using InterProScan (Mulder and Apweiler, 2007; Jones *et al.*, 2014) version 5.15.54.0. InterProScan results were saved in xml format. Both InterProScan and BLASTX results in xml format for each of the 7DS and WCS assemblies were loaded into the Blast2GO desktop application separately. Mapping of the InterProScan and BLASTX results with the most up to date GO terms online was done from within the Blast2GO software. The retrieved GO terms were then filtered. Only GO terms specific to green plants (taxonomic id: 33090) were used for functional annotation.

GO terms for the 7DS BAC and WCS assemblies were compared to determine if there were any enriched or depleted gene families. Enrichment analysis was performed using Fisher's exact test. Lower levels of GO terms in the GO ontology directed acyclic graph (DAG) have more precise GO term descriptions than higher level GO parent GO terms. GO terms across the same level usually differ in their specificity of the GO term description (Raychaudhuri *et al.*, 2002; Soldatova and King, 2005). GO terms from the enrichment analysis were further analysed for specificity, where higher level GO terms were removed if a lower level child GO term was present.

4.2.2 Chromosome arm 7DS repeat annotation

Repeats on the bread wheat chromosome arm 7DS pseudo molecules were annotated using Repeat Masker version 4.0.5 (http://repeatmasker.org) with the repeat annotation database Repbase update of 27/4/2016 (W., Bao *et al.*, 2015). Repeat Masker was run against *Triticeae* repeats. Repeats in the 7DS WCS assemblies were also annotated with the same parameters. Proportions of annotated repeat families in the 7DS BAC and WCS assemblies were compared.

4.3 Results

4.3.1 Chromosome arm 7DS BAC assembly annotation

Annotation of the 7DS BAC scaffolds using the MAKER annotation pipeline resulted in the annotation of 7,485 genes consisting of 24,246 exons, 955 5' UTRs and 1,506 3' UTRs resulting in a gene density of 18.3 per Mb. Annotated genes had on average 3.2 exons per gene with a median of 2. The average and median gene, intron, exon, 5'UTR and 3'UTR sizes are shown in Table 4-1.

Table 4-1: Genome feature lengths of 7DS BAC assemblies.

Feature	Mean bps	Median bps
Genes	1994.7	1280
Introns	425.7	152
Exons	297.1	193
5' UTR	186.8	120
3' UTR	325.9	247

A check of the quality of annotations of the 7, 485 annotated genes showed all genes had AED scores < 1, 90% had AED scores < 0.55 and 50% had AED scores < 0.27.

BLASTX hits of the 7, 485 BAC genes onto the NCBI non redundant (nr) database resulted in 7,234 genes having hits. 251 genes did not have a match in the nr database. The BLASTX hits corresponded to a total of 124,041 High-scoring Segment Pairs (HSPs). HSPs are local maximal alignments of the query sequences against the BLAST database sequences. The distribution of HSPs per BLAST hit showed 55.3 % (68,671) of the BLAST hits corresponded to single HSPs while 21.4 % (26,530) of the BLASTX hits corresponded to two HSPs. The remaining 4.8% - 0.6 % of the BLASTX hits corresponded to 4 – 10 HSPs respectively (Figure 4-1).

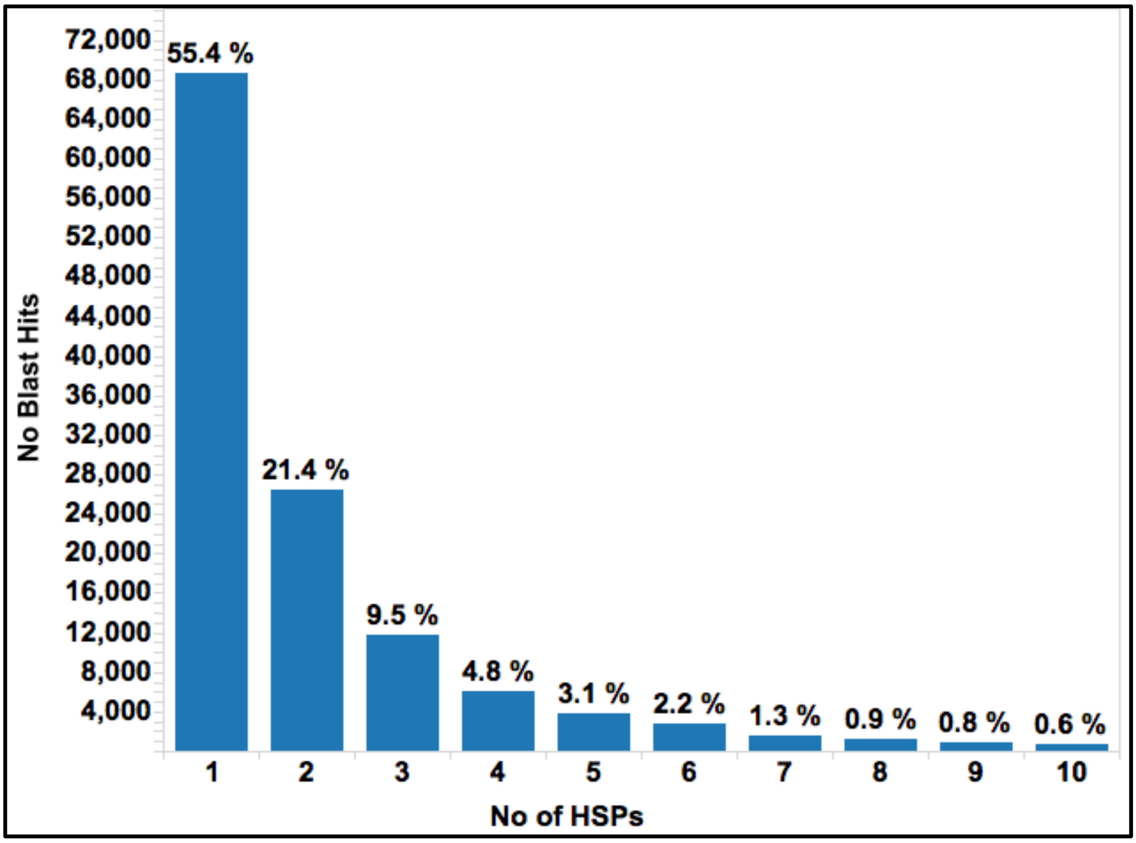

Figure 4-1: Distribution of HSPs for Blastx hits of 7DS BAC genes.

The species distribution of all top BLAST hits for each gene showed 3, 344 genes (42.6%) had hits to *Aegilops tauschii,* 1, 263 (17.5 %) to *Triticum urartu,* 884 (12.2%) to wheat *(Triticum aestivum),* 573 (7.9 %) to *Hordeum vulgare,* 325 (4.5 %) to *Brachypodium distachyon*, 243 (3.4 %) to *Oryza sativa* and 602 (8.3 %) to other species (Figure 4-2).

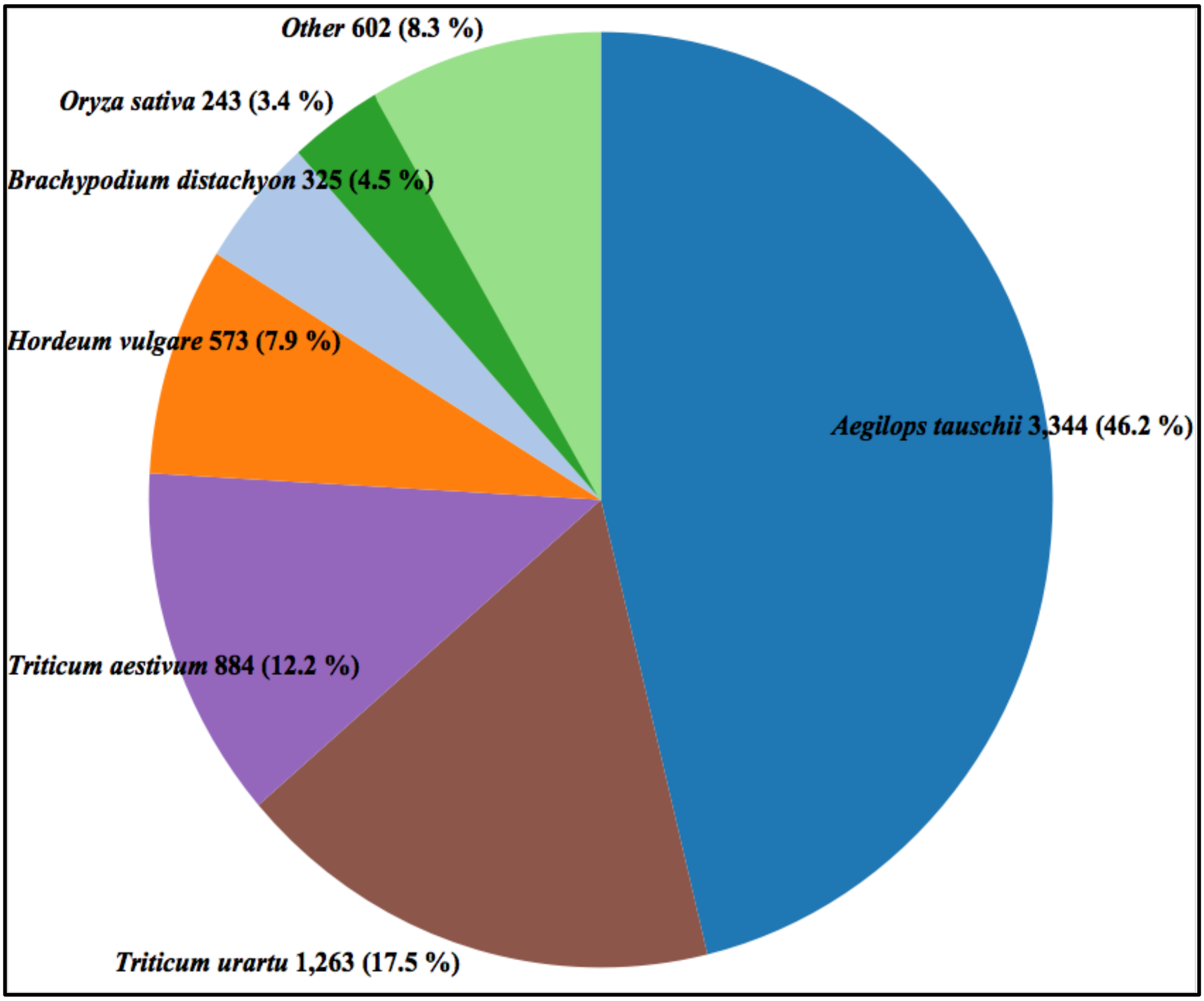

Figure 4-2: Species distribution of top BLASTX hits of 7DS BAC genes on the NCBI nr database.

BLASTX hits were mapped to 263,153 GO terms. Evidence codes for the GO terms retrieved showed 253, 040 (96.16%) of the GO terms were assigned to sequences in the databases as a result of Electronic Annotation (IEA) and 6,287 (2.39%) as a result of Reviewed Computational Analysis (RCA). The rest of the evidence codes constituted < 1% of the GO terms (Table 4-2).

Table 4-2:Distribution of evidence codes for 7DS BAC genes GO annotation.

Evidence Codes	No Hits	% Hits
Inferred from Electronic Annotation (IEA)	253040	96.16
Inferred from Reviewed Computational Analysis (RCA)	6287	2.39
Inferred from Biological aspect of Ancestor (IBA)	1994	0.76
Inferred from Direct Assay (IDA)	623	0.24
Inferred from Sequence or Structural Similarity (ISS)	304	0.12
Inferred from Mutant Phenotype (IMP)	232	0.09
Inferred by Curator (IC)	173	0.07
Traceable Author Statement (TAS)	119	0.05
Inferred from Expression Pattern (IEP)	111	0.04
Inferred from Expression Pattern (IPI)	95	0.04
Inferred from Sequence Model (ISM)	93	0.04
No Biological Data Available (ND)	35	0.01
Inferred from Genetic Interaction (IGI)	25	0.01
Inferred from Sequence Orthology (ISO)	13	0.00
Non-traceable Author Statement (NAS)	9	0.00

Most of the GO terms 255,422 (97 %) were retrieved from the UniprotKB database (Bairoch, 2008)and 6,383 (2.43%) were retrieved from the Gramene protein database (GR Protein) (Tello-Ruiz *et al.*, 2016). The remaining terms were retrieved from the *Arabidopsis* Information Resource (TAIR) database (Poole, 2007), the Mouse Genome Informatics (MGI) database (Bolser, 2014), the Zebrafish Information Network (ZFIN) database (Howe *et al.*, 2013) and the WormBase (WB) database (Table 4-3).

Table 4-3: Sources of GO terms for 7DS BAC gene annotation.

Database	No. of GOs	% GOs
UniprotKB	255422	97.06
GR Protein	6383	2.43
TAIR	1253	0.48
MGI	50	0.02
ZFIN	41	0.02
WB	4	0.00

Mapping of the 7, 485 BAC genes onto protein domain signature databases using InterProScan showed of the 7, 485 BAC genes, only 1, 150 (15.3%) had hits to InterProScan databases. Of these, only 314 had GO terms associated. From a combination of the nr and InterProScan mappings, 2,929 genes (39.1%) of the 7,485 7DS BAC genes did not have a GO term associated. The remaining genes were associated with 263,153 GO terms. A gene sequence can be associated with multiple GO terms and a GO term can also be associated with several genes. At GO term level 2, 18, 671 sequences were associate with 28 GO terms. The top 5 GO terms with the most sequences associated were binding

(2,845), catalytic activity (2,552), cellular process (2,306), metabolic process (2,275) and cell (2,241) (Figure 4-3).

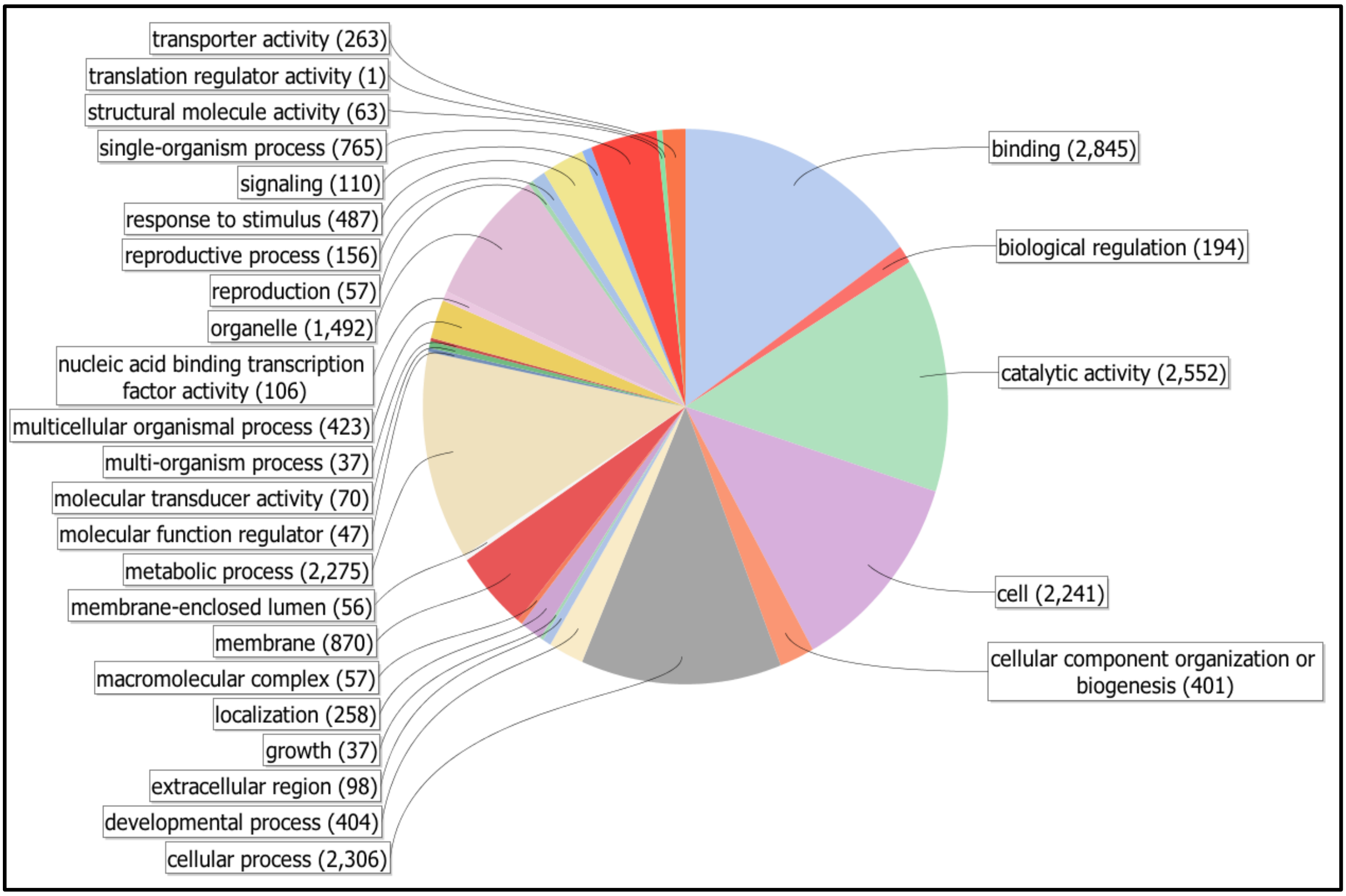

Figure 4-3:Sequence distribution at GO level 2 for 7DS BAC annotation.

4.3.2 Chromosome arm 7DS WCS assembly annotation

Annotation of the 7DS whole genome shotgun velvet assemblies using the MAKER annotation pipeline resulted in the annotation of 2,422 genes consisting of 7,258 exons, 359 5' UTRs and 530 3' UTRs. The subsequent gene density was 19.6 per Mb. Annotated genes had on average 2.9 exons per gene with a median of 2. The average and median gene, intron, exon, 5'UTR and 3'UTR sizes are shown in Table 4-4.

Table 4-4: Genome feature lengths of 7DS WCS assemblies.

Feature	Mean	Median
Genes	1478.4	994
Introns	145.3	129
Exons	259.7	168
5_UTR	276.6	142
3_UTR	414.4	282

Annotation quality of the 2, 422 genes showed all genes had AED scores < 1, 90% had AED scores < 0.49 and 50% had AED scores < 0.25.

BLASTX hits of the 2,422 WCS genes onto the NCBI non redundant (nr) database resulted in 2,352 genes with a database match while 70 WCS genes did not have a match. BLASTX hits corresponded to a total of 40,076 HSPs. The distribution of HSPs per BLAST hit showed 52.8 % (21,171) of the BLASTX hits corresponded to single HSPs while 23.7 % (9,514) of the BLASTX hits corresponded to two HSPs. The remaining 9.9% - 0.3 % of the BLASTX hits corresponded to 4 – 10 HSPs respectively (Figure 4-4).

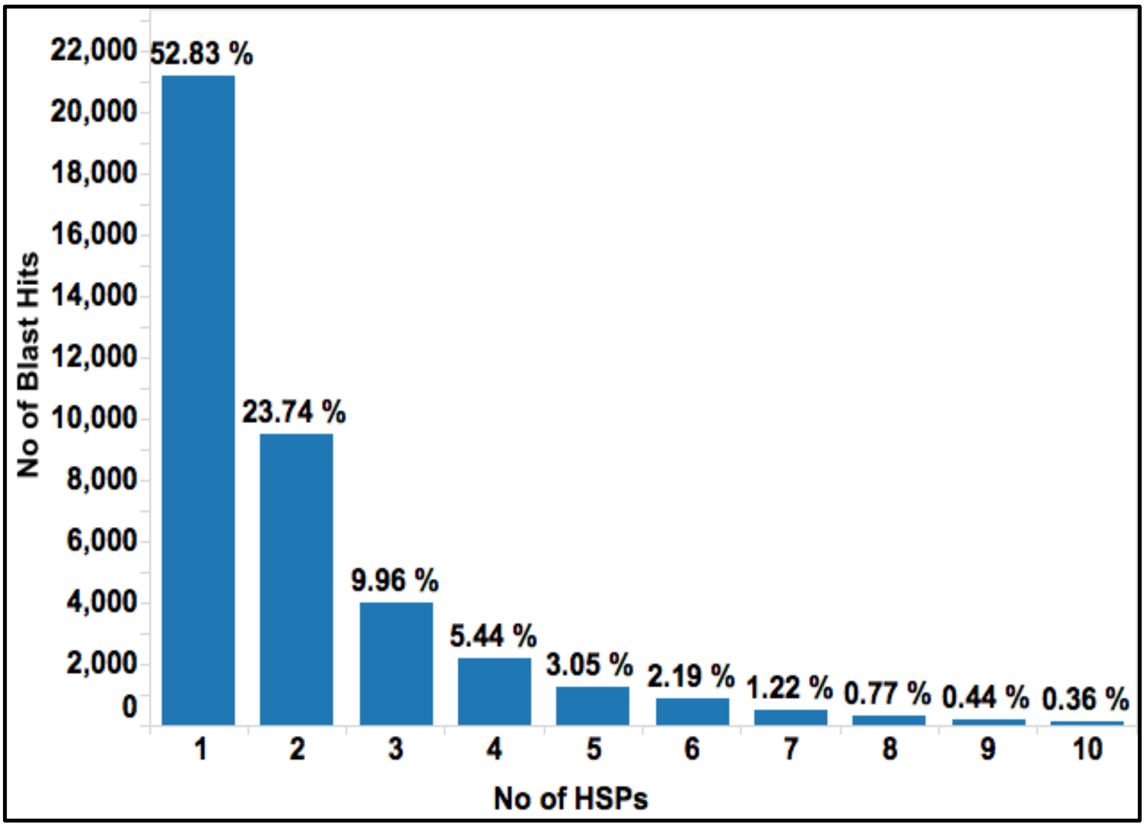

Figure 4-4: Distribution of HSPs for BLASTX hits of 7DS WCS genes.

The species distribution of all top BLAST hits for each gene showed 1, 217 genes (51.7 %) had hits to *Aegilops tauschii,* 445 (18.9 %) to *Triticum urartu,* 236 (10 %) to wheat *(Triticum aestivum)*, 95 (4 %) to *Hordeum vulgare,* 119 (5 %) to *Brachypodium distachyon*, 62 (2.6%) to *Oryza sativa* and 178 (7.5 %) to other species (Figure 4-5).

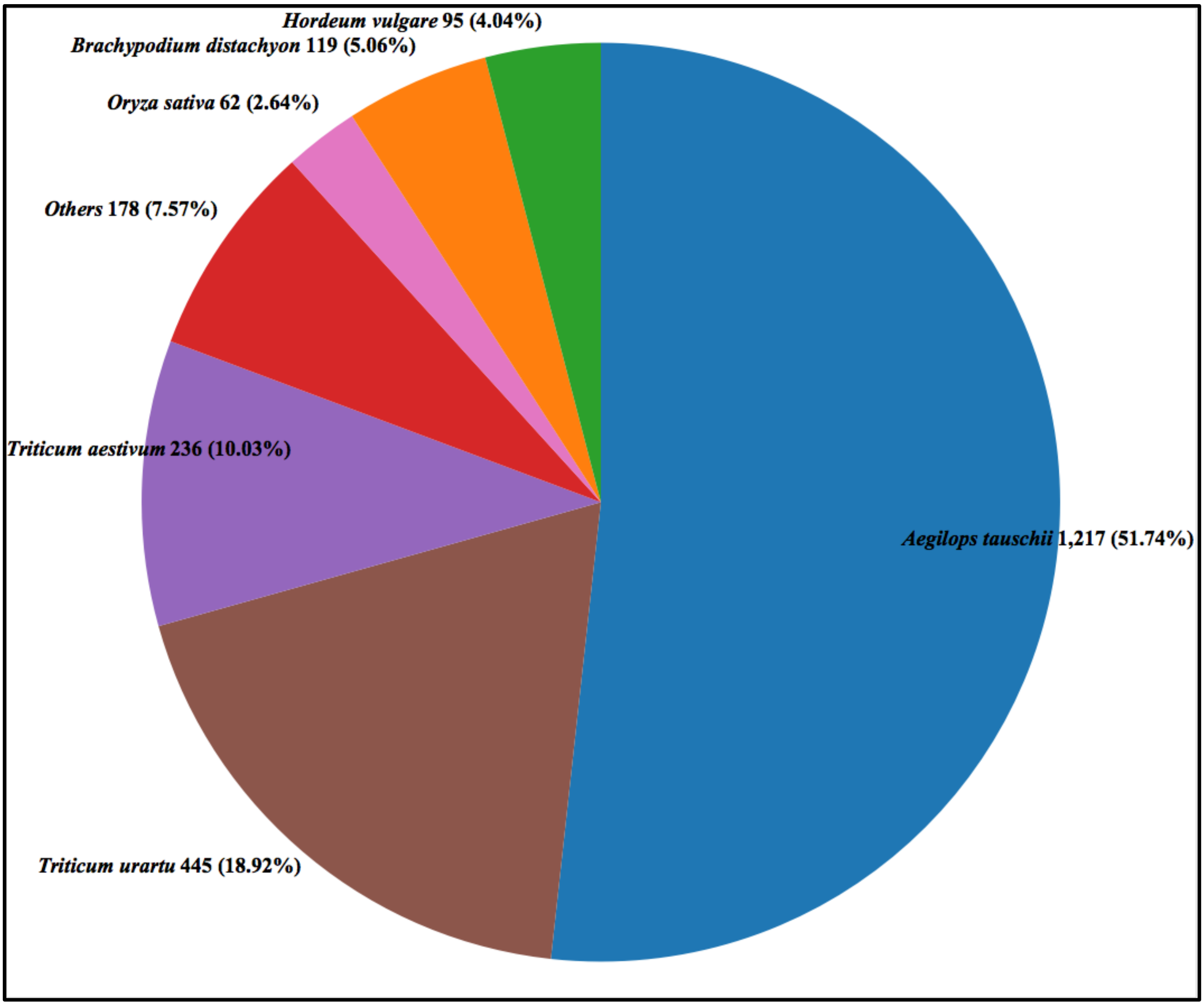

Figure 4-5: Species distribution of top BLASTX hits of 7DS WCS genes on the NCBI nr database

BLASTX hits were mapped to 80,286 GO terms. Evidence codes for the GO terms retrieved showed 77,186 (96.14 %) of the GO terms were assigned to sequences in the databases as a result of Electronic Annotation (IEA) and 1, 977 (2.46%) as a result of Reviewed Computational Analysis (RCA). The rest of the evidence codes constituted < 1% of the GO terms (Table 4-5).

Table 4-5: Distribution of evidence codes for 7DS WCS genes GO annotation.

Evidence Codes	No Hits	% Hits
Inferred from Electronic Annotation (IEA)	77186	96.14
inferred from Reviewed Computational Analysis (RCA)	1977	2.46
Inferred from Biological aspect of Ancestor (IBA)	727	0.91
Inferred from Direct Assay (IDA)	162	0.20
Inferred by Curator (IC)	74	0.09
Inferred from Mutant Phenotype (IMP)	38	0.05
Inferred from Sequence or Structural Similarity (ISS)	35	0.04
Traceable Author Statement (TAS)	31	0.04
Inferred from Expression Pattern (IEP)	28	0.03
Inferred from Sequence Model (ISM)	10	0.01
Inferred from Genetic Interaction (IGI)	10	0.01
Inferred from Expression Pattern (IPI)	7	0.01
No Biological Data Available (ND)	1	0.00

Most of the GO terms 78,057 (97.22 %) were retrieved from the UniprotKB database and 2,051 (2.55%) were retrieved from the Gramene Protein database. The remaining terms were retrieved from TAIR and FlyBase (FB) (Crosby *et al.*, 2007) databases (Table 4-6).

Table 4-6: Sources of GO terms for 7DS WCS gene annotation.

Database	**No. of GOs**	**% GOs**
UNIPROTKB	78057	97.22
GR_PROTEIN	2051	2.55
TAIR	177	0.22
FB	1	0.00

Mapping of the 2,422 WCS genes onto protein domain signature databases using InterProScan showed of the 2,422 WCS genes, only 186 (7.6. %) had hits in InterProScan databases. Of these, only 31 had GO terms associated. From the combined GO annotations retrieved from the nr and InterProScan matches, of the 2,422 total 7DS WCS genes, 925 (38.1%) did not have a GO term associated. The remaining genes were associated with 80,286 GO terms. At GO term level 2, 7,083 sequences were associate with 38 GO terms. The top 5 GO terms with the most sequences associated were metabolic processes (981), binding (935) , catalytic activity (803), cellular process (799), and cell (639) (Figure 4-6).

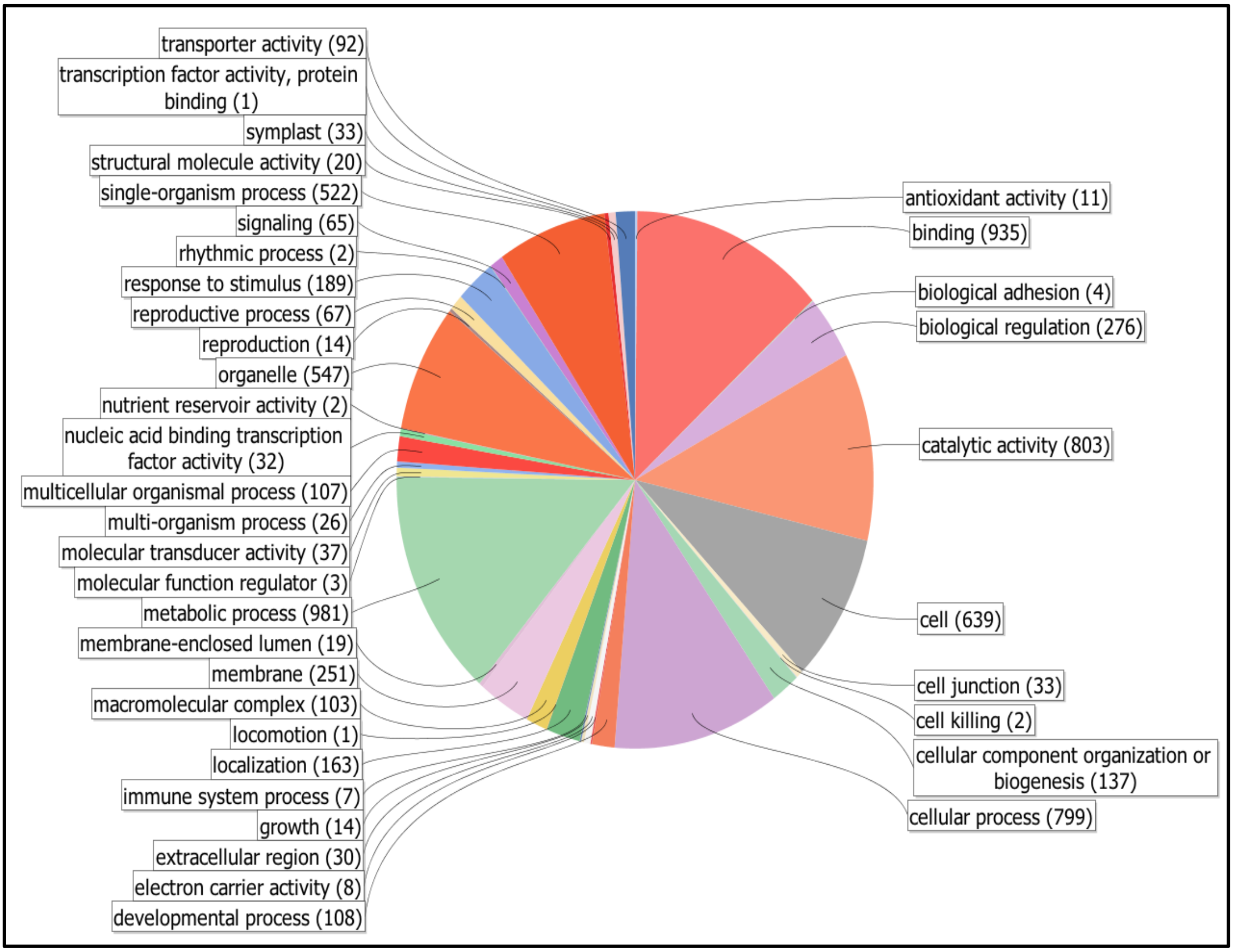

Figure 4-6: Sequence distribution at GO level 2 for 7DS WCS annotation.

4.3.3 A comparison of gene annotations of chromosome arm 7DS WCS and 7DS BAC-by-BAC assemblies

The 7DS WCS AED scores are significantly smaller (p < 0.0001, df 4388.9, t ratio - 6.7) than the BAC assemblies. The 7DS WCS assemblies had a higher density of AED scores between 0 – 0.33 than the 7DS BAC assemblies but the BAC assemblies had a higher density of AED scores above

0.33 (Figure 4-7).

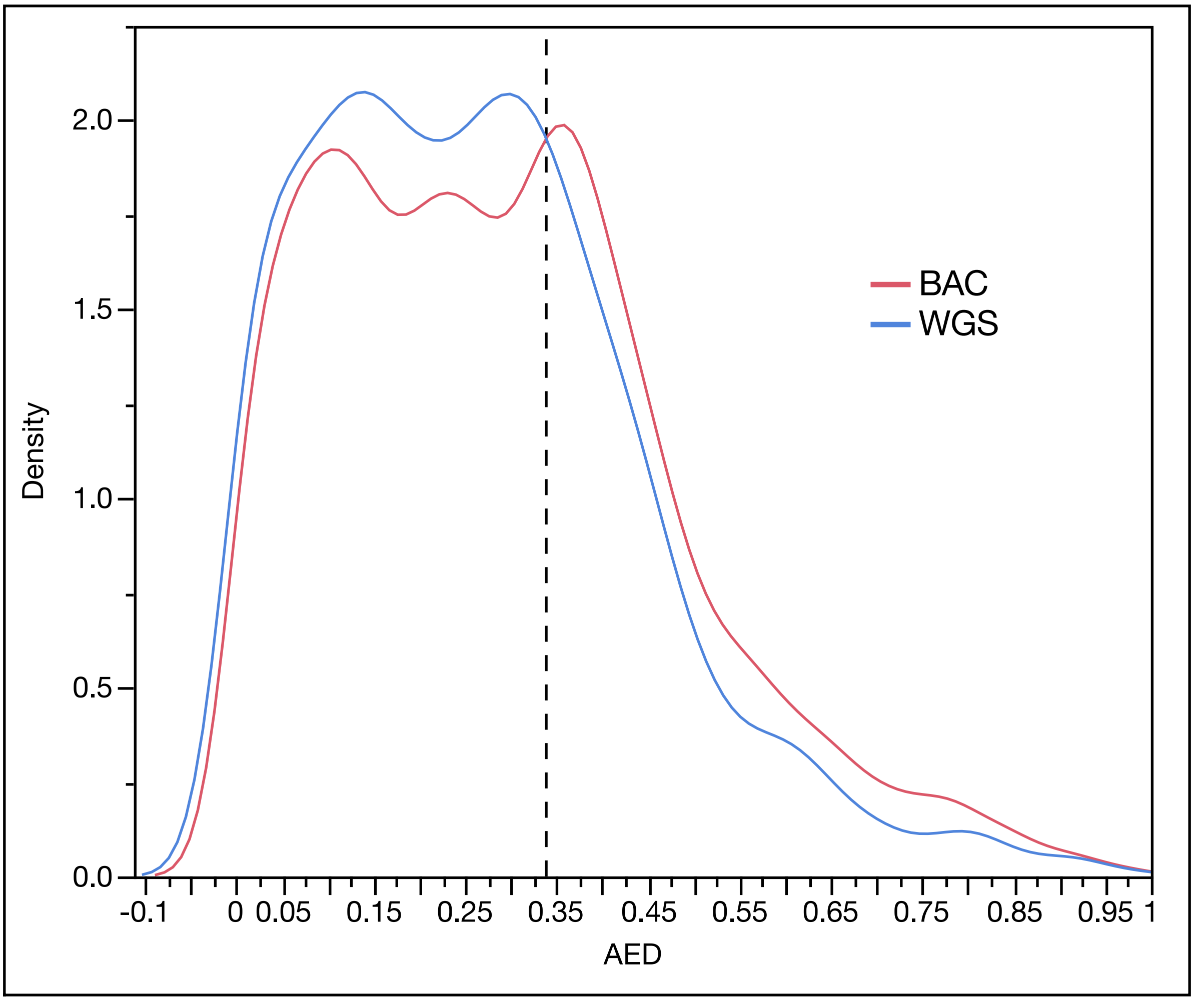

Figure 4-7: Kernel density function distributions of AED scores of 7DS BAC and WCS annotations.

Gene lengths of the 7DS BAC assemblies were significantly larger compared to the 7DS WCS assemblies (p < 0.0001, df 7734.8, t ratio -11.7). Exon lengths of the 7DS BAC assemblies were also significantly larger compared to those of the 7DS WCS assemblies (p < 0.0001, df 13075.4, t ratio -9.3). In contrast to the above observations, lengths of 5' and 3' UTR regions were significantly larger in the WCS assemblies than the BAC assemblies (p < 0.0001, df 524.7, t ratio 3.9) and (p < 0.0001, df 804.1, t ratio 4.1) respectively. A comparison of kernel densities of the lengths of WCS and BAC 5' and 3' UTRs showed although the WCS assembly UTR's were significantly larger than those of the BAC assembly, the size distributions were similar (Figure 4-8, Figure 4-9) . The peaks of the density plots of the 5' UTR's and 3' UTRs in both assemblies were at the same size values of 93.4 bps for the 5' UTR's (Figure 4-8) and 215.8 bps for the 3' UTRs (Figure 4-9), but the BAC assemblies had higher densities at these sizes as shown by read peaks in Figure 4-8 and Figure 4-9.

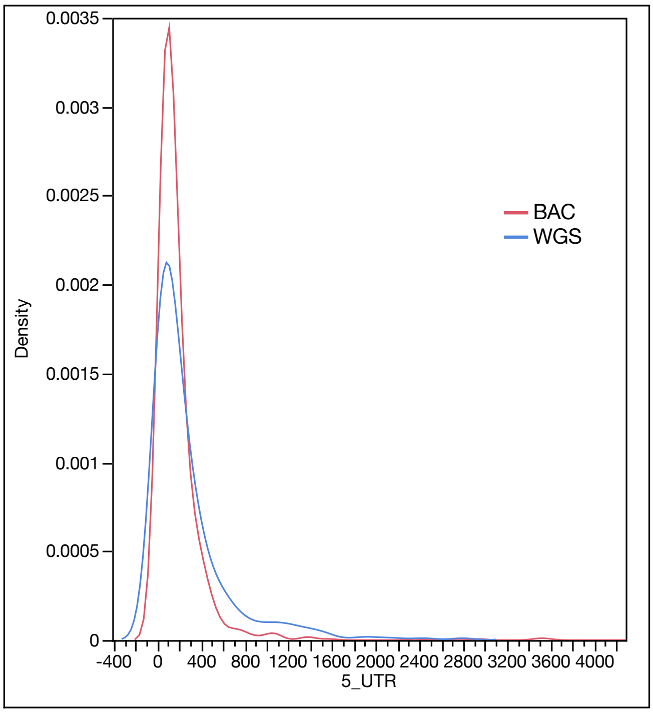

Figure 4-8: Kernel density function distributions of 5’ UTR sizes of 7DS BAC and WGS annotations.

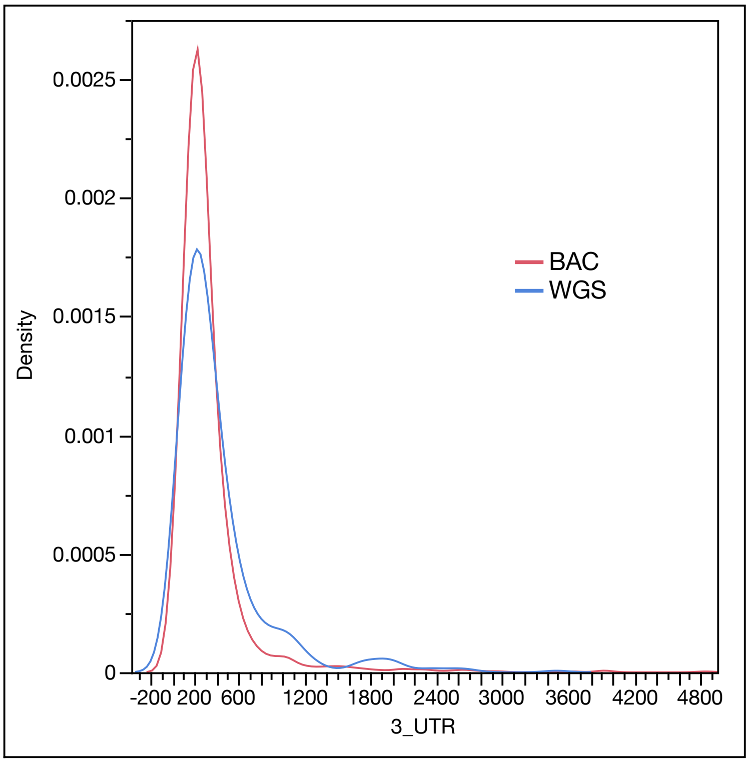

Figure 4-9: Kernel density function distributions of 3’ UTR sizes of 7DS BAC and WCS annotations.

A comparisons of lengths of Cufflinks assembled isoforms between the 7DS BACs and 7DS WCS assemblies showed that the 7DS BAC’s isoforms were significantly larger on average by 487 bps than those of the 7DS WCS assemblies across the stem, root, spike, leaf and grain samples (Table

4-7), (p < 0.0001, df 39152.8, t ratio 70). A total of 26,370 and 75,278 isoforms were generated from the assembly of RNA-seq datasets for the BAC and WCS assemblies respectively.

Table 4-7:Isoform length statistics on 7DS BACs and 7DS WCS assemblies.

Assembly	Organ	N	Avg. length (bps)	Median length (bps)	Std. dev (bps)
BAC	Stem	5,844	1,422	1,191	981
	Spike	6,439	1,568	1,319	1,110
	Root	5,310	1,277	1,097	892
	Leaf	5,229	1,453	1,214	1,000
	Grain	3,548	1,484	1,261	994
WCS	Stem	16,861	949	724	794
	Spike	18,722	1,030	771	868
	Root	15,000	844	609	756
	Leaf	14,860	961	721	817
	Grain	9,835	988	763	816

Reciprocal best BLAST hit analysis of 7,485 7DS BAC genes against 2,422 7DS WCS genes identified 411 shared genes. This represented 5.4 % of the 7DS BAC genes and 16.9 % of the 7DS WCS genes. The same gene numbers and pairs were retrieved by using either highest BLAST bit scores or lowest BLAST e-values as the selection criteria for a reciprocal best hit.

Similar percentages of the 7DS BAC genes and 7DS WCS genes had hits on the nr database, 96.6% vs 97.1 % respectively, even though the 7DS BAC genes were 3 x more than the 7DS WCS genes in number (7,485 vs 2,422). An equally higher number (3.5 x) of 7DS BAC genes did not have hits in the nr database compared to the 7DS WCS genes (251 vs 70 respectively). As expected, the higher number of genes in the 7DS BAC assembly resulted in a higher number of HSPs matching on the 7DS BAC genes compared to the 7DS WCS genes (124,041 HSPs vs 40,076 HSPs respectively). A slightly higher percentage of the 7DS BAC genes' BLAST hits 55.3% (68,671 genes) corresponded to single HSPs compared to 52.8 % (21,171 genes) from the 7DS WCS assembly (Figure 4-1, Figure 4-4).

BLAST hit distribution of the 7DS BAC and WCS genes per species showed similar mapping percentages on the same species (Figure 4-2, Figure 4-5). The highest hits were to *Aegilops tauschii* with hits from 42.6% of the 7DS BAC genes and 51.7 % of the 7DS WCS genes. This was followed by hits to *Triticum urartu* with 17.5 % of the hits from the 7DS BAC genes and 18.5 % of the 7DS WCS genes. Hits to wheat were at 12.2 % for the 7DS BAC genes and 10% for the 7DS WCS genes.

Hits of the 7DS BAC compared to the 7D WCS genes to *Hordeum vulgare* were (7.9 % vs 4 %), *Brachypodium distachyon* (4.5 % vs 5 %) and rice (3.4 % vs 2.6 %) respectively.

Table 4-8: Enrichment analysis of 7DS BAC vs 7DS WCS GO terms.

GO-ID	Term	Category	FDR	P-Value	Over/Under
GO:0043227	membrane-bounded organelle	C	1.16E-07	1.21E-09	UNDER
GO:0043231	intracellular membrane-bounded organelle	C	1.16E-07	1.21E-09	UNDER
GO:0005794	Golgi apparatus	C	4.67E-07	7.30E-09	UNDER
GO:0043229	intracellular organelle	C	6.39E-07	1.66E-08	UNDER
GO:0043226	organelle	C	6.39E-07	1.66E-08	UNDER
GO:0012505	endomembrane system	C	5.39E-06	1.68E-07	UNDER
GO:0044444	cytoplasmic part	C	2.32E-05	8.45E-07	UNDER
GO:0005623	cell	C	1.38E-04	5.74E-06	UNDER
GO:0044464	cell part	C	2.08E-04	9.77E-06	UNDER
GO:0044424	intracellular part	C	4.55E-04	2.37E-05	UNDER
GO:0005622	intracellular	C	1.21E-03	8.80E-05	UNDER
GO:0005737	cytoplasm	C	4.83E-03	4.78E-04	UNDER
GO:0009536	plastid	C	1.63E-02	1.95E-03	UNDER
GO:0016740	transferase activity	F	1.32E-03	1.07E-04	UNDER
GO:0030234	enzyme regulator activity	F	1.95E-03	1.83E-04	UNDER
GO:0098772	molecular function regulator	F	1.95E-03	1.83E-04	UNDER
GO:0030246	carbohydrate binding	F	7.82E-03	8.15E-04	UNDER
GO:0003682	***chromatin binding***	***F***	***2.76E-02***	***3.82E-03***	***OVER***
GO:0044877	***macromolecular complex binding***	***F***	***2.76E-02***	***3.82E-03***	***OVER***
GO:0043412	macromolecule modification	P	7.26E-04	4.91E-05	UNDER
GO:0036211	protein modification process	P	7.26E-04	4.91E-05	UNDER
GO:0006464	cellular protein modification process	P	7.26E-04	4.91E-05	UNDER
GO:0009056	catabolic process	P	1.32E-03	1.10E-04	UNDER
GO:0044267	cellular protein metabolic process	P	9.66E-03	1.06E-03	UNDER
GO:0044237	cellular metabolic process	P	1.55E-02	1.78E-03	UNDER
GO:0044260	cellular macromolecule metabolic process	P	2.39E-02	2.99E-03	UNDER
GO:0007275	multicellular organismal development	P	2.76E-02	4.02E-03	UNDER
GO:0044707	single-multicellular organism process	P	2.76E-02	4.02E-03	UNDER
GO:0044767	single-organism developmental process	P	3.21E-02	4.86E-03	UNDER
GO:0032502	developmental process	P	3.21E-02	5.02E-03	UNDER
GO:0032501	multicellular organismal process	P	4.44E-02	7.17E-03	UNDER

The 7DS BAC genes were associated with 263,153 GO terms compared 80,286 GO terms for the 7DS WCS functional annotation. In both cases, 96 % of the GO terms retrieved were from Electronic Annotation (IEA) and 2 % from Reviewed Computational Analysis (RCA) (Table 4-2, Table 4-5). Similarly, for both the 7DS BAC and 7DS WCS genes, 97 % of the GO terms were retrieved from the UniprotKB database and 2 % from Gramene protein database (Table 4-3,Table 4-6). InterProScan annotations generated only 314 GO terms for the 7DS BAC genes and 31 GO terms for the 7DS WCS genes.

The top 5 GO terms associated with the 7DS BAC gene sequences and the 7DS WCS gene sequences were the same. These were **“binding”**, **“catalytic”**, **“cellular processes”**, **“metabolic processes”** and **“cell component”** (Figure 4-3, Figure 4-6). The associated sequences for the 7DS BAC genes and 7DS WCS genes were **“binding”** (2,845 vs 935), **“catalytic activity”** (2,552 vs 803) , **“cellular process”** (2,306 vs 799), **“metabolic processes”** (2,275 vs 981) and **“cell component”** (2, 241 vs 639) (Figure 4-3, Figure 4-6).

Given the top 5 GO term associations at GO term level 2 between the 7DS BAC and WCS genes were the same, a comparison of enrichment of GO terms using Fishers Exact test (with multiple testing correction for FDR) between the two gene sets (7DS BAC genes and 7DS WCS genes) was done to determine if the choice of an assembly approach results in enrichment of specific gene families. Enrichment analysis showed 31 GO terms were significantly over represented or under represented between the two gene sets (Table 4-8, Figure 4-10). All the 7DS WCS gene GO terms were used as the test set against a background of all the 7DS BAC gene GO terms. Enrichment analysis identified 29 GO terms that were enriched in the 7DS BAC genes (depleted in the 7DS WCS genes) while two GO terms were enriched in the 7DS WCS genes (depleted in the 7DS BAC genes) with p value < 0.05 and FDR 0.05 (Table 4-8). For the 29 GO terms enriched in the 7DS BAC genes, 13 were cellular component terms (C), 4 were molecular function (F) terms and 12 were biological process (P) terms (Table 4-8).

The two GO terms enriched in the 7DS WCS genes were molecular function (F) terms (Figure 4-10). Although two GO terms, **“chromatin binding”** and **“macromolecular complex binding”** were the only enriched GO terms in the 7DS WCS genes compared to the 7DS BAC genes (p value 2.76E-02) (Table 4-8, Figure 4-10), these terms were associated with < 1% of the gene sequences. The GO terms “chromatin binding and macromolecular complex binding” were both associated with 0.31% of the 7DS BAC gene sequences and 0.93 % of the 7DS WCS gene sequences (Figure 4-10).

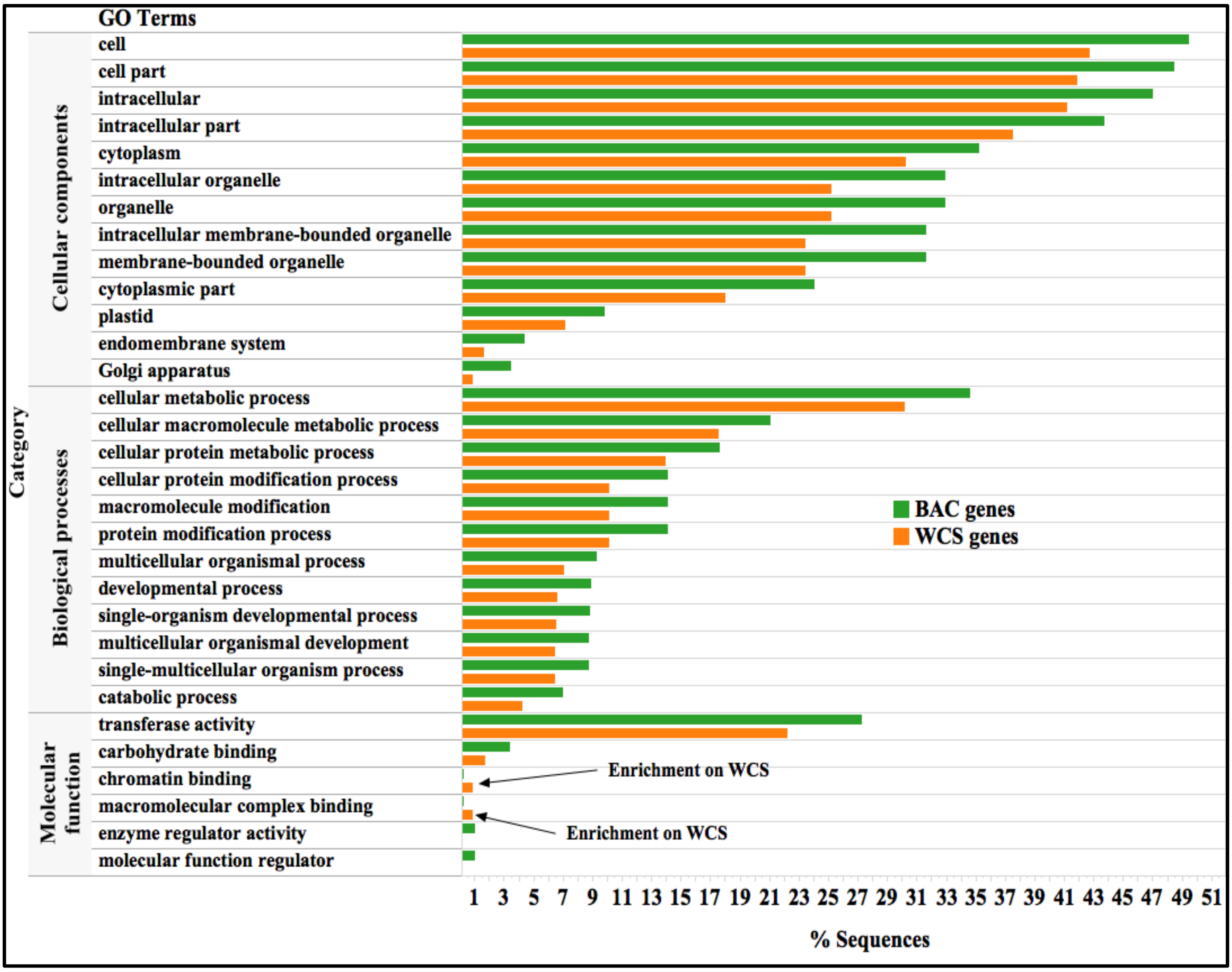

Figure 4-10: Percentage of sequences with GO Enriched terms.

While lower levels of GO terms in the GO directed acyclic graph (DAG) represent more precise GO terms, the specificity of GO terms across the same GO level is not uniform and can result in redundancy and ambiguity if a process or description is represented at multiple GO levels (Raychaudhuri *et al.*, 2002; Soldatova and King, 2005). For example, the two GO terms associated with molecular functions found to be enriched in the 7DS WCS gene sequences were, "**chromatin binding**" and "**macromolecular complex binding**". **Macromolecular complex binding *is a*** child term of **chromatin binding** in the GO DAG (red terms in Figure 4-11).

Similarly, three GO terms associated with molecular functions were found to be enriched in the 7DS BAC gene sequences. These were "**molecular function regulator**", "**enzyme regulator activity**" and "**transferase activity**". While the GO term **transferase activity** did not have a child or parent term identified in the enrichment analysis (green terms in Figure 4-11), the GO terms "**molecular function**" and "**enzyme regulator activity**" were linked with "**molecular function regulator**" being a parent term of "**enzyme regulator activity**" (red terms in Figure 4-11). Redundant parent GO terms were removed leaving the lower hierarchy more specific terms. In the above example, these parent GO terms were "**molecular function regulator**" and **"macromolecular complex binding".**

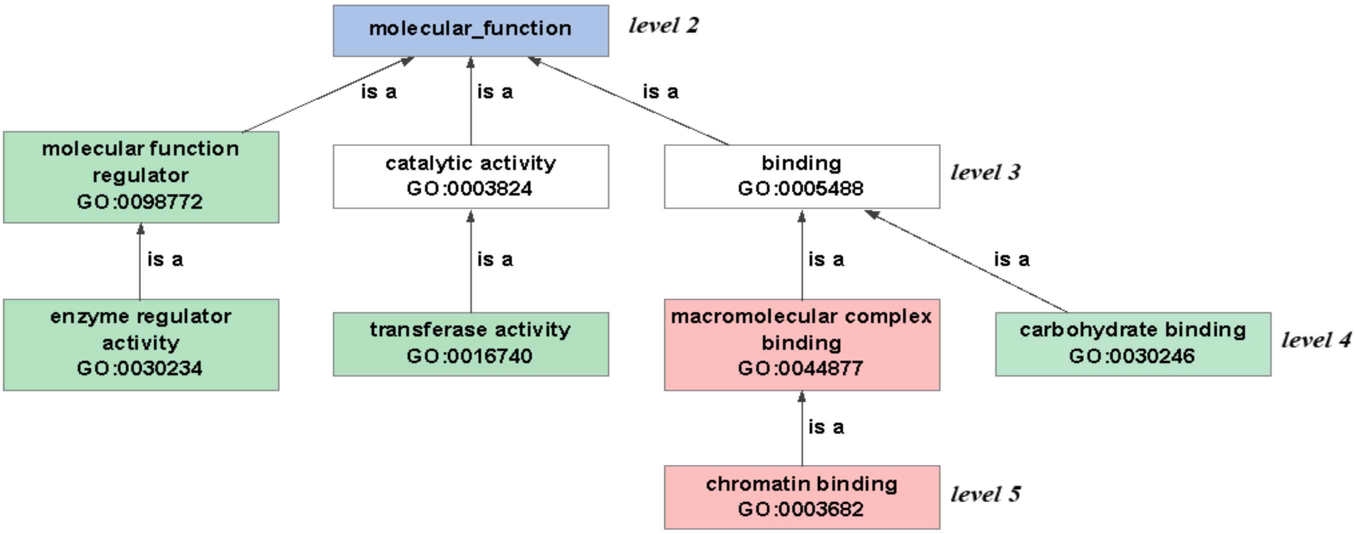
molecular_function
level 2
is a
is a
is a
molecular function regulator GO:0098772
catalytic activity GO:0003824
binding GO:0005488
level 3
is a
is a
is a
is a
enzyme regulator activity GO:0030234
transferase activity GO:0016740
macromolecular complex binding GO:0044877
carbohydrate binding GO:0030246
level 4
is a
chromatin binding GO:0003682
level 5

Figure 4-11: Hierarchy of molecular function GO enrichment terms. Terms in green are enriched in BAC gene sequences while terms in red are enriched in WCS gene sequences.

Capturing more specific GO enrichment terms by removal of general parental terms where present in the GO enriched terms resulted in a reduction of the 31 enriched GO terms to 10 GO terms (Table 4-9) with the corresponding associated gene sequence percentages (Figure 4-12).

Table 4-9: Specific GO enrichment analysis terms. Only chromatin binding (in italics) was over represented.

GO-ID	Term	Category	FDR	P-Value	Over/Under
GO:0005794	Golgi apparatus	C	4.67E-07	7.30E-09	UNDER
GO:0009536	plastid	C	1.63E-02	1.95E-03	UNDER
GO:0016740	transferase activity	F	1.32E-03	1.07E-04	UNDER
GO:0030234	enzyme regulator activity	F	1.95E-03	1.83E-04	UNDER
GO:0030246	carbohydrate binding	F	7.82E-03	8.15E-04	UNDER
GO:0003682	*chromatin binding*	*F*	*2.76E-02*	*3.82E-03*	*OVER*
GO:0006464	cellular protein modification process	P	7.26E-04	4.91E-05	UNDER
GO:0009056	catabolic process	P	1.32E-03	1.10E-04	UNDER
GO:0007275	multicellular organismal development	P	2.76E-02	4.02E-03	UNDER

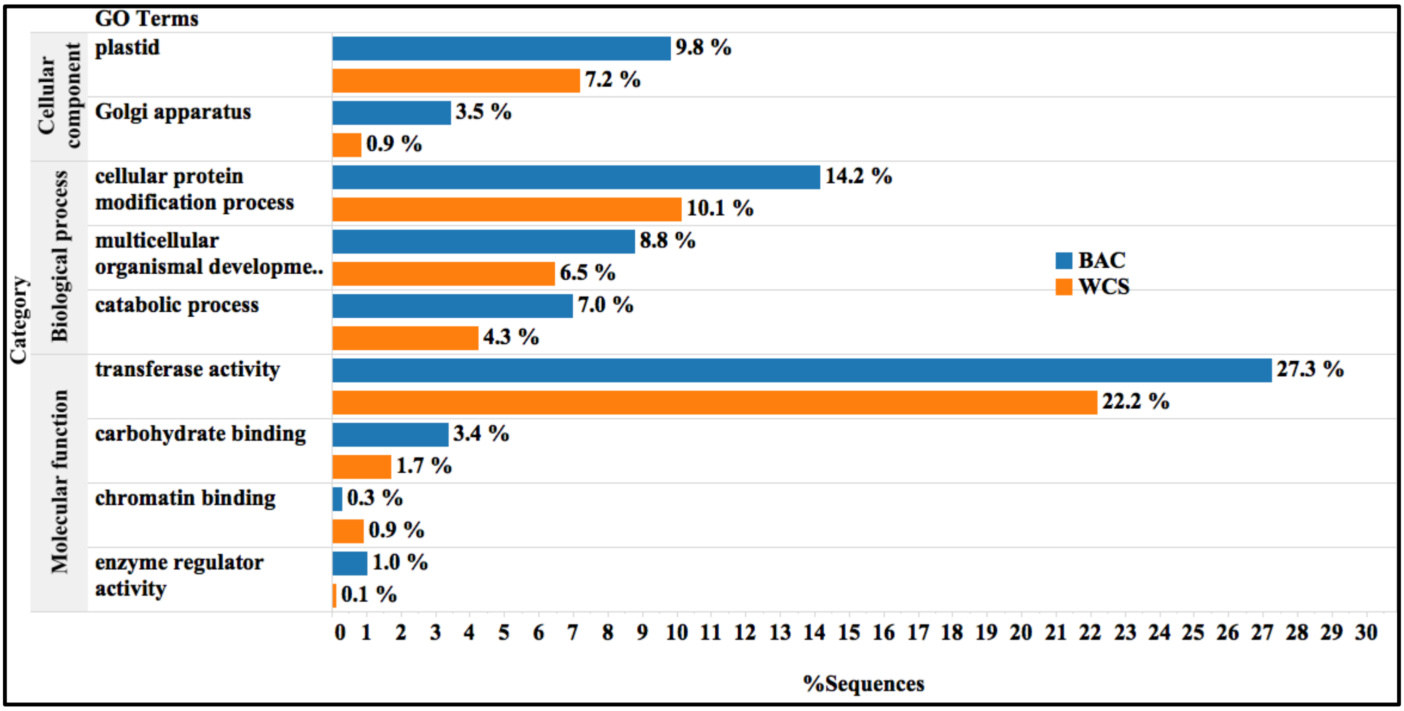

Figure 4-12: Percentages of sequences with specific GO enrichment analysis terms.

4.3.4 A comparison of repeat annotations on chromosome arm 7DS WCS and 7DS BAC-by-BAC assemblies

Annotations of repeat elements on chromosome arm 7DS BAC assemblies showed 49.59 % of the sequences contained repeats. Of these, 42.26 % were retroelements (Table 4-10). LTRs were the most prevalent retroelements in the 7DS BAC assembly at 41.65 %. The most common superfamily of retroelements was *Gypsy*/DIRS1 accounting for 29 % of the assembly. Among the DNA transposons annotated, transposons of the order Tc1-IS630-Pogo were the majority at 0.22 % (Table 4-10).

Table 4-10: Repeat annotation of 7DS BAC and 7DS WCS assemblies.

Assembly	% Bases masked	Class	Order	No. in order	Order %	Super family	No. in super family	Super family %
BAC	49.59	Retroelements	LTR elements	77906	41.65	*Gypsy*/DIRS1	59,374	28.88
						Ty1/*Copia*	18,317	12.74
						BEL/Pao	0	0
						Retroviral	0	0
			LINEs	2853	0.61	L1/CIN4	2,575	0.57
						RTE/Bov-B	278	0.04
			SINEs	65	0	Penelope	0	0
		DNA transposons	Tc1-IS630-Pogo	7171	0.22			
			Tourist/Harbinger	1420	0.12			
			hobo-Activator	115	0.01			
WCS	8.07	Retroelements	LTR elements	723	4.18	*Gypsy*/DIRS1	429	2.39
						Ty1/*Copia*	281	1.76
						BEL/Pao	0	0
						Retroviral	0	0
			LINEs	207	1.29	L1/CIN4	139	0.89
						RTE/Bov-B	68	0.40
			SINEs	26	0.05	Penelope	0	0
		DNA transposons	Tc1-IS630-Pogo	545	0.8			
			Tourist/Harbinger	33	0.13			
			hobo-Activator	13	0.07			

Repeat annotations on the 7DS WCS assemblies showed only 8.07 % of the sequences contained repeats. Of these, 5.52 % were retroelements (Table 4-10). Similar to the 7DS BAC assemblies, LTRs were the most prevalent retroelements in the 7DS WCS assembly contributing to 4.18 % of the assembly. The most common superfamily of retroelements in the 7DS WCS assembly was *Gypsy*/DIRS1 with 0.8 % of the assembly annotated. Among the DNA transposons annotated, transposons of the order Tc1-IS630-Pogo were the majority at 0.22 % (Table 4-10), similar to the 7DS BAC assembly.

Both the 7DS BAC and 7DS WCS assemblies showed a similar trend in the proportions of TE orders within the respective assemblies. The highest percentage of TEs in the 7DS BAC assembly were LTR retroelements specifically *Gypsy*/DIRS1 (29 %) followed by Ty1/*Copia* (13 %). This was also observed in the 7DS WCS assemblies were *Gypsy*/DIRS1 formed the highest percentage of repeats (2 %) followed by Ty1/*Copia* (2 %). In general, the 7DS WCS and BAC assemblies showed similar trends in the abundance of TE orders. TEs in decreasing abundance in the 7DS BAC assembly were LTRs > LINEs > SINEs > Tourist/Harbinger > hobo-Activator (Table 4-10). A similar trend was also observed for the 7DS WCS assembly. More repeats were however annotated on the 7DS BAC assemblies (59.59 %) compared to the 7DS WCS assemblies (8.07 %). Only the proportion of LTRs varied significantly (41.65 % vs 4.18 %) compared to other TE orders which remained below 2 % in the 7DS BAC compared to the 7DS WCS assembly.

4.4 Discussion

4.4.1 Chromosome arm 7DS annotation

The number of high quality genes annotated from the 7DS BAC assemblies (7, 485) was ~ 2.5 x more than the 3,014 high confidence genes reported for the 7DS isolated chromosome arm assembly by the international wheat genome sequencing consortium (IWGSC) (IWGSC *et al.*, 2014). The IWGSC assembled ~ 58% of the 7DS estimated size of 384 Mb, while the BAC assemblies constituted 239 Mb, consisting of non duplicate sequences. This was ~ 62 % of the estimated size of 7DS. The IWGSC's assemblies were generated from paired end short reads assembled with ABySS (Simpson *et al.*, 2009) and the CLC *de novo* assembler (http://www.clcdenovo.com). The IWGSC's assemblies were annotated using multi tissue RNA-Seq data, full length publically available cDNA's and reference proteins from *Brachypodium distachyon, Oryza sativa, Sorghum bicolor* and *Hordeum vulgare*. The 7DS BAC assemblies in contrast were assembled with the SASSY assembler and annotated using the MAKER annotation pipeline but with the same RNA-Seq data, and all full length publically available *Triticum* cDNA's.

While different gene annotation methods often result in varied numbers of gene calls, annotation quality filtering is often used to select high confidence gene calls. Both annotations were filtered for quality, with the IWGSC filtering based on high confidence gene calls (genes with similarity to related grass genes), while the BAC assemblies were based on AED scores < 1 (genes with supporting evidence). The selection of high quality annotations based on AED scores compared to annotation filtering based on homology to related grass proteins is better as filtering based on similarity to related genomes alone may result in real genes being assigned lower scores due to lack of homology to related grasses' annotations which may not necessarily be complete. Despite the varied approaches to annotation, its unlikely that annotation methods alone would result in ~ 2.5 x difference in number of genes called. This suggests that the nature of the assemblies and thus the assemblers used have contributed to the 2.5 x difference in gene numbers. The ABySS and CLC assemblers are de-Bruijn based assemblers. When assembling regions with complex repeats, de-Bruijn graphs result in bubbles representing numerous possible alternatives by which a region can be assembled. There are a variety of ways in which various implementations of de-Bruijn assemblers address this issue, which include selection of longest paths, selection of paths based on coverage or unique kmers within the repeat sequence. Such approaches may result in the generation of sequences that are not representative of the assembled genome.

Gene density on the 7DS BACs was 18.3 per Mb, higher than that reported for the wheat genome by the IWGSC with a range of 4.4 genes per Mb on chromosome 5AS to 10.4 per Mb on chromosome 2DL (IWGSC *et al.*, 2014). Based on the findings of the IWGSC, its likely that the wheat genome's gene content is much higher than the earlier estimate of between 32,000 - 38,000 for each diploid genome (IWGSC *et al.*, 2014).

An average gene length of 1,994 bps (± 2,761 bps) for the BAC assemblies is larger than the estimated average for the chromosome 3B BAC by BAC assembly at 1,095 bps (± 807 bps) (Choulet *et al.*, 2014). The BAC assembly average gene length was consistent with that reported for the IWGSC wheat assemblies at 2,008 bps (Choulet *et al.*, 2014). The average exon size of 297 bps for the BAC assemblies was also comparable to that of chromosome 3B BAC by BAC assembly at 307 bps (Choulet *et al.*, 2014).

The high percentage of genes mapping to *Aegilops tauschii,* (42.6%) followed by *Triticum urartu* (17.5 %) (Figure 4-2) is consistent with the evolutionary ancestry of wheat, specifically the wheat D genome as *Aegilops tauschii* and *Triticum urartu* the ancestral donors of the wheat D and A genome respectively (Cox, 1997; S., Huang *et al.*, 2002; Petersen *et al.*, 2006). The lower mapping percentage of the 7DS BAC genes to wheat at 7.9% suggests limited database entries for wheat, specifically wheat 7DS resources. Synteny between the wheat chromosome arm 7DS to *Hordeum vulgare, Brachypodium distachyon* and *Oryza sativa* was also reflected in the mappings (IWGSC *et al.*, 2014; Schmutz, Rokhsar, *et al.*, 2010; Berkman *et al.*, 2011). The consistency of the gene mappings onto published databases further validates the gene sequences of the 7DS BAC assemblies, although it is expected that if a higher number of wheat resources were available in the nr NCBI database , a greater percentage of the wheat genes would have mapped to wheat.

4.4.2 A comparison of gene annotations of chromosome arm 7DS WCS and BAC-by-BAC assemblies

The impact of assembly approaches on subsequent annotations was evidenced by comparisons of annotations of the 7DS BACs SASSY assemblies to those of 7DS WCS short reads using the Velvet assembler (Berkman *et al.*, 2011; Zerbino and Birney, 2008). More genes (7,485 vs 2, 422) were annotated on the 7DS BAC assemblies compared to the 7DS WCS assemblies as genic regions on the BAC assemblies were not as collapsed as those of the 7DS WCS. This resulted in better detection of genes by gene predictors and improved alignment of evidence during annotation. Gene and exon sizes were significantly larger in the BAC assemblies, suggesting that collapsing of repeats during assembly of the 7DS WCS resulted in the collapse of not only repeats but genic regions. The two assemblies differed only in the sequencing approach and assemblers used. Both assemblies were repeat masked and annotated with the same repeat databases and supporting evidence. The higher number of genes found on the BAC assemblies was expected as previous studies have shown that current published genome assemblies have erroneous estimates in gene counts (Denton *et al.*, 2014).

The gene density of the 7DS WCS assemblies was higher than that of BAC assemblies at 19.6 genes per Mb compared to 18.3 genes per Mb, as a result of a smaller and fragmented assembly. The AED values of the 7DS WCS assemblies were also significantly lower than those of the 7DS BAC annotations. If extra genomic features such as exons were suggested from the mapping of evidence as would be the case for alternative splicing, the AED score would be higher than if no alternate splicing was suggested from the mapping data. In such a case, a higher AED score would not necessarily indicate a lesser annotation quality, rather gene models that capture all the genomic features, in this case additional exons. In addition, assembled isoforms lengths were larger by ~ 487 bps on the 7DS BAC assemblies compared to the 7DS WCS assemblies, with the 7DS BAC assembly having fewer isoforms (26,370) compared to the 7DS WCS assembly with 75,278 isoforms. This may also explain why the AED scores of the WCS assemblies were significantly smaller than those of the BAC assemblies as the BAC assemblies may have captured more exons and fitting more genomic features in a gene model raised the AED score. This is reflected in the kernel density plot (Figure 4-7) that showed the density of AED scores, < 0.33 was higher in the WCS than the BAC assemblies but above 0.33, the inverse was observed. This is also evidenced by the average number of exons per gene being higher on the BAC assemblies (average 3.2 exons per gene with a median of 2) compared to the WCS assemblies (average 2.9 exons per gene with a median of 2) indicating more exons were detected on the 7DS assemblies and the 7DS WCS assemblies were more fragmented.

While genes and exons were significantly smaller in the WCS assemblies, the same was not observed with 5’ and 3’ UTRs. Comparing the Kernel density distributions of sizes of 5’ UTR and 3’UTR’s of the BAC and WCS showed that despite WCS UTR’s being larger than those of BACs, not only were more UTR’s detected in the BAC assemblies, but the UTR’s had a size distribution similar to those detected in the WCS but more centred towards the mean suggesting the UTR sizes among the BAC assemblies were more consistent than those of the WCS assemblies. These UTR’s may have been missed during the annotation of the 7DS WCS as a result of a fragmented assembly.

Reciprocal best BLAST hit analysis of shared genes between the BAC and WCS assemblies showed of the 7, 485 7DS BAC genes, 411 were also found in the 2,422 7DS WCS genes. The small number of shared genes suggests that sequence identity between the 7DS BAC and 7DS WCS genes varied greatly due to a collapse of sequences in the 7DS WCS genes. The variation number of genes absent but present in either assembly is as a result genome mis-assembly and local sequence variations caused by the assembly approaches and assemblers used. A similar observation was made in the assembly of the *Bos taurus* genome, where assembly improvements using the same data and assembler yielded a 40% variation in gene SNP calls with genes unique to both sets of assemblies being identified (Florea *et al.*, 2011). In addition, the BAC assemblies may have captured more genes as a result of a better reference and a more complete genome assembly. The WCS assemblies may also have captured regions not sequenced or assembled by the BAC assembly.

While reciprocal best BLAST hit analysis of 7, 485 7DS BAC genes against 2,422 7DS WCS genes identified only 411 shared genes, similar mapping percentages of the 7DS BAC genes and 7DS WCS genes on the nr database, 96.6% vs 97.1 % respectively validated the higher gene counts on the 7DS BAC assemblies as these were 3 x more than the 7DS WCS genes in number (7,485 vs 2,422). An equally higher number (3.5 x) of 7DS BAC genes did not have hits in the nr database compared to the 7DS WCS genes (251 vs 70 respectively).

The higher gene counts on the 7DS BAC assemblies may also be as a result of the percentage of the chromosome arm assembled by the 7DS WCS assemblies which was only 40 % (Berkman *et al.*, 2011).

As expected, the higher number of genes in the 7DS BAC assembly resulted in a higher number of HSPs matching on the 7DS BAC genes compared to the 7DS WCS genes (124,041 HSPs vs 40,076 HSPs respectively). A slightly higher percentage of the 7DS BAC genes' BLAST hits 55.3% (68,671 genes) corresponded to single HSPs compared to 52.8 % (21,171 genes) from the 7DS WCS assembly (Figure 4-1, Figure 4-4) further validating the better contiguity of the 7DS BAC genes compared to 7DS WCS genes.

The percentage of GO terms curated in sequence databases as a result of Electronic Annotation (IEA) at (96%) and Reviewed Computational Analysis (RCA) at 2% for both the 7DS BAC and 7DS WCS assemblies suggests limited experimentally derived GO's for wheat and plants in general. IEA are GO annotations inferred from sequence similarity to curated protein sequences without human curation (http://geneontology.org). RCA in contrast are GO annotations inferred computationally from large experimental datasets such as differential expression studies and protein protein interactions but are reviewed before addition onto public databases (http://geneontology.org). While computational analysis has greatly advanced crop genomics, more research and studies that yield experimental evidence for functional assignment are needed. These will lead to inference of gene function through evidence codes such as Experiment (EXP), Inferred from Direct Assay (IDA), Inferred from Physical Interaction (IPI), Inferred from Mutant Phenotype (IMP), Inferred from Genetic Interaction (IGI) and Inferred from Expression Pattern (IEP) (http://geneontology.org). Such evidence code categories are needed to further improve the accuracy and confidence of high throughput computational functional annotation pipelines. Currently, the UniprotKB database and the Gramene protein database were the most informative with 97 % and 2 % of all the GO terms retrieved for both the 7DS BAC and 7DS WCS genes. This shows the lack of highly curated plant specific resources for functional annotation inference.

Enrichment analysis of a GO term enables the identification of the probabilities of a set of genes annotated with a given GO term occurring by chance from a reference set of genes annotated with the same GO term. The enrichment of 29 GO terms in the 7DS BAC genes compared to the 7DS WCS genes and the enrichment of two GO terms in the 7DS WCS genes suggests that assembly approaches not only affect the contiguity, quality and number of genes annotated but also result in the enrichment/depletion of specific gene families.

For the enriched non redundant GO terms, only one GO term was enriched in the 7DS WCS gene sequences. This was **"chromatin binding"** (p value 2.76E-02, FDR 3.82E-03) (Table 4-9). The definition of the term **"chromatin binding"** with GO id: GO:0003682 on the gene ontology database (http://geneontology.org) is *"the interaction selectively or non selectively with chromatin, the network of fibers of DNA, protein, and sometimes RNA, that make up the chromosomes of the eukaryotic nucleus during interphase"*. 7DS BAC gene sequences associated with chromatin binding were 0.3% while 7DS WCS gene sequences associated with chromatin binding were 0.9 % (Figure 4-12).

Chromatin sequences are found on centromeres and telomeres of chromosomes and have been shown to contain large tandem repeats (Maluszynska and Heslop-Harrison, 1991; Murata *et al.*, 1994; Round *et al.*, 1997; Heslop-Harrison *et al.*, 1999; Copenhaver *et al.*, 1999). Centromeric regions have also been shown to contain genes (Qi *et al.*, 2010; B., Li *et al.*, 2013; Raats *et al.*, 2013; Palmer *et al.*, 1991; Dawe *et al.*, 1999; Henikoff *et al.*, 2001; Talbert *et al.*, 2002). It was expected that gene sequences associated with chromatin binding would be enriched in the 7DS BAC genes rather than the 7DS WCS as repeats are more collapsed in the 7DS WCS assemblies. This suggest other factors other than the assembly approach contributed to the enrichment of chromatin biding genes in the 7DS WCS assembly.

The 7DS WCS assembly was generated from random shearing followed by sequencing of isolated chromosome arm 7DS DNA (Berkman *et al.*, 2011) while the 7DS BAC assembly involved the generation of a minimum overlap of BAC clones, followed by BAC library construction (Safár *et al.*, 2010). The identification of BAC overlaps was based on the comparison of restriction fragment profiles using the SNaPshot HICF platform (Coulson *et al.*, 1986; Ding *et al.*, 2001; M., Luo *et al.*, 2003). The use of restriction profiles from restriction-digested genomic DNA has been shown to result in not only an under-representation of centromeric and telomeric regions as a result of problematic profiles of restriction sites in these regions (Riethman *et al.*, 2001; Mefford and Trask, 2002), but also in difficulties in cloning of AT rich regions as demonstrated in *Plasmodium falciparum* (Gardner *et al.*, 2002) and *Dictyostelium discoideum* (Eichinger *et al.*, 2005).

Protocols for BAC library preparation that do not use restriction enzymes have also been shown to capture previously missed regions that resulted from the use of restriction enzymes (Osoegawa *et al.*, 2007). More specific in wheat, HICF fingerprinting was shown to be unable to adequately detect clone overlaps in telomeric and centromeric regions due to uneven distribution of *Hind*III restriction sites in these regions (Raats *et al.*, 2013). Even though the number of genes associated with **"chromatin binding"** was low, 0.3 % in the 7DS BAC genes and 0.9 % in the 7DS WCS genes (Figure 4-12), an under representation of **"chromatin binding"** genes in an assembly has implications in downstream analysis as the low percentages is a reflection on the percentage of genes in the cell associated with **"chromatin binding"** and not the extend of enrichment. Genes associated with **"chromatin binding"** are also involved in growth inhibition, apoptosis (Mancinelli *et al.*, 2009) and in suppression of recombination in wheat (Corredor *et al.*, 2007), making their complete assembly and annotation important.

The GO cellular component enriched terms **"Golgi apparatus"** (p value 4.67E-07, FDR 7.30E-09) and **"plastid"** (p value 1.63E-02, FDR 1.95E-03) were found to be enriched in the 7DS BAC genes (Table 4-9). The definition of the term **"Golgi apparatus"** GO id: GO:0005794 on the gene ontology database (http://geneontology.org) is *"A compound membranous cytoplasmic organelle of eukaryotic cells, consisting of flattened, ribosome-free vesicles arranged in a more or less regular stack."* The definition of the term **"plastid"** GO id: GO:0009536 on the gene ontology database (http://geneontology.org) is *"Any member of a family of organelles found in the cytoplasm of plants and some protists, which are membrane-bounded and contain DNA. Plant plastids develop from a common type, the proplastid."* Both the Golgi apparatus and plastids are cytoplasmic organelles, with plastid covering chloroplasts, chromoplasts and leucoplasts. The Golgi apparatus is encoded in the nucleus while some plastids (chloroplasts) contain their own DNA. 7DS BAC gene sequences associated with Golgi apparatus were 3.5% while 7DS WCS gene sequences associated with Golgi apparatus were 0.9 % (Figure 4-12). 7DS BAC gene sequences associated with plastids were 9.8 % while 7DS WCS gene sequences associated with plastids were 7.2 % (Figure 4-12).

It is established that organelles occur in multiple copies within a cell, between 11 - 70 and the number of copies vary depending on the developmental stage and physiological state of the cell (Tymms *et al.*, 1983) with chloroplasts representing 17% - 23 % of the total DNA of a plant cell (Boffey and Leech, 1982). While mitochondrial and chloroplast DNA contamination of plant DNA extraction has been shown to be high , between 3 - 5% (Noir *et al.*, 2004; Moullet *et al.*, 1999), protocols used for the the preparation of 7DS BAC clones showed there was no mitochondrial or chloroplast contamination of BAC clones (Šimková *et al.*, 2003; Safár *et al.*, 2004). Thus prior to assembly, chloroplast wheat reference genomes were not filtered from the 7DS WCS and BAC reads. The *E coli* and the cloning vector pBACIndigo reference sequences were filter off from the BAC sequenced pool reads. No gene sequences in both the 7DS BAC genes and the 7DS WCS genes mapped to *E coli* genes from the BLASTX mapping results. The chloroplast genome is between 115 and 165 Kb long (Jansen *et al.*, 2006) with two inverted repeat sequences of approximately 20 kb where extensive rearrangement occurs (Ogihara *et al.*, 2002; Saski *et al.*, 2007). These inverted repeat regions have been shown to contain up to 18 duplicated genes in *H. vulgare*, *Sorghum bicolor* and *A. stolonifera* (Saski *et al.*, 2007). The integration of chloroplast gene sequences into the nuclear genome in plant cells is rapid, continuous and has been widely demonstrated (Scott and Timmis, 1984; Ayliffe and Timmis, 1992; C., Y., Huang *et al.*, 2003; Shahmuradov *et al.*, 2003; Stegemann *et al.*, 2003; Matsuo *et al.*, 2005) with such sequences often referred to as promiscuous DNAs (Ellis, 1982). The mechanism of transfer is not well understood but movements similar to transposable elements (Stegemann *et al.*, 2003) and chromosome double-strand break (DSB) repair mechanisms (Ricchetti *et al.*, 1999; X., Yu and Gabriel, 1999) have been proposed. In rice, the integration of chloroplast genes into the nuclear genome has been shown to be scattered throughout rice chromosomes with larger fragments integrating preferentially to repeat rich areas such as peri-centromeric regions of chromosomes. Peri-centromeric regions of chromosomes contain genes (Qi *et al.*, 2010; B., Li *et al.*, 2013; Raats *et al.*, 2013; Palmer *et al.*, 1991; Dawe *et al.*, 1999; Henikoff *et al.*, 2001; Talbert *et al.*, 2002) and large tandem repeats (Maluszynska and Heslop-Harrison, 1991; Murata *et al.*, 1994; Round *et al.*, 1997; Heslop-Harrison *et al.*, 1999; Copenhaver *et al.*, 1999). Given gene sequences within repeat regions would be more collapsed in the 7DS WCS assemblies compared to the 7DS BAC assemblies may explain the enrichment of gene sequences associated with plastid in the 7DS BAC genes as possible chloroplast genes in the wheat genome would not have been purified out during library preparation (due to integration into the genome) and were thus assembled in both the WCS and BAC assemblies. The preferential localization of genes from the chloroplast genome in repeat rich areas would result in these genes being more collapsed in the WCS assemblies but over represented in the 7DS BAC assemblies.

The Golgi apparatus was an enriched term in the 7DS BAC genes with a p value of 4.67E-07. The Golgi apparatus in a cell is responsible for processing, sorting and packaging proteins and lipids into vesicles for onward transport to other parts of the eukaryotic cell. The Golgi apparatus also referred to as Golgi complex consists of a stack of membranous cisternae (Barr and G., Warren, 1996). Proteins of the membranous cisternae that form the Golgi apparatus have common coiled-coil motifs which appear as rod-like structures in the cell and are often referred to as golgins (Burkhard *et al.*, 2001; Gillingham and Munro, 2003). The coiled-coil motifs of golgins have a common protein domain, the GRIP domain (Munro and Nichols, 1999; D., L., Brown *et al.*, 2001; Kjer-Nielsen *et al.*, 1999; Barr, 1999). The sequence of the GRIP domain proteins contain 75 % - 85% of seven long α helical amino acid repeats (Luke *et al.*, 2005; J., H., Brown *et al.*, 1996; Gruber *et al.*, 2006; Strelkov and Burkhard, 2002; Crick, 1952; Crick, 1953; Gruber and Lupas, 2003). Such a repeat would be fragmented or collapsed in the 7DS WCS assemblies, leading to an under representation of golgin genes in the subsequent annotations, and an enrichment in the 7DS BAC assemblies.

The number of genes encoding coiled-coil proteins in a eukaryotic cell has been estimated to be about 5% of all coding sequences, implying their involvement in numerous cell processes and functions (Wolf *et al.*, 1997; Newman *et al.*, 2000; Gillingham and Munro, 2003). This agrees with the 7DS BAC gene sequences associated with Golgi apparatus were 3.4% terms (Figure 4-12) as the initial estimate of ~ 5 % was determined in *Saccharomyces cerevisiae* which has a more compact genome compared to wheat and thus probably has a slightly higher percentage. The role of the Golgi apparatus in a cell is very central. Its involvement in the modification, packaging and sorting of secreted and membrane lipids and proteins destined for transport to other cellular regions explains why an enrichment of genes associated with the Golgi apparatus would result in the enrichment of terms associated with functions of the Golgi apparatus.

All the other 7DS BAC gene enriched GO terms are associated with functions of the Golgi apparatus. These GO terms were **"transferase activity"**, **"enzyme regulator activity"**, **"carbohydrate binding"**, **"cellular protein modification process"**, **"catabolic process"** and **"multicellular organismal development"**. The identification of "**carbohydrate binding", "cellular protein modification process"** and **"transferase activity"** in the Golgi apparatus is well established. Specific examples such as the oligomeric Golgi complex (COG) genes, COG5 and COG7, that are involved in glycosylation , a process of protein modification that involves the attachment of carbohydrates onto proteins and fats mediated by enzymes that are broadly categorized as transferases (Cailotto and Lories, 2013) (X., Wu *et al.*, 2004; Marquardt and Denecke, 2003) (Rymen *et al.*, 2015; J., Y., Ha *et al.*, 2014; Nikolovski *et al.*, 2012).The central role of the Golgi complex genes in a cell is further supported by the percentage of genes associated with **"transferase activity"** in the 7DS BAC genes at 27.2% compared to the 7DS WCS genes at 22.2 % , the highest of all the enriched GO terms. This was followed by "**cellular protein modification process"** with 14.1% of the 7DS BAC genes and 10.1% of the 7DS WCS genes associated (Figure 4-12).

The Golgi apparatus has also been directly linked to the control of membrane trafficking through enzymatic regulation and modification of lipid and phospholipids. Published examples include regulation and modification of lipid, phospholipids and proteins using phospholipase A (PLA) and lysophospholipid acyltransferase (LPAT) (K., D., Ha *et al.*, 2012), platelet-activating factor acetylhydrolase (PAFAH) (Bechler *et al.*, 2010) and lysophosphatidic acid-specific acyltransferase LPAAT3 (Schmidt and W., J., Brown, 2009) enzymes. Catabolic pathways involving in death and degradation of cell and cellular components have also been directly linked to the Golgi apparatus. Specific examples include neuronal differentiation and aging (Aureli *et al.*, 2014; H., J., Kim *et al.*, 2014) and macro autophagy a proteins catabolic degradation pathway in the cell (Yen *et al.*, 2010; J., Wang *et al.*, 2013). Multicellular organismal development processes have also been localised on the Golgi complex such as plant growth, cell wall and xylem synthesis (Gardiner and Chrispeels, 1975; Voiniciuc *et al.*, 2015; Parsons *et al.*, 2012) (Baocai Zhang *et al.*, 2011) and lipid biosynthesis (Binnington *et al.*, 2016; Ridgway, 2016).

While there appears to be no published literature on the effect of assembly approaches on enrichment or depletion of specific gene families, the results presented in this thesis suggest this is the case. The gene families and cellular component genes enriched in this analysis and their central role in all cell functions and processes shows how significant an assembly approach influences downstream analysis and characterization of a genome. Due to limited plant annotation resources in public databases, 39.1% (2,929) of 7DS BAC genes and 38.1% (925) of 7DS WCS genes were not functionally annotated. A full functional annotation with existing current available public resources is not yet possible but would probably reveal more enriched gene families.

The use of GO terms allows uniform transfer and inference of gene functional annotation across datasets and species, but there are limitations posed by the use of GO terms for accurate annotation. GO inference is reliant on accurate orthologue inference by sequence similarity. Incorrect orthologue assignment results in erroneous functional assignment. Such errors can arise through accurate identification of an orthologue whose function has diverged in the organism being annotated, the identification of an orthologue from an erroneous database match or as a result of the absence of sequences from a true orthologue in the database, thus erroneous functional annotations are inferred from a non optimal database match. Such errors can be reduced by inspection of annotation evidence and sequence matches, but this is not practical when annotating whole genomes where high throughput computational pipelines are used. Limiting the annotation to closely related species may reduce the effect of erroneous annotations, but for plant annotations, this is limiting due to the lack of highly curated plant resources compared to other species. When resources are limited, extending the species used in database matches may offer insights where non would otherwise be available from plant specific resources. For projects or analysis on specific genes, confirmatory analysis of the inferred functions should be done.

4.4.3 A comparison of repeat annotations on chromosome arm 7DS WCS and 7DS BAC-by-BAC assemblies

Repeats constituted 49 % of the 7DS BAC assembly. This was lower than the expected ~80% repeat content of wheat (Wicker *et al.*, 2011). This supports previous findings that showed there is varied repeat content across the bread wheat or sub genomes. The wheat D sub genome was shown to have fewer LTRs compared to the A and B sub genomes (IWGSC *et al.*, 2014). *Gypsy* and *Copia*-like elements were the most abundant in the 7DS BAC and WCS assemblies, similar to other findings in wheat's A genome progenitor *T. urartu* with 49% (Ling *et al.*, 2013), chromosome 5D of *A. tauschii* with 56.11% (Akpinar *et al.*, 2014), maize with 50 - 85% (SanMiguel *et al.*, 1996; Schnable *et al.*, 2009), and sorghum with 61% (Paterson *et al.*, 2009).

The 7DS BAC and 7DS WCS assemblies both showed a similar trend in the proportions of TE orders with only LTRs varying significantly (41.65% vs 4.18%) compared to other TE orders which remained below 2 %. LTRs have an average length of 8 Kb (9.7 Kb for *Gypsy*-like elements and 6.3 Kb for *Copia*-like elements). An average length 16 fold longer than the PE read insert size, coupled with low sequence complexity would pose a great challenge in assembly resulting in collapsing of LTR sequences during assembly. A similar observation was made during a comparison of repeats in sequence reads and assembled contigs of all wheat chromosome based WGS assemblies where a reduced representation of *Gypsy* LTRs were observed in the assembly compared to raw reads, from 81 % to 76 % (IWGSC *et al.*, 2014).

This results suggest SASSY assembled LTRs better compared to the Velvet assembler that was used in assembling the 7DS WGS assembly, enabling more LTRs to be annotated. Accurate annotation of LTR is importartant in crop genomes as LTRs (specifically *WIS2*, *Wilma*, *Daniela*, and *Fatima*) have been shown to shape the genome of *A. speltoides* in response to stressful environmental conditions (Hosid *et al.*, 2012). In addition, LTRs from the *Gypsy*-like family have been successfully used as genetic markers in wheat (Melnikova *et al.*, 2011) and have been associated with genome reorganization in wild wheats (Senerchia *et al.*, 2015). Accurate assembly of repeats is thus important as repeats have the potential of providing high density markers and insights into genome evolution due to their abundance in wheat. An unbiased annotation of existing and novel repeats would require *de-novo* identification of repeats from raw reads prior to assembly.

4.5 Summary

BAC-by-BAC assemblies with SASSY resulted in more genes, more exons, larger isoforms and more UTR regions being detected. This shows the choice of an assembly approach not only affects the completeness of the resulting assembly, but also the detection of genes and alternative splicing events. The results presented in this chapter have shown not only are more and longer genes annotated from BAC assemblies with SASSY, but specific gene families are captured that are otherwise depleted when using other approaches. This has significant implications in previous and current genome sequencing projects affecting the identification of markers for crop improvement. Without the use of accurate and validated assembly and annotation approaches, post assembly analysis is likely to proceed with a high proportion of genes missing, yielding incomplete or erroneous results.

5 Discussion and future directions

5.1 Discussion

Crop improvement is vital for sustainable food security as world populations continue to expand. Crop yield improvements in world staple foods such as wheat, rice and maize offer the best opportunity for improved food security, especially in developing countries where crop production systems are not mechanized resulting in under production. Genome complexity of crops such as wheat still poses challenges in sequencing and assembly slowing crop improvement efforts. From the analysis and results presented in this thesis, the sequencing and assembly of large repetitive polyploidy genomes to high accuracy is possible, though still challenging.

Advancements in single molecule-sequencing technologies promise longer reads (10kbp – 4.3 Mb) with fewer sequencing errors, GC bias and reduced sequencing costs in the near future (Lee *et al.*, 2014; Berlin *et al.*, 2015). Currently single molecule-sequencing platforms under development offer faster, easy to use platforms with minimal library preparation steps. These include Oxford Nanopore technologies (www.nanoporetech.com) and Single-molecule real-time (SMRT) sequencing, developed by Pacific Biosciences (PacBio) (http://www.pacb.com). The move to single molecule sequencing has similarly led to advancements in optical map technologies such as the Irys^{Tm} system from BioNano genomics (http://www.bionanogenomics.com). While there is a shift towards single molecule technologies, short read NGS technologies will continue to play a significant role in the sequencing and assembly of genomes.

At the time of this thesis writing, the application of multiple approaches for the sequencing and assembly of complex genomes, namely the use of BAC-by-BAC approaches, combined with chromosome isolation provided complexity reduction enabling the accurate and robust sequencing and assembly of bread wheat chromosome arm 7DS. Long range Nextera^{TM} mate pair libraries and optical maps enabled scaffolding and validation of resulting assemblies to high accuracy. While rapid advances in computing infrastructure and sequencing platforms will offer better, faster and cheaper methods for genome sequencing and assembly in the near future, improvements to the sequencing and assembly approach presented in this thesis can be made and applied to immediate upcoming sequencing projects.

Results from the assembly of BAC pools using paired end Illumina reads showed the need for scaffolding of resulting assemblies using mate pair data. As the cost of sequencing reduces and advances in automation of library preparation steps continue, it will become viable to prepare and sequence mate pair datasets to high coverage and use these for assembly with SASSY instead of using paired end libraries. The SASSY assembler already supports mate pair datasets.

The use of genome assemblies with optical maps has the potential to greatly advance genome assembly approaches. Already attempts at integrating BioNano optical maps with single molecule sequencing in the scaffolding of the human genome have showed great success (Pendleton *et al.*, 2015; Usher *et al.*, 2015; English *et al.*, 2015). For complex polyploid genomes, no attempts have been published as of the writing of this thesis, but integration of optical maps with sequence data for *de novo* assembly is now possible. While only 48% (Table 2-11) and 49% (Table 3-11) of the BAC scaffolds could be anchored onto the 7DS MTP and optical map respectively, these percentages were consistent with previous studies where a low density of markers or lack of an even distribution resulted in significant regions in assemblies remaining unanchored. In such studies (S., Liu *et al.*, 2014; X., Wang *et al.*, 2011; Varshney *et al.*, 2012; Varshney *et al.*, 2013; Schmutz *et al.*, 2014; Kang *et al.*, 2015; Kang *et al.*, 2014; Dohm *et al.*, 2014; K., Wang *et al.*, 2012), only 30 % – 60% of the assembled genomes could be anchored into pseudo molecules (Sakai *et al.*, 2015).

Current published and finished genome assemblies have been shown to be incomplete with mis-assemblies (Alkan *et al.*, 2010). This thesis has shown that not only are published genome assemblies not subjected to vigorous validation, but the choice of the assembly and annotation approaches greatly impacts on the number of genes detected and annotated. The primary goal of sequencing a genome is to capture the full set of genes within that organism. With improvements in sequencing and assembly technologies, there will be a significant increase in the number of genes estimated as well.

In particular, the choice of a sequencing and assembly approach has a significant impact on the structure of genic features. This thesis showed that WCS assemblies with Velvet led to significantly smaller exons, genes and isoforms when compared to BAC-by-BAC assemblies with SASSY. While the fraction of the chromosome arm assembled using the WCS approach differed from the BAC assembly, such a difference should not have led to a statistically significant difference in gene, exon and isoform sizes. For highly conserved genic features such as 5' and 3' UTR's, the BAC-by-BAC assemblies were significantly smaller with a smaller standard deviation than the WCS. This suggests that accurate assemblies can be used to infer the distribution, sizes and structural conservation of genomic features.

Current assembly algorithms are still not able to assemble genomic regions correctly across repeat rich regions leaving assemblies fragmented requiring scaffolding post assembly. Repeats confound assembly and have similarly been shown to complicate gene finding and estimation of gene numbers, especially where genes have been moved, duplicated or fragmented by transposon activity (N., Jiang *et al.*, 2004; J., Lai *et al.*, 2005; Morgante *et al.*, 2005; Paterson *et al.*, 2009). Long reads such as those generated by the PacBio sequencer will enable sequencing across such regions, enabling accurate and complete assembly of repeats and similarly genes within transposon sequences.

The use of experimental evidence to validate structural annotations is limited by the availability and suitability of such evidence. Species specific RNA-seq datasets to that of the assembled genome currently offer the best high throughput evidence for supporting automated annotation. While the generation of RNA-seq datasets from multiple organs at different developmental stages is primarily to study differential gene expression, it also provides unprecedented opportunities for identifying and validating novel genes and splice variants when used for annotation. As no current study has involved a complete capture of expressed transcripts using RNA-seq of all organs at all major developmental stages in any plant, it remains to be seen what such a study may reveal. Its most likely that more novel splice variants would be detected and not a major increase in genes counts for characterized genomes.

Annotation results of both the WCS and BAC assemblies showed that all identified gene models had supporting evidence. This suggests annotation evidence was exhaustive and that a thorough evaluation and improvement of ab initio gene finding tools may provide more candidate genes for validation using RNA-seq datasets. The correct use of RNA-seq datasets requires that the full genome reference is available for mapping prior to annotation. For the WCS and BAC 7DS assemblies, all other wheat arm assemblies were used. The use of an incomplete or poor reference may result in non optimal mapping of RNA-seq datasets leading to false positive validation of gene models. While the use of a full genome reference has been accounted for in the WCS and BAC 7DS annotations, the 7DS WCS assemblies with the velvet assembler showed the assemblies were fragmented and collapsed. This may contribute to sub optimal mapping of RNA-seq data given all other wheat arms except 3B were assembled in a similar manner to the 7DS WCS assemblies.

Collapsing repeats during assembly affects the detection of genes in repeat rich areas, affects structural and functional annotation of genes identified and the corresponding gene families with protein domains containing amino acid repeat sequences as demonstrated from the gene enrichment analysis were coiled-coil and golgin gene families were enriched on the 7DS BAC genes compared to the 7DS WCS genes. The only enriched gene family on the 7DS WCS genes was chromatin binding, but this was as a result of a difference in the library preparation protocols between BAC by BAC approaches and WCS approaches. Although not all 7DS BAC and 7DS WCS genes were functionally annotated, the number of genes functionally annotated in the 7DS BAC and 7DS WCS genes provided a statistically sufficient sample size on which enrichment analysis was done. Therefore, when more public curated databases resources are available to enable additional functional annotation from database matches, the enrichment analysis would still show the same enriched/depleted gene families as currently found.

GO enrichment analysis is limited by the number of GO terms available and by the nature of GO enrichment statistical tests. Enrichment analysis tests individual GO terms independently between a gene list of interest and a background list. A p value for each GO term is determined from the frequency of the GO term in the gene list and in the background. The selection of enriched terms is based on a p value cut-off of 0.05. A shortfall of enrichment analysis is that a GO term is considered independent of other GO terms, even though the terms are hierarchically related in the GO and may be biologically related due to possible protein protein interactions. The treatment of GO terms as independent entities in enrichment analysis results in redundancy in enriched GO terms (Khatri and Drăghici, 2005; D., W., Huang *et al.*, 2009; Yon Rhee *et al.*, 2008).

While a fully functionally annotated gene list was not possible, the reduction of redundancy of enriched GO terms was possible by retaining lower level GO terms where multiple terms were present within the same hierarchy. This enabled a more detailed view of enriched GO terms and enabled further inference of enriched GO terms based on literature searches, providing insight in the interaction between GO terms or gene families. This led to the identification of the Golgi complex and its associated genes and gene functions as being a central point of difference between the 7DS BAC and 7DS WCS gene families.

5.2 Future directions

As the cost of genome sequencing continues to drop and single molecule sequencing platforms become available, longer and more accurate reads will become the standard in genome sequencing projects. Current improvements in single molecule sequencing technologies have resulted in longer reads with a reduced error rate. One such platform is Single-molecule real-time (SMRT) sequencing developed by Pacific BioSciences (PacBio). The impact of long reads in plant genomics is already evident.

While long range MP reads can span repeats, assemblies using short reads have been shown to be fragmented with missing genes (Alkan *et al.*, 2010; Denton *et al.*, 2014). PacBio reads appear to have addressed this limitation. For example, long PacBio reads have enabled the assembly of 100 times longer contigs with 100 times fewer gaps compared to SGS-based assemblies of the Azuki bean (*Vigna angularis*) (Sakai *et al.*, 2015). In addition, thousands of genes that were previously missing or fragmented were also detected. The sequencing and assembly of the desiccation tolerant grass *Oropetium thomaeum* produced not only a draft genome assembly with no gaps, but also included transposable elements, centromeres and telomeres that were missing in previous assemblies (VanBuren *et al.*, 2015; Kellogg, 2015). A comprehensive identification of novel circular RNAs in rice was also possible as a result of long read transcripts sequenced using PacBio (Lu *et al.*, 2015). Long read PacBio reads are being adopted for sequencing of complex repetitive genomes. Current published analysis in wheat has shown successful use of PacBio reads in transcriptome sequencing (Wei Zhang *et al.*, 2014) with more applications in genome sequencing in future.

Genome assembly algorithms and data structures are similarly undergoing active development although assembly algorithms are generally developed and tested on non-plant species (Metzker, 2010) with the exception of the SASSY assembler. Advancements in algorithm development include algorithms based on probabilistic models. Current assembly algorithms are based on overlap layout consensus approaches, de Bruijn graphs or combinations of both. Assemblers based on probabilistic models such as maximum likelihood e.g. the GAML assembler (Boža *et al.*, 2015), the Computing Genome Assembly Likelihood (CGAL) assembler (Rahman and Pachter, 2012) and the Genome Assembly by Bayesian Inference (GABI) assembler (Howison *et al.*, 2013) have been shown to handle sequencing errors, variations in insert size distributions and repeat structures in sequenced datasets better or comparable to de Bruijn graphs and overlap layout consensus algorithms. However, the performance of probabilistic assemblers in assembling complex polyploidy genomes remains to be seen.

Assembly and mapping of NGS reads of large complex genomes still poses computing challenges. While there are advancements in NGS analysis algorithms as discussed above, computing hardware developments have lagged behind. However, recent advancements have focused on the adoption of graphic processing units (GPUs) and specialized processors such as DRAGEN (www.edicogenome.com/dragen/). These provide processing units with more cores than typical central processing units (CPUs) used on desktop and high performance computing clusters. The need for faster computing in NGS analysis is essential for the timely analysis of the massive data generated by sequencing platforms. This will be the bottle neck in future for comparative genomics of large complex genomes. For example, the largest plant genome assembled as of the writing of this thesis was that of the loblolly pine genome (22 Gb) (Zimin *et al.*, 2014). The assembly of such a large genome was only possible by filtering off redundant reads prior to assembly which resulted in loss of sequence information. The assembly, scaffolding and read mapping results in wheat reported in this thesis would have also been greatly accelerated if such computing facilities were available. Newer algorithms that run on GPUs also continue to be developed such as MaxSSmap (Turki and Roshan, 2014) and SW# (Korpar and Šikic, 2013) that exploit advantages of GPU's to deliver faster mapping of reads to a reference. There are also active modifications of common NGS tools for use on GPUs. Examples include G-BLASTN (Zhao and Chu, 2014), CUDA-BLASTP (W., Liu *et al.*, 2011), GPUBLAST(Vouzis and Sahinidis, 2011) and SOAP3 (C., Liu *et al.*, 2012).

In summary, future improvements in sequencing technologies, analysis algorithms and hardware promise to accelerate genomics. For plant genomics, such advancements will only enable faster identification of important agronomic traits leading to quicker improvements in crop yields through development of improved crop varieties.

References

Adey, A., Asan, A., Morrison, H.G., et al. (2010) Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology.*, **11**, R119.

Adhikari, T.B., Cavaletto, J.R., Dubcovsky, J., Gieco, J.O., Schlatter, A.R. and Goodwin, S.B. (2004) Molecular mapping of the Stb4 gene for resistance to *Septoria tritici* blotch in wheat. *Phytopathology.*, **94**, 1198–1206.

Akhunov, E.D., Akhunova, A.R. and Dvořák, J. (2005) BAC libraries of *Triticum urartu*, *Aegilops speltoides* and *Aegilops tauschii*, the diploid ancestors of polyploid wheat. *Theoretical and Applied Genetics.*, **111**, 1617–1622.

Akpinar, B.A., Lucas, S.J., Vrána, J., Doležel, J. and Budak, H. (2014) Sequencing chromosome 5D of *Aegilops tauschii* and comparison with its allopolyploid descendant bread wheat (*Triticum aestivum*). *Plant Biotechnology Journal.*, **13**, 740–752.

Alkan, C., Sajjadian, S. and Eichler, E.E. (2010) Limitations of next-generation genome sequence assembly. *Nature Methods.*, **8**, 61–65.

Altenhoff, A.M. and Dessimoz, C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. J. A. Eisen, ed. *PLoS Computational Biology.*, **5**, e1000262.

Altenhoff, A.M., kunca, N., Glover, N., et al. (2015) The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Research.*, **43**, D240–D249.

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

An, D., Zheng, Q., Zhou, Y., et al. (2013) Molecular cytogenetic characterization of a new wheat–rye 4R chromosome translocation line resistant to powdery mildew. *Chromosome Research.*, **21**, 419–432.

Anioł, A.M. (1995) Physiological aspects of aluminium tolerance associated with the long arm of chromosome 2D of the wheat (*Triticum aestivum L.*) genome. *Theoretical and Applied Genetics.*, **91**, 510–516.

Anon (2012) A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics.*, **13**, 329–342.

Apotikar, D.B., Venkateswarlu, D., Ghorade, R.B., Wadaskar, R.M., Patil, J.V. and Kulwal, P.L. (2011) Mapping of shoot fly tolerance loci in sorghum using SSR markers. *Journal of Genetics.*, **90**, 59–66.

Appels, R., Nystrom, J., Webster, H. and Keeble-Gagnère, G. (2015) Discoveries and advances in plant and animal genomics. *Functional & Integrative Genomics.*, **15**, 121–129.

Ariyaratne, P.N. and Sung, W.K. (2011) PE-Assembler: *de novo* assembler using short paired-end reads. *Bioinformatics.*, **27**, 167–174.

Arraiano, L.S., Worland, A.J., Ellerbrook, C. and Brown, J.K.M. (2001) Chromosomal location of a gene for resistance to septoria tritici blotch *(Mycosphaerella graminicola)* in the hexaploid wheat “Synthetic 6x.” *Theoretical and Applied Genetics., **103**, 758–764.*

Audano, P. and Vannberg, F. (2014) KAnalyze: a fast versatile pipelined K-mer toolkit. *Bioinformatics.*, **30**, 2070–2072.

Aureli, M., Samarani, M., Loberto, N., Bassi, R., Murdica, V., Prioni, S., Prinetti, A. and Sonnino, S. (2014) The glycosphingolipid hydrolases in the central nervous system. *Molecular Neurobiology.*, **50**, 76–87.

Ayliffe, M.A. and Timmis, J.N. (1992) Plastid DNA sequence homologies in the tobacco nuclear genome. *Molecular & General Genetics.*, **236**, 105–112.

Bairoch, A. (2008) UniProtKB/Swiss-Prot: new and future developments. In Berlin, Heidelberg: Springer-Verlag, pp. 204–206.

Bairoch, A., Boeckmann, B., Ferro, S. and Gasteiger, E. (2004) Swiss-Prot: juggling between evolution and stability. *Briefings in Bioinformatics.*, **5**, 39–55.

Bao, E., Jiang, T. and Girke, T. (2014) AlignGraph: algorithm for secondary *de novo* genome assembly guided by closely related references. *Bioinformatics.*, **30**, i319–i328.

Bao, W., Kojima, K.K. and Kohany, O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA.*, **6**, 11.

Bariana, H.S. and McIntosh, R.A. (1993) Cytogenetic studies in wheat. XV. Location of rust resistance genes in VPM1 and their genetic linkage with other disease resistance genes in chromosome 2A. *Genome.*, **36**, 476–482.

Barr, F. and Warren, G. (1996) Disassembly and reassembly of the golgi apparatus. *Seminars in Cell & Developmental Biology.*, **7**, 505–510.

Barr, F.A. (1999) A novel Rab6-interacting domain defines a family of golgi-targeted coiled-coil proteins. *Current Biology.*, **9**, 381–384.

Batzoglou, S. (2002) ARACHNE: a whole-genome shotgun assembler. *Genome Research.*, **12**, 177–189.

Baum, B.R., Edwards, T. and Johnson, D.A. (2009) Phylogenetic relationships among diploid *Aegilops* species inferred from 5S rDNA units. *Molecular Phylogenetics and Evolution.*, **53**, 34–44.

Bechler, M.E., Doody, A.M., Racoosin, E., Lin, L., Lee, K.H. and Brown, W.J. (2010) The phospholipase complex PAFAH Ib regulates the functional organization of the golgi complex. *The Journal of Cell Biology.*, **190**, 45–53.

Bennett, M.D. and Smith, J.B. (1991) Nuclear DNA amounts in angiosperms. *Philosophical Transactions: Biological Sciences.*, **334**, 309–345.

Bennetzen, J.L., Coleman, C., Liu, R., Ma, J. and Ramakrishna, W. (2004) Consistent over-estimation of gene number in complex plant genomes. *Current Opinions in Plant Biology.*, **7**, 732–736.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature.*, **456**, 53–59.

Berkman, P.J., Skarshewski, A., Lorenc, M.T., et al. (2011) 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., Manoli, S., et al. (2012) 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., et al. (2013) Dispersion and domestication shaped the genome of bread wheat. *Plant Biotechnology Journal.*, **11**, 564–571.

Berlin, K., Koren, S., Chin, C.-S., Drake, J.P., Landolin, J.M. and Phillippy, A.M. (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology.*, **33**, 623–630.

Binnington, B., Nguyen, L., Kamani, M., Hossain, D., Marks, D.L., Budani, M. and Lingwood, C.A. (2016) Inhibition of Rab prenylation by statins induces cellular glycosphingolipid remodeling. *Glycobiology.*, **26**, 166–180.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. and Pirovano, W. (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics.*, **27**, 578–579.

Boffey, S.A. and Leech, R.M. (1982) Chloroplast DNA levels and the control of chloroplast division in light-grown wheat leaves. *Plant Physiology.*, **69**, 1387–1391.

Boisvert, S., Laviolette, F. and Corbeil, J. (2010) Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of Computational Biology.*, **17**, 1519–

1533.

Bolser, D. (2014) *Mouse Genome Informatics (MGI), Mouse Genome Database (MGD)*, Chichester, UK: John Wiley & Sons, Ltd.

Bolser, D.M., Kerhornou, A., Walts, B. and Kersey, P. (2014) *Triticeae* resources in Ensembl Plants. *Plant and Cell Physiology.*, **56**, e3–e3.

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

Bowers, J.E., Chapman, B.A., Rong, J. and Paterson, A.H. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature.*, **422**, 433–438.

Boža, V., Brejová, B. and Vinar, T. (2015) GAML: genome assembly by maximum likelihood. *Algorithms for Molecular Biology.*, **10**, 18.

Bradnam, K.R., Fass, J.N., Alexandrov, A., et al. (2013) Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *Gigascience.*, **2**, 10.

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

Brown, D.L., Heimann, K., Lock, J., Kjer-Nielsen, L., van Vliet, C., Stow, J.L. and Gleeson, P.A. (2001) The GRIP domain is a specific targeting sequence for a population of trans-golgi network derived tubulo-vesicular carriers. *Traffic.*, **2**, 336–344.

Brown, J.H., Cohen, C. and Parry, D.A. (1996) Heptad breaks in alpha-helical coiled coils: stutters and stammers. *Proteins.*, **26**, 134–145.

Buermans, H.P.J. and Dunnen, den, J.T. (2014) Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta.*, **1842**, 1932–1941.

Burge, S., Kelly, E., Lonsdale, D., et al. (2012) Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. *Database.*, **2012**, bar068–bar068.

Burke, D., Carle, G. and Olson, M. (1987) Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science.*, **236**, 806–812.

Burkhard, P., Stetefeld, J. and Strelkov, S.V. (2001) Coiled coils: a highly versatile protein folding motif. *Trends in Cell Biology.*, **11**, 82–88.

Cabral, A.L., Jordan, M.C., McCartney, C.A., You, F.M., Humphreys, D.G., MacLachlan, R. and Pozniak, C.J. (2014) Identification of candidate genes, regions and markers for pre-harvest sprouting resistance in wheat (*Triticum aestivum L.*). *BMC Plant Biology.*, **14**, 340.

Cailotto, F. and Lories, R.J. (2013) Inhibition of COG5, a gene from the C7q22 osteoarthritis susceptibility locus, induces glycosylation defects and affects chondrogenesis and osteogenesis by disturbing wnt signaling. *Osteoarthritis and Cartilage.*, **21**, S128–S129.

Campbell, M.S., Holt, C., Moore, B. and Yandell, M. (2013) Genome annotation and curation using MAKER and MAKER-P. *Current Protocols in Bioinformatics.*, **48**, 4–39.

Cantor, C.R., Smith, C.L. and Mathew, M.K. (1988) Pulsed-field gel electrophoresis of very large DNA molecules. *Annual Review of Biophysics and Biophysical Chemistry.*, **17**, 287–304.

Cao, H., Hastie, A.R., Cao, D., et al. (2014) Rapid detection of structural variation in a human genome using nano channel-based genome mapping technology. *Gigascience.*, **3**, 34.

Carneiro, M.O., Russ, C., Ross, M.G., Gabriel, S.B., Nusbaum, C. and DePristo, M.A. (2012) Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics.*, **13**, 375.

Carter, A.H., Garland-Campbell, K., Morris, C.F. and Kidwell, K.K. (2011) Chromosomes 3B and 4D are associated with several milling and baking quality traits in a soft white spring wheat (*Triticum aestivum L.*) population. *Theoretical and Applied Genetics.*, **124**, 1079–1096.

Castro, A.M., Vasicek, A., Ellerbrook, C., Giménez, D.O., Tocho, E., Tacaliti, M.S., Clua, A. and Snape, J.W. (2004) Mapping quantitative trait loci in wheat for resistance against greenbug and Russian wheat aphid. *Crop and Pasture Science.*, **123**, 361–365.

Catchen, J.M., Conery, J.S. and Postlethwait, J.H. (2009) Automated identification of conserved synteny after whole-genome duplication. *Genome Research.*, **19**, 1497–1505.

Cativelli, M., Lewis, S. and Appendino, M.L. (2013) A fusarium head blight resistance quantitative trait locus on chromosome 7D of the spring wheat cultivar Catbird. *Crop Science.*, **53**, 1464.

Chaisson, M.J. and Tesler, G. (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics.*, **13**, 238.

Chaisson, M.J., Brinza, D. and Pevzner, P.A. (2008) *De novo* fragment assembly with short mate-paired reads: does the read length matter? *Genome Research.*, **19**, 336–346.

Chalhoub, B., Belcram, H. and Caboche, M. (2004) Efficient cloning of plant genomes into bacterial artificial chromosome (BAC) libraries with larger and more uniform insert size. *Plant Biotechnology Journal.*, **2**, 181–188.

Chan, E.Y. (2004) DNA mapping using microfluidic stretching and single-molecule detection of fluorescent site-specific tags. *Genome Research.*, **14**, 1137–1146.

Chantret, N., Salse, J., Sabot, F., et al. (2005) Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *The Plant Cell.*, **17**, 1033–1045.

Chao, D.-Y., Dilkes, B., Luo, H., Douglas, A., Yakubova, E., Lahner, B. and Salt, D.E. (2013) Polyploids exhibit higher potassium uptake and salinity tolerance in *Arabidopsis*. *Science.*, **341**, 658–659.

Chen, M., Presting, G., Barbazuk, W.B., et al. (2002) An integrated physical and genetic map of the rice genome. *The Plant Cell.*, **14**, 537–545.

Chin, C.-S., Alexander, D.H., Marks, P., et al. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods.*, **10**, 563–569.

Choulet, F., Alberti, A., Theil, S., et al. (2014) Structural and functional partitioning of bread wheat chromosome 3B. *Science.*, **345**, 1249721–1249721.

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

Church, D.M., Goodstadt, L., Hillier, L.W., et al. (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse R. J. Roberts, ed. *PLoS Biology.*, **7**, e1000112.

Conesa, A. and Götz, S. (2008) Blast2GO: a comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics.*, **2008**, 619832–12.

Copenhaver, G.P., Nickel, K., Kuromori, T., et al. (1999) Genetic definition and sequence analysis of *Arabidopsis* centromeres. *Science.*, **286**, 2468–2474.

Corredor, E., Lukaszewski, A.J., Pachón, P., Allen, D.C. and Naranjo, T. (2007) Terminal regions of wheat chromosomes select their pairing partners in meiosis. *Genetics.*, **177**, 699–706.

Cossegal, M., Chambrier, P., Mbelo, S., et al. (2008) Transcriptional and metabolic adjustments in ADP-glucose pyrophosphorylase-deficient bt2 maize kernels. *Plant Physiology.*, **146**, 1553–1570.

Coulson, A., Sulston, J., Brenner, S. and Karn, J. (1986) Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences.*, **83**, 7821–7825.

Cox, T.S. (1997) Deepening the wheat gene pool. *Journal of Crop Production.*, **1**, 1–25.

Crawford, A.C. and Francki, M.G. (2013) Chromosomal location of wheat genes of the carotenoid biosynthetic pathway and evidence for a catalase gene on chromosome 7A functionally associated with flour b* colour variation. *Molecular Genetics and Genomics.*, **288**, 483–493.

Crawford, A.C. and Francki, M.G. (2012) Lycopene-ε-cyclase (e-LCY3A) is functionally associated with quantitative trait loci for flour b* colour on chromosome 3A in wheat (*Triticum aestivum* L.). *Molecular Breeding.*, **31**, 737–741.

Crick, F.H.C. (1952) Is alpha-keratin a coiled coil? *Nature.*, **170**, 882–883.

Crick, F.H.C. (1953) The packing of α-helices: simple coiled-coils. *Acta Crystallographica.*, **6**, 689–697.

Crosby, M.A., Goodman, J.L., Strelets, V.B., Zhang, P., Gelbart, W.M.The FlyBase Consortium (2007) FlyBase: genomes by the dozen. *Nucleic Acids Research.*, **35**, D486–D491.

D rsquo Hont, A. (2005) Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. *Cytogenetic and Genome Research.*, **109**, 27–33.

Dadkhodaie, N.A., Karaoglou, H., Wellings, C.R. and Park, R.F. (2010) Mapping genes Lr53 and Yr35 on the short arm of chromosome 6B of common wheat with microsatellite markers and studies of their association with Lr36. *Theoretical and Applied Genetics.*, **122**, 479–487.

Das, S.K., Austin, M.D., Akana, M.C., Deshpande, P., Cao, H. and Xiao, M. (2010) Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic Acids Research.*, **38**, e177–e177.

Dawe, R.K., Reed, L.M., Yu, H.G., Muszynski, M.G. and Hiatt, E.N. (1999) A maize homolog of mammalian CENPC is a constitutive component of the inner kinetochore. *The Plant Cell.*, **11**, 1227–1238.

Denton, J.F., Lugo-Martinez, J., Tucker, A.E., Schrider, D.R., Warren, W.C. and Hahn, M.W. (2014) Extensive error in the number of genes inferred from draft genome assemblies R. Guigó, ed. *PLoS Computational Biology.*, **10**, e1003998.

Deorowicz, S., Debudaj-Grabysz, A. and Grabowski, S. (2013) Disk-based k-mer counting on a PC. *BMC Bioinformatics.*, **14**, 160.

DeSalle, R., Gregory, T.R. and Johnston, J.S. (2005) Preparation of samples for comparative studies of arthropod chromosomes: visualization, in situ hybridization, and genome size estimation. In *Molecular evolution: Producing the biochemical data*. Methods in Enzymology. Elsevier, pp. 460–488.

Devos, K.M., Dubcovsky, J., Dvořák, J., Chinoy, C.N. and Gale, M.D. (1995) Structural evolution of wheat chromosomes 4A, 5A, and 7B and its impact on recombination. *Theoretical and Applied Genetics.*, **91**, 282–288.

Ding, Y., Johnson, M.D., Chen, W.Q., Wong, D., Chen, Y.J., Benson, S.C., Lam, J.Y., Kim, Y.M. and Shizuya, H. (2001) Five-color-based high-information-content fingerprinting of bacterial artificial chromosome clones using type IIS restriction endonucleases. *Genomics.*, **74**, 142–154.

Dohm, J.C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for *de novo* genomic sequencing. *Genome Research.*, **17**, 1697–1706.

Dohm, J.C., Minoche, A.E., Holtgräwe, D., et al. (2014) The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature.*, **505**, 546–549.

Doležel, J. and Lucretti, S. (1995) High-resolution flow karyotyping and chromosome sorting in *Vicia faba* lines with standard and reconstructed karyotypes. *Theoretical and Applied Genetics.*, **90**, 797–802.

Dubcovsky, J. and Dvořák, J. (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science.*, **316**, 1862–1866.

Dumur, J., Branlard, G., Tanguy, A.-M., Dardevet, M., Coriton, O., Huteau, V., Lemoine, J. and Jahier, J. (2009) Homoeologous recombination within bread wheat to develop novel combinations of HMW-GS genes: transfer of the Glu-A1 locus to chromosome 1D. *Planta.*, **231**, 57–65.

Earl, D., Bradnam, K., St John, J., et al. (2011) Assemblathon 1: a competitive assessment of *de novo* short read assembly methods. *Genome Research.*, **21**, 2224–2241.

Eichinger, L., Pachebat, J.A., Glöckner, G., et al. (2005) The genome of the social amoeba *Dictyostelium discoideum*. *Nature.*, **435**, 43–57.

Eid, J., Fehr, A., Gray, J., et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science.*, **323**, 133–138.

Eilbeck, K., Moore, B., Holt, C. and Yandell, M. (2009) Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics.*, **10**, 67.

Ellis, J. (1982) Promiscuous DNA--chloroplast genes inside plant mitochondria. *Nature.*, **299**, 678–679.

Endo, T.R. and Gill, B.S. (1996) The deletion stocks of common wheat. *Journal of Heredity.*, **87**, 295–307.

English, A.C., Salerno, W.J., Hampton, O.A., et al. (2015) Assessing structural variation in a personal genome-towards a human reference diploid genome. *BMC Genomics.*, **16**, 286.

Estill, J.C. and Bennetzen, J.L. (2008) The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes. *Plant Methods.*, **5**, 8–8.

Ewing, B., Hillier, L., Wendl, M.C. and Green, P. (1998) Base-calling of automated sequencer traces using Phred I. Accuracy assessment. *Genome Research.*, **8**, 175–185.

Feldman, M. and Levy, A.A. (2005) Allopolyploidy--a shaping force in the evolution of wheat genomes. *Cytogenetic and Genome Research.*, **109**, 250–258.

Finn, R.D., Bateman, A., Clements, J., et al. (2014) Pfam: the protein families database. *Nucleic Acids Research.*, **42**, D222–30.

Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Systematic Zoology.*, **19**, 99.

Florea, L., Souvorov, A., Kalbfleisch, T.S. and Salzberg, S.L. (2011) Genome assembly has a major impact on gene content: a comparison of annotation in two *Bos taurus* assemblies. N. M. El-Sayed, ed. *PloS ONE.*, **6**, e21400.

Fu, S., Ren, Z., Chen, X., Yan, B., Tan, F., Fu, T. and Tang, Z. (2014) New wheat-rye 5DS-4RS·4RL and 4RS-5DS·5DL translocation lines with powdery mildew resistance. *Journal of Plant Research.*, **127**, 743–753.

Gardiner, M. and Chrispeels, M.J. (1975) Involvement of the golgi apparatus in the synthesis and secretion of hydroxyproline-rich cell wall glycoproteins. *Plant Physiology.*, **55**, 536–541.

Gardner, M.J., Hall, N., Fung, E., et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature.*, **419**, 498–511.

Gaut, B.S. and Ross-Ibarra, J. (2008) Selection on major components of angiosperm genomes. *Science.*, **320**, 484–486.

Genc, Y., Taylor, J., Rongala, J. and Oldach, K. (2014) A major locus for chloride accumulation on chromosome 5A in bread wheat T. Wang, ed. *PloS ONE.*, **9**, e98845.

Gerdol, M., Venier, P. and Pallavicini, A. (2015) The genome of the pacific oyster *Crassostrea gigas* brings new insights on the massive expansion of the C1q gene family in *Bivalvia*. *Developmental & Comparative Immunology.*, **49**, 59–71.

Gill, B.S. (2004) A workshop report on wheat genome sequencing: International Genome Research on wheat consortium. *Genetics.*, **168**, 1087–1096.

Gillingham, A.K. and Munro, S. (2003) Long coiled-coil proteins and membrane traffic. *Biochimica et Biophysica Acta.*, **1641**, 71–85.

Glazer, A.N. and Rye, H.S. (1992) Stable dye–DNA intercalation complexes as reagents for high-sensitivity fluorescence detection. *Nature.*, **359**, 859–861.

Gnerre, S., MacCallum, I., Przybylski, D., et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences.*, **108**, 1513–1518.

Goff, S.A. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science.*, **296**, 92–100.

Golan, D. and Medvedev, P. (2013) Using state machines to model the Ion Torrent sequencing process and to improve read error rates. *Bioinformatics.*, **29**, i344–i351.

Grabherr, M.G., Haas, B.J., Yassour, M., et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology.*, **29**, 644–652.

Groth, D., Lehrach, H. and Hennig, S. (2004) GOblet: a platform for gene ontology annotation of anonymous sequence data. *Nucleic Acids Research.*, **32**, W313–7.

Gruber, M. and Lupas, A.N. (2003) Historical review: another 50th anniversary--new periodicities in coiled coils. *Trends in Biochemical Sciences.*, **28**, 679–685.

Gruber, M., Söding, J. and Lupas, A.N. (2006) Comparative analysis of coiled-coil prediction methods. *Journal of Structural Biology.*, **155**, 140–145.

Gulick, P., Wanjugi, H., Coleman-Derr, D., Huo, N., Kianian, S.F., Luo, M., Wu, J., Anderson, O. and Gu, Y.Q. (2009) Rapid development of PCR-based genome-specific repetitive DNA junction markers in wheat. *Genome.*, **52**, 576–587.

Guo, X.-H., Huff, E.J. and Schwartz, D.C. (1992) Sizing single DNA molecules. *Nature.*, **359**, 783–784.

Ha, J.Y., Pokrovskaya, I.D., Climer, L.K., Shimamura, G.R., Kudlyk, T., Jeffrey, P.D., Lupashin, V.V. and Hughson, F.M. (2014) Cog5-Cog7 crystal structure reveals interactions essential for the function of a multisubunit tethering complex. *Proceedings of the National Academy of Sciences.*, **111**, 15762–15767.

Ha, K.D., Clarke, B.A. and Brown, W.J. (2012) Regulation of the golgi complex by phospholipid remodeling enzymes. *Biochimica et Biophysica Acta.*, **1821**, 1078–1088.

Hall, A.R., Scott, A., Rotem, D., Mehta, K.K., Bayley, H. and Dekker, C. (2010) Hybrid pore formation by directed insertion of α-haemolysin into solid-state nanopores. *Nature Nanotechnology.*, **5**, 874–877.

Han, Y. and Wessler, S.R. (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research.*, **38**, e199–e199.

Hastie, A.R., Dong, L., Smith, A., et al. (2013) Rapid genome mapping in nanochannel arrays for highly complete and accurate *de novo* sequence assembly of the complex *Aegilops tauschii* genome J. C. Nelson, ed. *PloS ONE.*, **8**, e55864.

Haudry, A., Cenci, A., Ravel, C., et al. (2007) Grinding up wheat: a massive loss of nucleotide diversity since domestication. *Molecular Biology and Evolution.*, **24**, 1506–1517.

Henikoff, S., Ahmad, K. and Malik, H.S. (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science.*, **293**, 1098–1102.

Heslop-Harrison, J.S., Murata, M., Ogura, Y., Schwarzacher, T. and Motoyoshi, F. (1999) Polymorphisms and genomic organization of repetitive DNA from centromeric regions of *Arabidopsis* chromosomes. *The Plant Cell.*, **11**, 31–42.

Hosid, E., Brodsky, L., Kalendar, R., Raskina, O. and Belyayev, A. (2012) Diversity of long terminal repeat retrotransposon genome distribution in natural populations of the wild diploid wheat *Aegilops speltoides* D. Voytas, ed. *Genetics.*, **190**, 263–274.

Howe, D.G., Bradford, Y.M., Conlin, T., et al. (2013) ZFIN, the zebrafish model organism database: increased support for mutants and transgenics. *Nucleic Acids Research.*, **41**, D854–60.

Howison, M., Zapata, F., Edwards, E.J. and Dunn, C.W. (2013) Bayesian genome assembly and assessment by markov chain monte carlo sampling. M. Rattray, ed. *PloS ONE.*, **9**, e99497–e99497.

Hsam, N.B.O., Kowalczyk, K., Zeller, F.J. and Hsam, S.L.K. (2014) Characterization of powdery mildew resistance and linkage studies involving the Pm3 locus on chromosome 1A of common wheat (*Triticum aestivum* L.). *Journal of Applied Genetics.*, **56**, 37–44.

Huang, C.Y., Ayliffe, M.A. and Timmis, J.N. (2003) Direct measurement of the transfer rate of chloroplast DNA into the nucleus. *Nature.*, **422**, 72–76.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research.*, **37**, 1–13.

Huang, Q., Li, X., Chen, W.Q., et al. (2014) Genetic mapping of a putative *Thinopyrum* intermedium-derived stripe rust resistance gene on wheat chromosome 1B. *Theoretical and Applied Genetics.*, **127**, 843–853.

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

Huang, X. and Yang, S.-P. (2002) *Generating a genome assembly with PCAP*, Hoboken, NJ, USA: John Wiley & Sons, Inc.

Huang, X.Q., Kempf, H., Ganal, M.W. and R der, M.S. (2004) Advanced backcross QTL analysis in progenies derived from a cross between a German elite winter wheat variety and a synthetic wheat (*Triticum aestivum* L.). *Theoretical and Applied Genetics.*, **109**, 933–943.

Huang, Z., Petty, J.T., O'Quinn, B., Longmire, J.L., Brown, N.C., Jett, J.H. and Keller, R.A. (1996) Large DNA fragment sizing by flow cytometry: Application to the characterization of P1 artificial chromosome (PAC) clones. *Nucleic Acids Research.*, **24**, 4202–4209.

Huda, A. and Jordan, I.K. (2009) Analysis of transposable element sequences using CENSOR and RepeatMasker. *Methods in Molecular Biology.*, **537**, 323–336.

Idury, R.M. and Waterman, M.S. (1995) A new algorithm for DNA sequence assembly. *Journal of Computational Biology.*, **2**, 291–306.

Iehisa, J.C.M., Matsuura, T., Mori, I.C. and Takumi, S. (2014) Identification of quantitative trait locus for abscisic acid responsiveness on chromosome 5A and association with dehydration tolerance in common wheat seedlings. *Journal of Plant Physiology.*, **171**, 25–34.

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature.*, **436**, 793–800.

IWGSC, Mayer, K.F.X., Rogers, J., et al. (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science.*, **345**, 1251788–1251788.

Jacoby, W.G. (2000) Loess. *Electoral Studies.*, **19**, 577–613.

Jaillon, O., Aury, J.-M., Noel, B., et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature.*, **449**, 463–467.

Janda, J., Bartos, J., Safár, J., et al. (2004) Construction of a subgenomic BAC library specific for chromosomes 1D, 4D and 6D of hexaploid wheat. *Theoretical and Applied Genetics.*, **109**, 1337–1345.

Janda, J., Safár, J., Kubaláková, M., et al. (2006) Advanced resources for plant genomics: a BAC library specific for the short arm of wheat chromosome 1B. *The Plant Journal.*, **47**, 977–986.

Jansen, R.K., Kaittanis, C., Saski, C., Lee, S.-B., Tomkins, J., Alverson, A.J. and Daniell, H. (2006) Phylogenetic analyses of *Vitis (Vitaceae)* based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. *BMC Evolutionary Biology.*, **6**, 32.

Jayatilake, D.V., Bai, G.H. and Dong, Y.H. (2011) A novel quantitative trait locus for fusarium head blight resistance in chromosome 7A of wheat. *Theoretical and Applied Genetics.*, **122**, 1189–1198.

Jeck, W.R., Reinhardt, J.A., Baltrus, D.A., Hickenbotham, M.T., Magrini, V., Mardis, E.R., Dangl, J.L. and Jones, C.D. (2007) Extending assembly of short DNA sequences to handle error. *Bioinformatics.*, **23**, 2942–2944.

Jiang, H., Lei, R., Ding, S.-W. and Zhu, S. (2014) Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics.*, **15**, 182.

Jiang, N., Bao, Z., Zhang, X., Eddy, S.R. and Wessler, S.R. (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature.*, **431**, 569–573.

Jing, J., Reed, J., Huang, J., et al. (1998) Automated high resolution optical mapping using arrayed, fluid-fixed DNA molecules. *Proceedings of the National Academy of Sciences.*, **95**, 8046–8051.

Jones, P., Binns, D., Chang, H.-Y., et al. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics.*, **30**, 1236–1240.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome*

Research., **110**, 462–467.

Kadam, S., Singh, K., Shukla, S., Goel, S., Vikram, P., Pawar, V., Gaikwad, K., Khanna-Chopra, R. and Singh, N. (2012) Genomic associations for drought tolerance on the short arm of wheat chromosome 4B. *Functional & Integrative Genomics.*, **12**, 447–464.

Kajitani, R., Toshimoto, K., Noguchi, H., et al. (2014) Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research.*, **24**, 1384–1395.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2015) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research.*, **44**, gkv1070–D462.

Kang, Y.J., Kim, S.K., Kim, M.Y., et al. (2014) Genome sequence of mungbean and insights into evolution within *Vigna* species. *Nature Communications.*, **5**, 5443.

Kang, Y.J., Satyawan, D., Shim, S., et al. (2015) Draft genome sequence of adzuki bean, *Vigna* angularis. *Scientific Reports.*, **5**, 8069.

Kapustin, Y., Souvorov, A., Tatusova, T. and Lipman, D. (2008) Splign: algorithms for computing spliced alignments with identification of paralogs. *Biology Direct.*, **3**, 20.

Kasianowicz, J.J., Brandin, E., Branton, D. and Deamer, D.W. (1996) Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences.*, **93**, 13770–13773.

Kazakoff, S.H., Imelfort, M., Edwards, D., Koehorst, J., Biswas, B., Batley, J., Scott, P.T. and Gresshoff, P.M. (2011) Capturing the biofuel wellhead and powerhouse: the chloroplast and mitochondrial genomes of the leguminous feedstock tree *Pongamia pinnata*. J. H. Badger, ed. *PloS ONE.*, **7**, e51687–e51687.

Kelley, D.R. and Salzberg, S.L. (2010) Detection and correction of false segmental duplications caused by genome mis-assembly. *Genome Biology.*, **11**, R28.

Kellogg, E.A. (2015) Genome sequencing: long reads for a short plant. *Nature Plants.*, **1**, 15169.

Khatri, P. and Drăghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics.*, **21**, 3587–3595.

Kidd, J.M., Cooper, G.M., Donahue, W.F., et al. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature.*, **453**, 56–64.

Kim, H.J., Kim, J., Kang, K.S., Lee, K.T. and Yang, H.O. (2014) Neuroprotective effect of chebulagic acid via autophagy induction in SH-SY5Y cells. *Biomolecules & Therapeutics.*, **22**, 275–281.

Kim, J., Roh, J., Kwon, D., et al. (2014) Estimation of the genome sizes of the chigger mites *Leptotrombidium pallidum* and *Leptotrombidium scutellare* based on quantitative PCR and k-mer analysis. *Parasites & Vectors.*, **7**, 279.

Kjer-Nielsen, L., Teasdale, R.D., van Vliet, C. and Gleeson, P.A. (1999) A novel golgi-localisation domain shared by a class of coiled-coil peripheral membrane proteins. *Current Biology.*, **9**, 385–388.

Kocsy, G., Athmer, B., Perovic, D., Himmelbach, A., Szucs, A., Vashegyi, I., Schweizer, P., Galiba, G. and Stein, N. (2010) Regulation of gene expression by chromosome 5A during cold hardening in wheat. *Molecular Genetics and Genomics.*, **283**, 351–363.

Koonin, E.V., Mushegian, A.R. and Bork, P. (1996) Non-orthologous gene displacement. *Trends in Genetics.*, **12**, 334–336.

Korf, I. (2004) Gene finding in novel genomes. *BMC Bioinformatics.*, 59.

Korlach, J., Marks, P.J., Cicero, R.L., et al. (2008) Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proceedings of the National Academy of Sciences.*, **105**, 1176–1181.

Korpar, M. and Šikic, M. (2013) SW#-GPU-enabled exact alignments on genome scale. *Bioinformatics.*, **29**, 2494–2495.

Krattinger, S.G., Lagudah, E.S., Spielmeyer, W., Singh, R.P., Huerta-Espino, J., McFadden, H., Bossolini, E., Selter, L.L. and Keller, B. (2009) A putative ABC transporter confers durable resistance to multiple fungal pathogens in wheat. *Science.*, **323**, 1360–1363.

Kruppa, K., Sepsi, A., Szakács, É., Röder, M.S. and Molnár-Láng, M. (2013) Characterization of a 5HS-7DS.7DL wheat-barley translocation line and physical mapping of the 7D chromosome using SSR markers. *Journal of Applied Genetics.*, **54**, 251–258.

Kubaláková, M., Vrána, J., Cíhalíková, J., Šimková, H. and Doležel, J. (2002) Flow karyotyping and chromosome sorting in bread wheat (*Triticum aestivum* L.). *Theoretical and Applied Genetics.*, **104**, 1362–1372.

Kumar, A., Seetan, R., Mergoum, M., et al. (2015) Radiation hybrid maps of the D-genome of *Aegilops tauschii* and their application in sequence assembly of large and complex plant genomes. *BMC Genomics.*, **16**, 800.

Kurtz, S., Narechania, A., Stein, J.C. and Ware, D. (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics.*, **9**, 517.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. and Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biology.*, **5**, R12.

la Bastide, de, M. and McCombie, W.R. (2002) *Assembling genomic DNA sequences with PHRAP*, Hoboken, NJ, USA: John Wiley & Sons, Inc.

Lai, J., Li, Y., Messing, J. and Dooner, H.K. (2005) Gene movement by helitron transposons contributes to the haplotype variability of maize. *Proceedings of the National Academy of Sciences.*, **102**, 9068–9073.

Lai, Z., Jing, J., Aston, C., et al. (1999) A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nature Genetics.*, **23**, 309–313.

Lam, E.T., Hastie, A., Lin, C., et al. (2012) Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature Biotechnology.*, **30**, 771–776.

Landegren, U., Kaiser, R., Sanders, J. and Hood, L. (1988) A ligase-mediated gene detection technique. *Science.*, **241**, 1077–1080.

Lee, H., Gurtowski, J., Yoo, S., Marcus, S., McCombie, W.R. and Schatz, M. (2014) *Error correction and assembly complexity of single molecule sequencing reads*, Cold Spring Harbor Labs Journals.

Leggett, R.M., Clavijo, B.J., Clissold, L., Clark, M.D. and Caccamo, M. (2014) NextClip: an analysis and read preparation tool for Nextera long mate pair libraries. *Bioinformatics.*, **30**, 566–568.

Leroy, P., Guilhot, N., Sakai, H., et al. (2012) TriAnnot: a versatile and high performance pipeline for the automated annotation of plant genomes. *Frontiers in Plant Science.*, **3**, 5.

Levene, M.J. (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. *Science.*, **299**, 682–686.

Li, B., Choulet, F., Heng, Y., et al. (2013) Wheat centromeric retrotransposons: the new ones take a major role in centromeric structure. *The Plant Journal.*, **73**, 952–965.

Li, H. and Homer, N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics.*, **11**, 473–483.

Li, L., Stoeckert, C.J. and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research.*, **13**, 2178–2189.

Li, Q., Chen, X.M., Wang, M.N. and Jing, J.X. (2010) Yr45, a new wheat gene for stripe rust resistance on the long arm of chromosome 3D. *Theoretical and Applied Genetics.*, **122**, 189–197.

Li, R., Fan, W., Tian, G., et al. (2009) The sequence and *de novo* assembly of the giant panda genome. *Nature.*, **463**, 311–317.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K. and Wang, J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics.*, **25**, 1966–1967.

Li, R., Zhu, H., Ruan, J., et al. (2010) *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research.*, **20**, 265–272.

Li, Z., Chen, Y., Mu, D., et al. (2012) Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and *de-bruijn* graph. *Briefings in Functional Genomics.*, **11**, 25–37.

Lin, J. (1999) Whole-genome shotgun optical mapping of *Deinococcus radiodurans*. *Science.*, **285**, 1558–1562.

Lin, L., Pierce, G.J., Bowers, J.E., et al. (2010) A draft physical map of a D-genome cotton species (*Gossypium raimondii*). *BMC Genomics.*, **11**, 395.

Ling, H.-Q., Zhao, S., Liu, D., et al. (2013) Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature.*, **496**, 87–90.

Lippman, Z., Gendrel, A.-V., Black, M., et al. (2004) Role of transposable elements in heterochromatin and epigenetic control. *Nature.*, **430**, 471–476.

Liu, B., Shi, Y., Yuan, J., et al. (2013) Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects. *arXiv*, **1308**, 2012.

Liu, B., Yuan, J., Yiu, S.M., et al. (2012) COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. *Bioinformatics.*, **28**, 2870–2874.

Liu, C., Wong, T., Wu, E., et al. (2012) SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. *Bioinformatics.*, **28**, 878–879.

Liu, C.J., Atkinson, M.D., Chinoy, C.N., Devos, K.M. and Gale, M.D. (1992) Non-homoeologous translocations between group 4, 5 and 7 chromosomes within wheat and rye. *Theoretical and Applied Genetics.*, **83**.

Liu, L., sup, L.W.S.1., sup, J.Y.S.2., sup, Y.Z.S.2. and sup, C.Z.S.1. (2010) Association mapping of six agronomic traits on chromosome 4A of wheat (*Triticum aestivum L.*). *Molecular Plant Breeding.*, **1**.

Liu, S., Liu, Y., Yang, X., et al. (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nature Communications.*, **5**, 3930.

Liu, W., Schmidt, B. and Müller-Wittig, W. (2011) CUDA-BLASTP: accelerating BLASTP on CUDA-enabled graphics hardware. *Transactions on Computational Biology and Bioinformatics.*, **8**, 1678–1684.

Liu, X.M., Smith, C.M., Friebe, B.R. and Gill, B.S. (2005) Molecular mapping and allelic relationships of Russian wheat aphid–resistance genes. *Crop Science.*, **45**, 2273–2280.

Lockton, S. and Gaut, B.S. (2005) Plant conserved non-coding sequences and paralogue evolution. *Trends in Genetics.*, **21**, 60–65.

Loman, N.J., Misra, R.V., Dallman, T.J., Constantinidou, C., Gharbia, S.E., Wain, J. and Pallen, M.J. (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology.*, **30**, 434–439.

Lu, T., Cui, L., Zhou, Y., et al. (2015) Transcriptome-wide investigation of circular RNAs in rice. *RNA*, **21**, 2076–2087.

Luke, M.R., Houghton, F., Perugini, M.A. and Gleeson, P.A. (2005) The trans-golgi network GRIP-domain proteins form alpha-helical homodimers. *The Biochemical Journal.*, **388**, 835–841.

Lundquist, P.M., Zhong, C.F., Zhao, P., et al. (2008) Parallel confocal detection of single molecules in real time. *Optics Letters.*, **33**, 1026.

Luo, M., Thomas, C., You, F.M., et al. (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics.*, **82**, 378–389.

Luo, M.C., Deal, K.R., Akhunov, E.D., et al. (2009) Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in *Triticeae*. *Proceedings of the National Academy of Sciences.*, **106**, 15780–15785.

Luo, P., Hu, X., Zhang, H. and Ren, Z. (2009) Genes for resistance to stripe rust on chromosome 2B and their application in wheat breeding. *Progress in Natural Science.*, **19**, 9–15.

Lupski, J.R. (2007) Structural variation in the human genome. *New England Journal of Medicine.*, **356**, 1169–1171.

Ma, Hongqi, Kong, Z., Fu, B., Li, N., Zhang, L., Jia, H. and Ma, Z. (2011) Identification and mapping of a new powdery mildew resistance gene on chromosome 6D of common wheat. *Theoretical and Applied Genetics.*, **123**, 1099–1106.

Ma, Jian, Stiller, J., Berkman, P.J., et al. (2013) Sequence-based analysis of translocations and inversions in bread wheat (*Triticum aestivum* L.) J. C. Nelson, ed. *PloS ONE.*, **8**, e79329.

Ma, Jiong, Morrow, D.J., Fernandes, J. and Walbot, V. (2006) Comparative profiling of the sense and antisense transcriptome of maize lines. *Genome Biology.*, **7**, R22.

Magoc, T. and Salzberg, S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics.*, **27**, 2957–2963.

Maluszynska, J. and Heslop-Harrison, J.S. (1991) Localization of tandemly repeated DNA sequences in *Arabidopsis thaliana*. *The Plant Journal.*, **1**, 159–166.

Mancinelli, L., De Angelis, P.M., Annulli, L., Padovini, V., Elgjo, K. and Gianfranceschi, G.L. (2009) A class of DNA-binding peptides from wheat bud causes growth inhibition, G2 cell cycle arrest and apoptosis induction in HeLa cells. *Molecular Cancer.*, **8**, 55.

Marcais, G. and Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.*, **27**, 764–770.

Margulies, M., Egholm, M., Altman, W.E., et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature.*, **437**, 376–380.

Marine, R., Polson, S.W., Ravel, J., Hatfull, G., Russell, D., Sullivan, M., Syed, F., Dumas, M. and Wommack, K.E. (2011) Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. *Applied and Environmental Microbiology.*, **77**, 8071–8079.

Marquardt, T. and Denecke, J. (2003) Congenital disorders of glycosylation: Review of their molecular bases, clinical presentations and specific therapies. *European Journal of Pediatrics.*, **162**, 359–379.

Marra, M., Kucaba, T., Sekhon, M., et al. (1999) A map for sequence analysis of the *Arabidopsis thaliana* genome. *Nature Genetics.*, **22**, 265–270.

Marra, M.A., Kucaba, T.A., Dietrich, N.L., et al. (1997) High throughput fingerprint analysis of large-insert clones. *Genome Research.*, **7**, 1072–1084.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal.*, **17**, 10.

Martinez-Perez, E., Shaw, P. and Moore, G. (2001) The Ph1 locus is needed to ensure specific somatic and meiotic centromere association. *Nature.*, **411**, 204–207.

Masterson, J. (1994) Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science.*, **264**, 421–424.

Matsuo, M., Ito, Y., Yamauchi, R. and Obokata, J. (2005) The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux. *The Plant Cell.*, **17**, 665–675.

Mayer, K.E., Taudien, S., Martis, M., et al. (2009) Gene content and virtual gene order of barley chromosome 1H. *Plant Physiology.*, **151**, 496–505.

Mäkinen, V., Salmela, L. and Ylinen, J. (2012) Normalized N50 assembly metric using gap-restricted co-linear chaining. *BMC Bioinformatics.*, **13**, 255.

McClure, M.A., Richardson, H.S., Clinton, R.A., Hepp, C.M., Crowther, B.A. and Donaldson, E.F. (2005) Automated characterization of potentially active retroid agents in the human genome. *Genomics.*, **85**, 512–523.

McGinnis, S. and Madden, T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research.*, **32**, W20–W25.

McPherson, J.D., Marra, M., Hillier, L., et al. (2001) A physical map of the human genome. *Nature.*, **409**, 934–941.

Mefford, H.C. and Trask, B.J. (2002) The complex structure and dynamic evolution of human subtelomeres. *Nature Reviews Genetics.*, **3**, 91–102.

Melnikova, N.V., Konovalov, F.A. and Kudryavtsev, A.M. (2011) Long terminal repeat retrotransposon Jeli provides multiple genetic markers for common wheat (*Triticum aestivum*). *Plant Genetic Resources.*, **9**, 163–165.

Melsted, P. and Pritchard, J.K. (2011) Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC Bioinformatics.*, **12**, 333.

Merriman, B., R D Team, I.T. and Rothberg, J.M. (2012) Progress in Ion Torrent semiconductor chip based sequencing V. M. Ugaz, ed. *Electrophoresis.*, **33**, 3397–3417.

Metzker, M.L. (2010) Sequencing technologies — the next generation. *Nature Reviews Genetics.*, **11**, 31–46.

Miftahudin (2004) Analysis of expressed sequence tag loci on wheat chromosome group 4. *Genetics.*, **168**, 651–663.

Mikheyev, A.S. and Tin, M.M.Y. (2014) A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources.*, **14**, 1097–1102.

Miller, C.A., Altinkut, A. and Lapitan, N.L.V. (2001) A microsatellite marker for tagging , a wheat gene conferring resistance to the Russian wheat aphid. *Crop Science.*, **41**, 1584.

Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A. and Rafalski, A. (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature Genetics.*, **37**, 997–1002.

Morgulis, A., Gertz, E.M., Schäffer, A.A. and Agarwala, R. (2005) WindowMasker: window based masker for sequenced genomes. *Bioinformatics.*, **22**, 134–141.

Moullet, O., Zhang, H.B. and Lagudah, E.S. (1999) Construction and characterisation of a large DNA insert library from the D genome of wheat. *Theoretical and Applied Genetics.*, **99**, 305–313.

Mulder, N. and Apweiler, R. (2007) InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods in Molecular Biology.*, **396**, 59–70.

Mullikin, J.C. (2002) The Phusion assembler. *Genome Research.*, **13**, 81–90.

Multer, G.L. and Boynton, K.A. (1995) PCR bias in amplification of androgen receptor alleles, a trinucleotide repeat marker used in clonality studies. *Nucleic Acids Research.*, **23**, 1411–1418.

Munro, S. and Nichols, B.J. (1999) The GRIP domain - a novel golgi-targeting domain found in several coiled-coil proteins. *Current Biology.*, **9**, 377–380.

Murata, M., Ogura, Y. and Motoyoshi, F. (1994) Centromeric repetitive sequences in *Arabidopsis thaliana*. *The Japanese Journal of Genetics.*, **69**, 361–370.

Myers, E.W. (2000) A whole-genome assembly of *Drosophila*. *Science.*, **287**, 2196–2204.

Nakano, M., Komatsu, J., Matsuura, S.-I., Takashima, K., Katsura, S. and Mizuno, A. (2003) Single-molecule PCR using water-in-oil emulsion. *Journal of Biotechnology.*, **102**, 117–124.

Nalam, V.J., Vales, M.I., Watson, C.J.W., Johnson, E.B. and Riera-Lizarazu, O. (2007) Map-based analysis of genetic loci on chromosome 2D that affect glume tenacity and threshability, components of the free-threshing habit in common wheat (*Triticum aestivum L.*). *Theoretical and Applied Genetics.*, **116**, 135–145.

Nehrt, N.L., Clark, W.T., Radivojac, P. and Hahn, M.W. (2011) Testing the ortholog conjecture with comparative functional genomic data from mammals A. Rzhetsky, ed. *PLoS Computational Biology.*, **7**, e1002073.

Nelson, J.C., Sorrells, M.E., Van Deynze, A.E., Lu, Y.H., Atkinson, M., Bernard, M., Leroy, P., Faris, J.D. and Anderson, J.A. (1995) Molecular mapping of wheat: major genes and rearrangements in homoeologous groups 4, 5, and 7. *Genetics.*, **141**, 721–731.

Nelson, R.T. and Shoemaker, R. (2006) Identification and analysis of gene families from the duplicated genome of soybean using EST sequences. *BMC Genomics.*, **7**, 204.

Nelson, W. and Soderlund, C. (2009) Integrating sequence with FPC fingerprint maps. *Nucleic Acids Research.*, **37**, e36–e36.

Nelson, W.M., Dvorak, J., Luo, M., Messing, J., Wing, R.A. and Soderlund, C. (2007) Efficacy of clone fingerprinting methodologies. *Genomics.*, **89**, 160–165.

Newman, J.R.S., Wolf, E. and Kim, P.S. (2000) A computationally directed screen identifying interacting coiled coils from *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences.*, **97**, 13203–13208.

Nikolovski, N., Rubtsov, D., Segura, M.P., Miles, G.P., Stevens, T.J., Dunkley, T.P.J., Munro, S., Lilley, K.S. and Dupree, P. (2012) Putative glycosyltransferases and other plant golgi apparatus proteins are revealed by LOPIT proteomics. *Plant Physiology.*, **160**, 1037–1051.

Noir, S., Patheyron, S., Combes, M.-C., Lashermes, P. and Chalhoub, B. (2004) Construction and characterisation of a BAC library for genome analysis of the allotetraploid coffee species (*Coffea arabica L.*). *Theoretical and Applied Genetics.*, **109**, 225–230.

O'Bleness, M., Searles, V.B., Dickens, C., et al. (2014) Finished sequence and assembly of the

DUF1220-rich 1q21 region using a haploid human genome. *BMC Genomics.*, **15**, 387.

O'Connell, J., Schulz-Trieglaff, O., Carlson, E., Hims, M.M., Gormley, N.A. and Cox, A.J. (2014) NxTrim: optimized trimming of Illumina mate pair reads. *bioRxiv*, 007666. Available at: http://biorxiv.org/lookup/doi/10.1101/007666.

O'Sullivan, H. (2007) GrainGenes. *Methods in Molecular Biology.*, **406**, 301–314.

Ogihara, Y., Isono, K., Kojima, T., et al. (2002) Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA. *Molecular Genetics and Genomics.*, **266**, 740–746.

Osoegawa, K., Vessere, G.M., Li Shu, C., Hoskins, R.A., Abad, J.P., de Pablos, B., Villasante, A. and de Jong, P.J. (2007) BAC clones generated from sheared DNA. *Genomics.*, **89**, 291–299.

Ozkan, H., Levy, A.A. and Feldman, M. (2001) Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops-Triticum*) group. *The Plant Cell.*, **13**, 1735–1747.

Palmer, D.K., O'Day, K., Trong, H.L., Charbonneau, H. and Margolis, R.L. (1991) Purification of the centromere-specific protein CENP-A and demonstration that it is a distinctive histone. *Proceedings of the National Academy of Sciences.*, **88**, 3734–3738.

Parisod, C., Holderegger, R. and Brochmann, C. (2010) Evolutionary consequences of autopolyploidy. *New Phytologist.*, **186**, 5–17.

Parsons, H.T., Christiansen, K., Knierim, B., et al. (2012) Isolation and proteomic characterization of the *Arabidopsis* golgi defines functional and novel components involved in plant cell wall biosynthesis. *Plant Physiology.*, **159**, 12–26.

Paterson, A.H., Bowers, J.E., Bruggmann, R., et al. (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature.*, **457**, 551–556.

Paux, E., Sourdille, P., Salse, J., et al. (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. *Science.*, **322**, 101–104.

Pendleton, M., Sebra, R., Pang, A.W.C., et al. (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods.*, **12**, 780–786.

Peng, J.H., Bai, Y., Haley, S.D. and Lapitan, N.L.V. (2008) Microsatellite-based molecular diversity of bread wheat germplasm and association mapping of wheat resistance to the Russian wheat aphid. *Genetica.*, **135**, 95–122.

Petersen, G., Seberg, O., Yde, M. and Berthelsen, K. (2006) Phylogenetic relationships of *Triticum* and *Aegilops* and evidence for the origin of the A, B, and D genomes of common wheat (*Triticum aestivum*). *Molecular Phylogenetics and Evolution.*, **39**, 70–82.

Pevzner, P.A., Tang, H. and Waterman, M.S. (2001) An eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences.*, **98**, 9748–9753.

Philippe, R., Choulet, F., Paux, E., et al. (2012) Whole genome profiling provides a robust framework for physical mapping and sequencing in the highly complex and repetitive wheat genome. *BMC Genomics.*, **13**, 47.

Phillippy, A.M., Schatz, M.C. and Pop, M. (2008) Genome assembly forensics: finding the elusive mis-assembly. *Genome Biology.*, **9**, R55.

Piegu, B., Guyot, R., Picault, N., et al. (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Research.*, **16**, 1262–1269.

Poole, R.L. (2007) The TAIR database. *Methods in Molecular Biology.*, **406**, 179–212.

Price, A.L., Jones, N.C. and Pevzner, P.A. (2005) *De novo* identification of repeat families in large genomes. *Bioinformatics.*, **21 Suppl 1**, i351–8.

Qi, L., Friebe, B., Wu, J., Gu, Y., Qian, C. and Gill, B.S. (2010) The compact *Brachypodium* genome conserves centromeric regions of a common ancestor with wheat and rice. *Functional & Integrative Genomics.*, **10**, 477–492.

Raats, D., Frenkel, Z., Krugman, T., et al. (2013) The physical map of wheat chromosome 1BS provides insights into its gene space organization and evolution. *Genome Biology.*, **14**, R138.

Ragupathy, R., Rathinavelu, R. and Cloutier, S. (2011) Physical mapping and BAC-end sequence analysis provide initial insights into the flax (*Linum usitatissimum* L.) genome. *BMC Genomics.*, **12**, 217.

Rahman, A. and Pachter, L. (2012) CGAL: computing genome assembly likelihoods. *Genome Biology.*, **14**, R8–R8.

Raychaudhuri, S., Chang, J.T., Sutphin, P.D. and Altman, R.B. (2002) Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Research.*, **12**, 203–214.

Reisner, W., Morton, K.J., Riehn, R., et al. (2005) Statics and dynamics of single DNA molecules confined in nanochannels. *Physical Review Letters.*, **94**, 196101–196101.

Ricchetti, M., Fairhead, C. and Dujon, B. (1999) Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature.*, **402**, 96–100.

Ridgway, N.D. (2016) Analysis of sphingolipid synthesis and transport by metabolic labeling of cultured cells with [(3)H] serine. *Methods in Molecular Biology.*, **1376**, 195–202.

Riethman, H.C., Xiang, Z., Paul, S., Morse, E., Hu, X.L., Flint, J., Chi, H.C., Grady, D.L. and Moyzis, R.K. (2001) Integration of telomere sequences with the draft human genome sequence. *Nature.*, **409**, 948–951.

Rizk, G., Lavenier, D. and Chikhi, R. (2013) DSK: k-mer counting with very low memory usage. *Bioinformatics.*, **29**, 652–653.

Roberts, R.J., Carneiro, M.O. and Schatz, M.C. (2013) The advantages of SMRT sequencing. *Genome Biology.*, **14**, 405.

Rodrigue, S., Materna, A.C., Timberlake, S.C., Blackburn, M.C., Malmstrom, R.R., Alm, E.J. and Chisholm, S.W. (2010) Unlocking short read sequencing for metagenomics J. A. Gilbert, ed. *PloS ONE.*, **5**, e11840.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. and Nyrén, P. (1996) Real-time

DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry.*, **242**, 84–89.

Ronaghi, M., Uhlén, M. and Nyrén, P. (1998) A sequencing method based on real-time pyrophosphate. *Science.*, **281**, 363–365.

Round, E.K., Flowers, S.K. and Richards, E.J. (1997) *Arabidopsis thaliana* centromere regions: genetic map positions and repetitive DNA structure. *Genome Research.*, **7**, 1045–1053.

Roy, R.S., Bhattacharya, D. and Schliep, A. (2014) Turtle: identifying frequent k-mers with cache-efficient algorithms. *Bioinformatics.*, **30**, 1950–1957.

Röder, M.S., Huang, X.-Q. and Börner, A. (2007) Fine mapping of the region on wheat chromosome 7D controlling grain weight. *Functional & Integrative Genomics.*, **8**, 79–86.

Ruan, J., Jiang, L., Chong, Z., et al. (2013) Pseudo-Sanger sequencing: massively parallel production of long and near error-free reads using NGS technology. *BMC Genomics.*, **14**, 711.

Rustgi, S., Shafqat, M.N., Kumar, N., Baenziger, P.S., Ali, M.L., Dweikat, I., Campbell, B.T. and Gill, K.S. (2013) Genetic dissection of yield and its component traits using high-density composite map of wheat chromosome 3A: bridging gaps between QTLs and underlying genes T. Yin, ed. *PloS ONE.*, **8**, e70526.

Rymen, D., Winter, J., Van Hasselt, P.M., et al. (2015) Key features and clinical variability of COG6-CDG. *Molecular Genetics and Metabolism.*, **116**, 163–170.

Safár, J., Bartos, J., Janda, J., et al. (2004) Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. *The Plant Journal.*, **39**, 960–968.

Safár, J., Šimková, H., Kubaláková, M., Cíhalíková, J., Suchánková, P., Bartoš, J. and Doležel, J. (2010) Development of chromosome-specific BAC resources for genomics of bread wheat. *Cytogenetic and Genome Research.*, **129**, 211–223.

Sakai, H., Naito, K., Ogiso-Tanaka, E., et al. (2015) The power of single molecule real-time sequencing technology in the *de novo* assembly of a eukaryotic genome. *Scientific Reports.*, **5**, 16780.

Sakurai, T. and Husimi, Y. (1992) Real-time monitoring of DNA polymerase reactions by a micro ISFET pH sensor. *Analytical Chemistry.*, **64**, 1996–1997.

Salzberg, S.L. and Yorke, J.A. (2005) Beware of mis-assembled genomes. *Bioinformatics.*, **21**, 4320–4321.

Salzberg, S.L., Phillippy, A.M., Zimin, A., et al. (2012) GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Research.*, **22**, 557–567.

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

SanMiguel, P., Tikhonov, A., Jin, Y.K., et al. (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science.*, **274**, 765–768.

Saski, C., Lee, S.-B., Fjellheim, S., et al. (2007) Complete chloroplast genome sequences of

Hordeum vulgare, *Sorghum bicolor* and *Agrostis stolonifera*, and comparative analyses with other grass genomes. *Theoretical and Applied Genetics.*, **115**, 571–590.

Schatz, M.C., Delcher, A.L. and Salzberg, S.L. (2010) Assembly of large genomes using second-generation sequencing. *Genome Research.*, **20**, 1165–1173.

Schmidt, J.A. and Brown, W.J. (2009) Lysophosphatidic acid acyltransferase 3 regulates golgi complex structure and function. *The Journal of Cell Biology.*, **186**, 211–218.

Schmutz, J., Cannon, S.B., Schlueter, J., et al. (2010) Genome sequence of the palaeopolyploid soybean. *Nature.*, **465**, 120–120.

Schmutz, J., McClean, P.E., Mamidi, S., et al. (2014) A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genetics.*, **46**, 707–713.

Schmutz, J., Rokhsar, D., Barry, K., et al. (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature.*, **463**, 763–768.

Schnable, P.S., Ware, D., Fulton, R.S., et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science.*, **326**, 1112–1115.

Schneider, M., Tognolli, M. and Bairoch, A. (2004) The Swiss-Prot protein knowledgebase and ExPASy: providing the plant community with high quality proteomic data and tools. *Plant Physiology and Biochemistry.*, **42**, 1013–1021.

Schwartz, D., Li, X., Hernandez, L., Ramnarain, S., Huff, E. and Wang, Y. (1993) Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science.*, **262**, 110–114.

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

Scott, N.S. and Timmis, J.N. (1984) Homologies between nuclear and plastid DNA in spinach. *Theoretical and Applied Genetics.*, **67**, 279–288.

Senerchia, N., Felber, F., North, B., Sarr, A., Guadagnuolo, R. and Parisod, C. (2015) Differential introgression and reorganization of retrotransposons in hybrid zones between wild wheats. *Molecular Ecology.*, 2518–2528.

Shahmuradov, I.A., Akbarova, Y.Y., Solovyev, V.V. and Aliyev, J.A. (2003) Abundance of plastid DNA insertions in nuclear genomes of rice and *Arabidopsis*. *Plant Molecular Biology.*, **52**, 923–934.

Shearer, L.A., Anderson, L.K., de Jong, H., Smit, S., Goicoechea, J.L., Roe, B.A., Hua, A., Giovannoni, J.J. and Stack, S.M. (2014) Fluorescence in situ hybridization and optical mapping to correct scaffold arrangement in the tomato genome. *G3.*, **4**, 1395–1405.

Shulaev, V., Sargent, D.J., Crowhurst, R.N., et al. (2010) The genome of woodland strawberry (*Fragaria vesca*). *Nature Genetics.*, **43**, 109–116.

Simmonds, J., Scott, P., Leverington-Waite, M., Turner, A.S., Brinton, J., Korzun, V., Snape, J. and Uauy, C. (2014) Identification and independent validation of a stable yield and thousand grain weight QTL on chromosome 6A of hexaploid wheat (*Triticum aestivum* L.). *BMC Plant Biology.*, **14**, 191.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M. and Birol, I. (2009) ABySS: A parallel assembler for short read sequence data. *Genome Research.*, **19**, 1117–1123.

Singh, R.S., Jiang, J. and Gill, B.S. (2006) Current status and the future of fluorescence in situ hybridization (FISH) in plant genome research. *Genome.*, **49**, 1057–1068.

Slater, G.S.C. and Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics.*, **6**, 31.

Soderlund, C., Humphray, S., Dunham, A. and French, L. (2000) Contigs built with fingerprints, markers, and FPC V4.7. *Genome Research.*, **10**, 1772–1787.

Soderlund, C., Longden, I. and Mott, R. (1997) FPC: a system for building contigs from restriction fingerprinted clones. *Bioinformatics.*, **13**, 523–535.

Soldatova, L.N. and King, R.D. (2005) Are the current ontologies in biology good ontologies? *Nature Biotechnology.*, **23**, 1095–1098.

Soltis, D.E., Soltis, P.S. and Tate, J.A. (2004) Advances in the study of polyploidy since plant speciation. *New Phytologist.*, **161**, 173–191.

Somyong, S., Ishikawa, G., Munkvold, J.D., Tanaka, J., Benscher, D., Cho, Y.-G. and Sorrells, M.E. (2014) Fine mapping of a preharvest sprouting QTL interval on chromosome 2B in white wheat. *Theoretical and Applied Genetics.*, **127**, 1843–1855.

Sonnhammer, E.L.L. and Koonin, E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics.*, **18**, 619–620.

Spielmeyer, W., Hyles, J., Joaquim, P., Azanza, F., Bonnett, D., Ellis, M.E., Moore, C. and Richards, R.A. (2007) A QTL on chromosome 6A in bread wheat (*Triticum aestivum*) is associated with longer coleoptiles, greater seedling vigour and final plant height. *Theoretical and Applied Genetics.*, **115**, 59–66.

Stanke, M. and Waack, S. (2003) Gene prediction with a hidden markov model and a new intron submodel. *Bioinformatics.*, **19 Suppl 2**, ii215–25.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. and Morgenstern, B. (2006) AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Research.*, **34**, W435–9.

Stegemann, S., Hartmann, S., Ruf, S. and Bock, R. (2003) High-frequency gene transfer from the chloroplast genome to the nucleus. *Proceedings of the National Academy of Sciences.*, **100**, 8828–8833.

Stein, Ponelies, Musket, McMullen and Weber (1998) Chromosome micro-dissection and region-specific libraries from pachytene chromosomes of maize (*Zea mays* L.). *The Plant Journal.*, **13**, 281–289.

Sternberg, N. (1990) Bacteriophage P1 cloning system for the isolation, amplification, and recovery of DNA fragments as large as 100 kilobase pairs. *Proceedings of the National Academy of Sciences.*, **87**, 103–107.

Stoddart, D., Heron, A.J., Mikhailova, E., Maglia, G. and Bayley, H. (2009) Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings* *of the National Academy of Sciences.*, **106**, 7702–7707.

Strelkov, S.V. and Burkhard, P. (2002) Analysis of alpha-helical coiled coils with the program TWISTER reveals a structural mechanism for stutter compensation. *Journal of Structural Biology.*, **137**, 54–64.

Šimková, H., Cíhalíková, J., Vrána, J., Lysák, M.A. and Doležel, J. (2003) Preparation of HMW DNA from plant nuclei and chromosomes isolated from root tips. *Biologia Plantarum.*, **46**, 369–373.

Šimková, H., Safár, J., Kubaláková, M., et al. (2011) BAC libraries from wheat chromosome 7D: efficient tool for positional cloning of aphid resistance genes. *Journal of Biomedicine and Biotechnology.*, **2011**, 1–11.

Škunca, N., Altenhoff, A. and Dessimoz, C. (2012) Quality of computationally inferred gene ontology annotations L. J. Jensen, ed. *PLoS Computational Biology.*, **8**, e1002533.

Talbert, P.B., Masuelli, R., Tyagi, A.P., Comai, L. and Henikoff, S. (2002) Centromeric localization and adaptive evolution of an *Arabidopsis* histone H3 variant. *The Plant Cell.*, **14**, 1053–1066.

Tamburic-Ilincic, L. (2012) Effect of 3B, 5A and 3A QTL for fusarium head blight resistance on agronomic and quality performance of Canadian winter wheat T. Miedaner, ed. *Plant Breeding*, **131**, 722–727.

Tao, Q. (1998) Cloning and stable maintenance of DNA fragments over 300 kb in *Escherichia coli* with conventional plasmid-based vectors. *Nucleic Acids Research.*, **26**, 4901–4909.

Teague, B., Waterman, M.S., Goldstein, S., et al. (2010) High-resolution human genome structure by single-molecule analysis. *Proceedings of the National Academy of Sciences.*, **107**, 10848–10853.

Tegenfeldt, J.O., Prinz, C., Cao, H., et al. (2004) The dynamics of genomic-length DNA molecules in 100-nm channels. *Proceedings of the National Academy of Sciences.*, **101**, 10979–10983.

Tello-Ruiz, M.K., Stein, J., Wei, S., Youens-Clark, K., Jaiswal, P. and Ware, D. (2016) Gramene: a resource for comparative analysis of plants genomes and pathways. *Methods in Molecular Biology.*, **1374**, 141–163.

Tempel, S. (2012) Using and understanding RepeatMasker. *Methods in Molecular Biology.*, **859**, 29–51.

Tenaillon, M.I., Hollister, J.D. and Gaut, B.S. (2010) A triptych of the evolution of plant transposable elements. *Trends in Plant Science.*, **15**, 471–478.

The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature.*, **408**, 796–815.

Thomas, P.D., Mi, H. and Lewis, S. (2007) Ontology annotation: mapping genomic regions to biological function. *Current Opinion in Chemical Biology*, **11**, 4–11.

Timp, W., Comer, J. and Aksimentiev, A. (2012) DNA base-calling from a Nanopore using a viterbi algorithm. *Biophysical Journal.*, **102**, L37–L39.

Tomkins, J.P., Yu, Y., Miller-Smith, H., Frisch, D.A., Woo, S.S. and Wing, R.A. (1999) A bacterial artificial chromosome library for sugarcane. *Theoretical and Applied Genetics.*, **99**, 419–424.

Trapnell, C., Roberts, A., Goff, L., et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols.*, **7**, 562–578.

Tsilo, T.J., Chao, S., Jin, Y. and Anderson, J.A. (2008) Identification and validation of SSR markers linked to the stem rust resistance gene Sr6 on the short arm of chromosome 2D in wheat. *Theoretical and Applied Genetics.*, **118**, 515–524.

Turki, T. and Roshan, U. (2014) MaxSSmap: a GPU program for mapping divergent short reads to genomes with the maximum scoring subsequence. *BMC Genomics.*, **15**, 969.

Tymms, M.J., Scott, N.S. and Possingham, J.V. (1983) DNA content of *Beta vulgaris* chloroplasts during leaf cell expansion. *Plant Physiology.*, **71**, 785–788.

Usher, C.L., Handsaker, R.E., Esko, T., et al. (2015) Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. *Nature Genetics.*, **47**, 921–925.

Valouev, A., Ichikawa, J., Tonthat, T., et al. (2008) A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Research.*, **18**, 1051–1063.

van Oeveren, J., de Ruiter, M., Jesse, T., et al. (2011) Sequence-based physical mapping of complex genomes by whole genome profiling. *Genome Research.*, **21**, 618–625.

VanBuren, R., Bryant, D., Edger, P.P., et al. (2015) Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature.*, **527**, 508–511.

Varshney, R.K., Chen, W., Li, Y., et al. (2012) Draft genome sequence of pigeonpea (*Cajanus cajan)*, an orphan legume crop of resource-poor farmers. *Nature Biotechnology.*, **30**, 83–89.

Varshney, R.K., Song, C., Saxena, R.K., et al. (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nature Biotechnology.*, **31**, 240–246.

Vielle-Calzada, J.P., Martinez de la Vega, O., Hernandez-Guzman, G., et al. (2009) The Palomero genome suggests metal effects on domestication. *Science.*, **326**, 1078–1078.

Visendi, P., Batley, J. and Edwards, D. (2013) Next generation characterisation of cereal genomes for marker discovery. *Biology.*, **2**, 1357–1377.

Vítámvás, P., Prášil, I.T., Kosová, K., Planchon, S. and Renaut, J. (2011) Analysis of proteome and frost tolerance in chromosome 5A and 5B reciprocal substitution lines between two winter wheats during long-term cold acclimation. *Proteomics.*, **12**, 68–85.

Voiniciuc, C., Günl, M., Schmidt, M.H.-W. and Usadel, B. (2015) Highly branched xylan made by irregular xylem14 and mucilage-related21 links mucilage to *Arabidopsis* seeds. *Plant Physiology.*, **169**, 2481–2495.

Vouzis, P.D. and Sahinidis, N.V. (2011) GPU-BLAST: using graphics processors to accelerate protein sequence alignment. *Bioinformatics.*, **27**, 182–188.

Wang, J., Menon, S., Yamasaki, A., Chou, H.-T., Walz, T., Jiang, Y. and Ferro-Novick, S.

(2013) Ypt1 recruits the Atg1 kinase to the preautophagosomal structure. *Proceedings of the National Academy of Sciences.*, **110**, 9800–9805.

Wang, K., Wang, Z., Li, F., et al. (2012) The draft genome of a diploid cotton *Gossypium raimondii*. *Nature Genetics.*, **44**, 1098–1103.

Wang, X., Wang, H., Wang, J., et al. (2011) The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics.*, **43**, 1035–1039.

Warren, R.L., Sutton, G.G., Jones, S.J.M. and Holt, R.A. (2007) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics.*, **23**, 500–501.

Wei, F., Coe, E., Nelson, W., et al. (2007) Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genetics.*, **3**, e123.

Wicker, T., Buchmann, J.P. and Keller, B. (2010) Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Research.*, **20**, 1229–1237.

Wicker, T., Matthews, D.E. and Keller, B. (2002) TREP: a database for *Triticeae* repetitive elements. *Trends in Plant Science.*, **7**, 561–562.

Wicker, T., Mayer, K.F.X., Gundlach, H., et al. (2011) Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *The Plant Cell.*, **23**, 1706–1718.

Wickham, H. (2009) *ggplot2*, New York, NY: Springer New York.

Wise, R.P., Graner, A., Sato, K., et al. (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature.*, **491**, 711–716.

Wolf, E., Kim, P.S. and Berger, B. (1997) MultiCoil: A program for predicting two- and three-stranded coiled coils. *Protein Science.*, **6**, 1179–1189.

Wu, T.D. and Nacu, S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics.*, **26**, 873–881.

Wu, X., Steet, R.A., Bohorov, O., Bakker, J., Newell, J., Krieger, M., Spaapen, L., Kornfeld, S. and Freeze, H.H. (2004) Mutation of the COG complex subunit gene COG7 causes a lethal congenital disorder. *Nature Medicine.*, **10**, 518–523.

Xiao, M., Phong, A., Ha, C., et al. (2007) Rapid DNA mapping by fluorescent single molecule detection. *Nucleic Acids Research.*, **35**, e16–e16.

Xu, X., Pan, S., Cheng, S., et al. (2011) Genome sequence and analysis of the tuber crop potato. *Nature.*, **475**, 189–195.

Xu, Z., Sun, S., Covaleda, L., Ding, K., Zhang, A., Wu, C., Scheuring, C. and Zhang, H.-B. (2004) Genome physical mapping with large-insert bacterial clones by fingerprint analysis: methodologies, source clone genome coverage, and contig map quality. *Genomics.*, **84**, 941–951.

Yen, W.-L., Shintani, T., Nair, U., Cao, Y., Richardson, B.C., Li, Z., Hughson, F.M., Baba, M. and Klionsky, D.J. (2010) The conserved oligomeric golgi complex is involved in double-membrane vesicle formation during autophagy. *The Journal of Cell Biology.*, **188**, 101–114.

Yin, T., Cook, D. and Lawrence, M. (2012) ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biology.*, **13**, R77.

Yon Rhee, S., Wood, V., Dolinski, K. and Drăghici, S. (2008) Use and misuse of the gene ontology annotations. *Nature Reviews Genetics.*, **9**, 509–515.

Young, N.D., Debellé, F., Oldroyd, G.E.D., et al. (2011) The medicago genome provides insight into the evolution of rhizobial symbioses. *Nature.*, 520–524.

Yu, J. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science.*, **296**, 79–92.

Yu, X. and Gabriel, A. (1999) Patching broken chromosomes with extranuclear cellular DNA. *Molecular Cell.*, **4**, 873–881.

Zehetner, G. (2003) OntoBlast function: from sequence similarities directly to potential functional annotations by ontology terms. *Nucleic Acids Research.*, **31**, 3799–3803.

Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for *de novo* short read assembly using *de Bruijn* graphs. *Genome Research.*, **18**, 821–829.

Zhang, Baocai, Liu, X., Qian, Q., Liu, L., Dong, G., Xiong, G., Zeng, D., Zhou, Y. and Delmer, D.P. (2011) Golgi nucleotide sugar transporter modulates cell wall biosynthesis and plant growth in rice. *Proceedings of the National Academy of Sciences.*, **108**, 5110–5115.

Zhang, Gengyun, Liu, X., Quan, Z., et al. (2012) Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nature Biotechnology.*, **30**, 549–554.

Zhang, Guizhi, Wang, Y., Guo, Y., Zhao, Y., Kong, F. and Li, S. (2015) Characterization and mapping of QTLs on chromosome 2D for grain size and yield traits using a mutant line induced by EMS in wheat. *The Crop Journal.*, **3**, 135–144.

Zhang, Hong, Tan, E., Suzuki, Y., et al. (2014) Dramatic improvement in genome assembly achieved using doubled-haploid genomes. *Scientific Reports.*, **4**, 6780.

Zhang, Qingpeng, Pell, J., Canino-Koning, R., Howe, A.C. and Brown, C.T. (2014) These are not the K-mers you are looking for: efficient online K-mer counting using a probabilistic data structure D. Zhu, ed. *PloS ONE.*, **9**, e101271.

Zhang, Wei, Ciclitira, P. and Messing, J. (2014) PacBio sequencing of gene families - a case study with wheat gluten genes. *Gene.*, **533**, 541–546.

Zhang, Xiaojun, Scheuring, C.F., Zhang, M., et al. (2009) A BAC/BIBAC-based physical map of chickpea, *Cicer arietinum* L. *BMC Genomics.*, **11**, 501–501.

Zhao, K. and Chu, X. (2014) G-BLASTN: accelerating nucleotide alignment by graphics processors. *Bioinformatics.*, **30**, 1384–1391.

Zhou, S. (2003) Whole-genome shotgun optical mapping of *Rhodobacter sphaeroides* strain 2.4.1 and its use for whole-genome shotgun sequence assembly. *Genome Research.*, **13**, 2142–2151.

Zhou, S., Bechner, M.C., Place, M., et al. (2007) Validation of rice genome sequence by optical mapping. *BMC Genomics.*, **8**, 278.

Zhou, S., Wei, F., Nguyen, J., et al. (2009) A single molecule scaffold for the maize genome J. R. Ecker, ed. *PLoS Genetics.*, **5**, e1000711.

Zhou, Y., Xia, X., He, Z., Li, X., Li, Z. and Liu, D. (2013) Fine mapping of leaf rust resistance gene LrZH84 using expressed sequence tag and sequence-tagged site markers, and allelism with other genes on wheat chromosome 1B. *Phytopathology.*, **103**, 169–174.

Zimin, A., Stevens, K.A., Crepeau, M.W., et al. (2014) Sequencing and assembly of the 22-Gb loblolly pine genome. *Genetics.*, **196**, 875–890.

Zody, M.C., Garber, M., Adams, D.J., et al. (2006) DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage. *Nature.*, **440**, 1045–1049.

Appendix 1

BACs Assembly pipeline (BacA)

“BacA” is a perl wrapper for the Sassy assembler.

Parameters and linux commands in this document are in bold and italics

Dependencies

Install or check that you have the following installed on your linux system:

1. Perl plus non-standard libraries (available from CPAN)
    Config::IniFiles
    Bioperl version 1.6.9 and above
    Perl version 5.14.2 and above
2. fastx-toolkit version 0.0.13.2
3. Blast+ version 2.2.28
4. SASSY version 1.0.1 (https://github.com/minillinim/SaSSY)
5. Seqclean
6. GNU Awk version 3.1.8
7. GNU sed version 4.1.5
8. Soap2 version 2.21
9. DSK version 1.5798

Pipeline setup

Unzip and place the BaCA folder into a suitable location on your linux system. Edit fields in the BACA.config file located at "BACS_Assembly_pipeline/bin/ “, according to your system setup. There are several sections in the config file: GENERAL, PROG and ADVANCED.

The GENERAL section of the config file contains absolute paths to all your datasets such paired end reads, mate pair reads and BES.

For your paired end reads for assembly, the reads for each pool or dataset are organized into sub folders, with each subfolder containing only two files, the read A and read B fastq files. The parent folder for all the sub folders is then entered in the config as below. In this example, the parent folder is 7DS

INPUT_DIR=/home/visendi/7DS

The location of BES and MP datasets is also denoted as below;

INPUT_BES=/home/visendi/7DS_BES

INPUT_MP=/home/visendi/7DS_MP

The PROG section of the config file contains absolute paths to all executable programs called by the pipeline. Example.

SOAP=/opt/bin

The ADVANCED section of the config file contains specific assembly and pipeline parameters, detailed in the pipeline's README file.

In the Pipeline's installed directory, under "BACS_Assembly_pipeline/data" directory, you will find multiple fasta files "vector.fasta and Ecoli.fasta"
These fasta files contain the vector and *E coli* reference genomes used in the cloning and maintaining of BACs. Edit these according to your sequencing project.

Running pipeline

Cd into bin directory of the installed pipeline "BACS_Assembly_pipeline/bin"
Run pipeline with the configuration file as first argument:

BacA BACA.conf

Outputs

If your project directory was named chr1, you will find a directory named:

"Chr1.your_config_name.TL.NOF.EOF.Assemblies"

TL, NOF and EOF are the trimmed lengths, Naïve offsets and Extension offset parameters used in the assembly process.
This directory will have a directory or sub directories with: assembled contigs and pipeline log files for debugging purposes.
Assembly statistics and BES mapping files will be found in the pipeline directory "BACS_Assembly_pipeline/"

Plotting BES mappings

To visualize the mapping positions of BES on contigs and scaffolds, an R script (Plot_BACs.R) was used to automate this process.

Dependencies

Install or check that you have the following installed on your linux system.

- R version 3.1.2
- R Libraries.
  - Bioconductor version 3.1
  - Ggbio version 1.16
  - GenomicRanges 1.16

Once assemblies or scaffolding has been done, the BES plotting file for either the assembly or scaffolds will be available under the pipeline directory “BACS_Assembly_pipeline/”. The BES mapping files named as “Configuration_filename_BES_mappings.txt” can then be mapped as below on the command line by directly executing the R script.

```
. /Plot_BACs.R   Configuration_filename_BES_mappings.txt
```

Each of the assemblies BES mappings will be plotted and saved as a PNG file.

Appendix 2

		Sequencing Strategy: Table represents 12 plates each with 384 cells. Each cell (eg cell A,1) contained a pool of four non overlapping BACs.											
Plate													
P1		1	2	3	4	5	6	7	8	9	10	11	12
	A	1-A1-C1-E1-G1	1-A9-C9-E9-G9	1-A17-C17-E17-G17	1-A2-C2-E2-G2	1-A10-C10-E10-G10	1-A18-C18-E18-G18	1-B1-D1-F1-H1	1-B9-D9-F9-H9	1-B17-D17-F17-H17	4-B2-D2-F2-H2	4-B10-D10-F10-H10	4-B18-D18-F18-H18
	B	1-I1-K1-M1-O1	1-I9-K9-M9-O9	1-I17-K17-M17-O17	1-I2-K2-M2-O2	1-I10-K10-M10-O10	1-I18-K18-M18-O18	1-J1-L1-N1-P1	1-J9-L9-N9-P9	1-J17-L17-N17-P17	4-J2-L2-N2-P2	4-J10-L10-N10-P10	4-J18-L18-N18-P18
	C	1-A3-C3-E3-G3	1-A11-C11-E11-G11	1-A19-C19-E19-G19	1-A4-C4-E4-G4	1-A12-C12-E12-G12	1-A20-C20-E20-G20	1-B3-D3-F3-H3	1-B11-D11-F11-H11	1-B19-D19-F19-H19	4-B4-D4-F4-H4	4-B12-D12-F12-H12	4-B20-D20-F20-H20
	D	1-I3-K3-M3-O3	1-I11-K11-M11-O11	1-I19-K19-M19-O19	1-I4-K4-M4-O4	1-I12-K12-M12-O12	1-I20-K20-M20-O20	1-J3-L3-N3-P3	1-J11-L11-N11-P11	1-J19-L19-N19-P19	4-J4-L4-N4-P4	4-J12-L12-N12-P12	4-J20-L20-N20-P20
	E	1-A5-C5-E5-G5	1-A13-C13-E13-G13	1-A21-C21-E21-G21	1-A6-C6-E6-G6	1-A14-C14-E14-G14	1-A22-C22-E22-G22	1-B5-D5-F5-H5	1-B13-D13-F13-H13	1-B21-D21-F21-H21	4-B6-D6-F6-H6	4-B14-D14-F14-H14	4-B22-D22-F22-H22
	F	1-I5-K5-M5-O5	1-I13-K13-M13-O13	1-I21-K21-M21-O21	1-I6-K6-M6-O6	1-I14-K14-M14-O14	1-I22-K22-M22-O22	1-J5-L5-N5-P5	1-J13-L13-N13-P13	1-J21-L21-N21-P21	4-J6-L6-N6-P6	4-J14-L14-N14-P14	4-J22-L22-N22-P22
	G	1-A7-C7-E7-G7	1-A15-C15-E15-G15	1-A23-C23-E23-G23	1-A8-C8-E8-G8	1-A16-C16-E16-G16	1-A24-C24-E24-G24	1-B7-D7-F7-H7	1-B15-D15-F15-H15	1-B23-D23-F23-H23	4-B8-D8-F8-H8	4-B16-D16-F16-H16	4-B24-D24-F24-H24
	H	1-I7-K7-M7-O7	1-I15-K15-M15-O15	1-I23-K23-M23-O23	1-I8-K8-M8-O8	1-I16-K16-M16-O16	1-I24-K24-M24-O24	1-J7-L7-N7-P7	1-J15-L15-N15-P15	1-J23-L23-N23-P23	4-J8-L8-N8-P8	4-J16-L16-N16-P16	4-J24-L24-N24-P24
P2		1	2	3	4	5	6	7	8	9	10	11	12
	A	2-A1-C1-E1-G1	2-A9-C9-E9-G9	2-A17-C17-E17-G17	2-A2-C2-E2-G2	2-A10-C10-E10-G10	2-A18-C18-E18-G18	2-B1-D1-F1-H1	2-B9-D9-F9-H9	2-B17-D17-F17-H17	2-B2-D2-F2-H2	2-B10-D10-F10-H10	2-B18-D18-F18-H18
	B	2-I1-K1-M1-O1	2-I9-K9-M9-O9	2-I17-K17-M17-O17	2-I2-K2-M2-O2	2-I10-K10-M10-O10	2-I18-K18-M18-O18	2-J1-L1-N1-P1	2-J9-L9-N9-P9	2-J17-L17-N17-P17	2-J2-L2-N2-P2	2-J10-L10-N10-P10	2-J18-L18-N18-P18
	C	2-A3-C3-E3-G3	2-A11-C11-E11-G11	2-A19-C19-E19-G19	2-A4-C4-E4-G4	2-A12-C12-E12-G12	2-A20-C20-E20-G20	2-B3-D3-F3-H3	2-B11-D11-F11-H11	2-B19-D19-F19-H19	2-B4-D4-F4-H4	2-B12-D12-F12-H12	2-B20-D20-F20-H20
	D	2-I3-K3-M3-O3	2-I11-K11-M11-O11	2-I19-K19-M19-O19	2-I4-K4-M4-O4	2-I12-K12-M12-O12	2-I20-K20-M20-O20	2-J3-L3-N3-P3	2-J11-L11-N11-P11	2-J19-L19-N19-P19	2-J4-L4-N4-P4	2-J12-L12-N12-P12	2-J20-L20-N20-P20
	E	2-A5-C5-E5-G5	2-A13-C13-E13-G13	2-A21-C21-E21-G21	2-A6-C6-E6-G6	2-A14-C14-E14-G14	2-A22-C22-E22-G22	2-B5-D5-F5-H5	2-B13-D13-F13-H13	2-B21-D21-F21-H21	2-B6-D6-F6-H6	2-B14-D14-F14-H14	2-B22-D22-F22-H22
	F	2-I5-K5-M5-O5	2-I13-K13-M13-O13	2-I21-K21-M21-O21	2-I6-K6-M6-O6	2-I14-K14-M14-O14	2-I22-K22-M22-O22	2-J5-L5-N5-P5	2-J13-L13-N13-P13	2-J21-L21-N21-P21	2-J6-L6-N6-P6	2-J14-L14-N14-P14	2-J22-L22-N22-P22
	G	2-A7-C7-E7-G7	2-A15-C15-E15-G15	2-A23-C23-E23-G23	2-A8-C8-E8-G8	2-A16-C16-E16-G16	2-A24-C24-E24-G24	2-B7-D7-F7-H7	2-B15-D15-F15-H15	2-B23-D23-F23-H23	2-B8-D8-F8-H8	2-B16-D16-F16-H16	2-B24-D24-F24-H24
	H	2-I7-K7-M7-O7	2-I15-K15-M15-O15	2-I23-K23-M23-O23	2-I8-K8-M8-O8	2-I16-K16-M16-O16	2-I24-K24-M24-O24	2-J7-L7-N7-P7	2-J15-L15-N15-P15	2-J23-L23-N23-P23	2-J8-L8-N8-P8	2-J16-L16-N16-P16	2-J24-L24-N24-P24
P3		1	2	3	4	5	6	7	8	9	10	11	12
	A	3-A2-C2-E2-G2	3-A1-C1-E1-G1	3-A9-C9-E9-G9	3-A17-C17-E17-G17	3-B2-D2-F2-H2	3-B10-D10-F10-H10	3-B18-D18-F18-H18	4-B1-D1-F1-H1	4-B9-D9-F9-H9	4-B17-D17-F17-H17	-	-
	B	3-I2-K2-M2-O2	3-I1-K1-M1-O1	3-I9-K9-M9-O9	3-I17-K17-M17-O17	3-J2-L2-N2-P2	3-J10-L10-N10-P10	3-J18-L18-N18-P18	4-J1-L1-N1-P1	4-J9-L9-N9-P9	4-J17-L17-N17-P17	-	-
	C	3-A4-C4-E4-G4	3-A3-C3-E3-G3	3-A11-C11-E11-G11	3-A19-C19-E19-G19	3-B4-D4-F4-H4	3-B12-D12-F12-H12	3-B20-D20-F20-H20	4-B3-D3-F3-H3	4-B11-D11-F11-H11	4-B19-D19-F19-H19	-	-
	D	3-I4-K4-M4-O4	3-I3-K3-M3-O3	3-I11-K11-M11-O11	3-I19-K19-M19-O19	3-J4-L4-N4-P4	3-J12-L12-N12-P12	3-J20-L20-N20-P20	4-J3-L3-N3-P3	4-J11-L11-N11-P11	4-J19-L19-N19-P19	-	-
	E	3-A6-C6-E6-G6	3-A5-C5-E5-G5	3-A13-C13-E13-G13	3-A21-C21-E21-G21	3-B6-D6-F6-H6	3-B14-D14-F14-H14	3-B22-D22-F22-H22	4-B5-D5-F5-H5	4-B13-D13-F13-H13	4-B21-D21-F21-H21	-	-
	F	3-I6-K6-M6-O6	3-I5-K5-M5-O5	3-I13-K13-M13-O13	3-I21-K21-M21-O21	3-J6-L6-N6-P6	3-J14-L14-N14-P14	3-J22-L22-N22-P22	4-J5-L5-N5-P5	4-J13-L13-N13-P13	4-J21-L21-N21-P21	-	-
	G	3-A8-C8-E8-G8	3-A7-C7-E7-G7	3-A15-C15-E15-G15	3-A23-C23-E23-G23	3-B8-D8-F8-H8	3-B16-D16-F16-H16	3-B24-D24-F24-H24	4-B7-D7-F7-H7	4-B15-D15-F15-H15	4-B23-D23-F23-H23	-	-
	H	3-I8-K8-M8-O8	3-I7-K7-M7-O7	3-I15-K15-M15-O15	3-I23-K23-M23-O23	3-J8-L8-N8-P8	3-J16-L16-N16-P16	3-J24-L24-N24-P24	4-J7-L7-N7-P7	4-J15-L15-N15-P15	4-J23-L23-N23-P23	-	-
P4		1	2	3	4	5	6	7	8	9	10	11	12
	A	1-B2-D2-F2-H2	1-B10-D10-F10-H10	1-B18-D18-F18-H18	3-B1-D1-F1-H1	3-B9-D9-F9-H9	3-B17-D17-F17-H17	4-A9-C9-E9-G9	4-A17-C17-E17-G17	4-A2-C2-E2-G2	4-A10-C10-E10-G10	4-A18-C18-E18-G18	-
	B	1-J2-L2-N2-P2	1-J10-L10-N10-P10	1-J18-L18-N18-P18	3-J1-L1-N1-P1	3-J9-L9-N9-P9	3-J17-L17-N17-P17	4-I9-K9-M9-O9	4-I17-K17-M17-O17	4-I2-K2-M2-O2	4-I10-K10-M10-O10	4-I18-K18-M18-O18	-
	C	1-B4-D4-F4-H4	1-B12-D12-F12-H12	1-B20-D20-F20-H20	3-B3-D3-F3-H3	3-B11-D11-F11-H11	3-B19-D19-F19-H19	4-A11-C11-E11-G11	4-A19-C19-E19-G19	4-A4-C4-E4-G4	4-A12-C12-E12-G12	4-A20-C20-E20-G20	-
	D	1-J4-L4-N4-P4	1-J12-L12-N12-P12	1-J20-L20-N20-P20	3-J3-L3-N3-P3	3-J11-L11-N11-P11	3-J19-L19-N19-P19	4-I11-K11-M11-O11	4-I19-K19-M19-O19	4-I4-K4-M4-O4	4-I12-K12-M12-O12	4-I20-K20-M20-O20	-
	E	1-B6-D6-F6-H6	1-B14-D14-F14-H14	1-B22-D22-F22-H22	3-B5-D5-F5-H5	3-B13-D13-F13-H13	3-B21-D21-F21-H21	4-A13-C13-E13-G13	4-A21-C21-E21-G21	4-A6-C6-E6-G6	4-A14-C14-E14-G14	4-A22-C22-E22-G22	-
	F	1-J6-L6-N6-P6	1-J14-L14-N14-P14	1-J22-L22-N22-P22	3-J5-L5-N5-P5	3-J13-L13-N13-P13	3-J21-L21-N21-P21	4-I13-K13-M13-O13	4-I21-K21-M21-O21	4-I6-K6-M6-O6	4-I14-K14-M14-O14	4-I22-K22-M22-O22	-
	G	1-B8-D8-F8-H8	1-B16-D16-F16-H16	1-B24-D24-F24-H24	3-B7-D7-F7-H7	3-B15-D15-F15-H15	3-B23-D23-F23-H23	4-A15-C15-E15-G15	4-A23-C23-E23-G23	4-A8-C8-E8-G8	4-A16-C16-E16-G16	4-A24-C24-E24-G24	-
	H	1-J8-L8-N8-P8	1-J16-L16-N16-P16	1-J24-L24-N24-P24	3-J7-L7-N7-P7	3-J15-L15-N15-P15	3-J23-L23-N23-P23	4-I15-K15-M15-O15	4-I23-K23-M23-O23	4-I8-K8-M8-O8	4-I16-K16-M16-O16	4-I24-K24-M24-O24	-
P5		1	2	3	4	5	6	7	8	9	10	11	12
	A	5-A1-C1-E1-G1	5-A9-C9-E9-G9	5-A17-C17-E17-G17	5-A2-C2-E2-G2	5-A10-C10-E10-G10	5-A18-C18-E18-G18	5-B1-D1-F1-H1	5-B9-D9-F9-H9	5-B17-D17-F17-H17	5-B2-D2-F2-H2	5-B10-D10-F10-H10	5-B18-D18-F18-H18
	B	5-I1-K1-M1-O1	5-I9-K9-M9-O9	5-I17-K17-M17-O17	5-I2-K2-M2-O2	5-I10-K10-M10-O10	5-I18-K18-M18-O18	5-J1-L1-N1-P1	5-J9-L9-N9-P9	5-J17-L17-N17-P17	5-J2-L2-N2-P2	5-J10-L10-N10-P10	5-J18-L18-N18-P18
	C	5-A3-C3-E3-G3	5-A11-C11-E11-G11	5-A19-C19-E19-G19	5-A4-C4-E4-G4	5-A12-C12-E12-G12	5-A20-C20-E20-G20	5-B3-D3-F3-H3	5-B11-D11-F11-H11	5-B19-D19-F19-H19	5-B4-D4-F4-H4	5-B12-D12-F12-H12	5-B20-D20-F20-H20
	D	5-I3-K3-M3-O3	5-I11-K11-M11-O11	5-I19-K19-M19-O19	5-I4-K4-M4-O4	5-I12-K12-M12-O12	5-I20-K20-M20-O20	5-J3-L3-N3-P3	5-J11-L11-N11-P11	5-J19-L19-N19-P19	5-J4-L4-N4-P4	5-J12-L12-N12-P12	5-J20-L20-N20-P20

	E	5-A5-C5-E5-G5	5-A13-C13-E13-G13	5-A21-C21-E21-G21	5-A6-C6-E6-G6	5-A14-C14-E14-G14	5-A22-C22-E22-G22	5-B5-D5-F5-H5	5-B13-D13-F13-H13	5-B21-D21-F21-H21	5-B6-D6-F6-H6	5-B14-D14-F14-H14	5-B22-D22-F22-H22
	F	5-I5-K5-M5-O5	5-I13-K13-M13-O13	5-I21-K21-M21-O21	5-I6-K6-M6-O6	5-I14-K14-M14-O14	5-I22-K22-M22-O22	5-J5-L5-N5-P5	5-J13-L13-N13-P13	5-J21-L21-N21-P21	5-J6-L6-N6-P6	5-J14-L14-N14-P14	5-J22-L22-N22-P22
	G	5-A7-C7-E7-G7	5-A15-C15-E15-G15	5-A23-C23-E23-G23	5-A8-C8-E8-G8	5-A16-C16-E16-G16	5-A24-C24-E24-G24	5-B7-D7-F7-H7	5-B15-D15-F15-H15	5-B23-D23-F23-H23	5-B8-D8-F8-H8	5-B16-D16-F16-H16	5-B24-D24-F24-H24
	H	5-I7-K7-M7-O7	5-I15-K15-M15-O15	5-I23-K23-M23-O23	5-I8-K8-M8-O8	5-I16-K16-M16-O16	5-I24-K24-M24-O24	5-J7-L7-N7-P7	5-J15-L15-N15-P15	5-J23-L23-N23-P23	5-J8-L8-N8-P8	5-J16-L16-N16-P16	5-J24-L24-N24-P24
P6		1	2	3	4	5	6	7	8	9	10	11	12
	A	6-A1-C1-E1-G1	6-A9-C9-E9-G9	6-A17-C17-E17-G17	6-A2-C2-E2-G2	6-A10-C10-E10-G10	6-A18-C18-E18-G18	6-B1-D1-F1-H1	6-B9-D9-F9-H9	6-B17-D17-F17-H17	6-B2-D2-F2-H2	6-B10-D10-F10-H10	6-B18-D18-F18-H18
	B	6-I1-K1-M1-O1	6-I9-K9-M9-O9	6-I17-K17-M17-O17	6-I2-K2-M2-O2	6-I10-K10-M10-O10	6-I18-K18-M18-O18	6-J1-L1-N1-P1	6-J9-L9-N9-P9	6-J17-L17-N17-P17	6-J2-L2-N2-P2	6-J10-L10-N10-P10	6-J18-L18-N18-P18
	C	6-A3-C3-E3-G3	6-A11-C11-E11-G11	6-A19-C19-E19-G19	6-A4-C4-E4-G4	6-A12-C12-E12-G12	6-A20-C20-E20-G20	6-B3-D3-F3-H3	6-B11-D11-F11-H11	6-B19-D19-F19-H19	6-B4-D4-F4-H4	6-B12-D12-F12-H12	6-B20-D20-F20-H20
	D	6-I3-K3-M3-O3	6-I11-K11-M11-O11	6-I19-K19-M19-O19	6-I4-K4-M4-O4	6-I12-K12-M12-O12	6-I20-K20-M20-O20	6-J3-L3-N3-P3	6-J11-L11-N11-P11	6-J19-L19-N19-P19	6-J4-L4-N4-P4	6-J12-L12-N12-P12	6-J20-L20-N20-P20
	E	6-A5-C5-E5-G5	6-A13-C13-E13-G13	6-A21-C21-E21-G21	6-A6-C6-E6-G6	6-A14-C14-E14-G14	6-A22-C22-E22-G22	6-B5-D5-F5-H5	6-B13-D13-F13-H13	6-B21-D21-F21-H21	6-B6-D6-F6-H6	6-B14-D14-F14-H14	6-B22-D22-F22-H22
	F	6-I5-K5-M5-O5	6-I13-K13-M13-O13	6-I21-K21-M21-O21	6-I6-K6-M6-O6	6-I14-K14-M14-O14	6-I22-K22-M22-O22	6-J5-L5-N5-P5	6-J13-L13-N13-P13	6-J21-L21-N21-P21	6-J6-L6-N6-P6	6-J14-L14-N14-P14	6-J22-L22-N22-P22
	G	6-A7-C7-E7-G7	6-A15-C15-E15-G15	6-A23-C23-E23-G23	6-A8-C8-E8-G8	6-A16-C16-E16-G16	6-A24-C24-E24-G24	6-B7-D7-F7-H7	6-B15-D15-F15-H15	6-B23-D23-F23-H23	6-B8-D8-F8-H8	6-B16-D16-F16-H16	6-B24-D24-F24-H24
	H	6-I7-K7-M7-O7	6-I15-K15-M15-O15	6-I23-K23-M23-O23	6-I8-K8-M8-O8	6-I16-K16-M16-O16	6-I24-K24-M24-O24	6-J7-L7-N7-P7	6-J15-L15-N15-P15	6-J23-L23-N23-P23	6-J8-L8-N8-P8	6-J16-L16-N16-P16	6-J24-L24-N24-P24
P7		1	2	3	4	5	6	7	8	9	10	11	12
	A	7-A1-C1-E1-G1	7-A9-C9-E9-G9	7-A17-C17-E17-G17	7-A2-C2-E2-G2	7-A10-C10-E10-G10	7-A18-C18-E18-G18	7-B1-D1-F1-H1	7-B9-D9-F9-H9	7-B17-D17-F17-H17	7-B2-D2-F2-H2	7-B10-D10-F10-H10	7-B18-D18-F18-H18
	B	7-I1-K1-M1-O1	7-I9-K9-M9-O9	7-I17-K17-M17-O17	7-I2-K2-M2-O2	7-I10-K10-M10-O10	7-I18-K18-M18-O18	7-J1-L1-N1-P1	7-J9-L9-N9-P9	7-J17-L17-N17-P17	7-J2-L2-N2-P2	7-J10-L10-N10-P10	7-J18-L18-N18-P18
	C	7-A3-C3-E3-G3	7-A11-C11-E11-G11	7-A19-C19-E19-G19	7-A4-C4-E4-G4	7-A12-C12-E12-G12	7-A20-C20-E20-G20	7-B3-D3-F3-H3	7-B11-D11-F11-H11	7-B19-D19-F19-H19	7-B4-D4-F4-H4	7-B12-D12-F12-H12	7-B20-D20-F20-H20
	D	7-I3-K3-M3-O3	7-I11-K11-M11-O11	7-I19-K19-M19-O19	7-I4-K4-M4-O4	7-I12-K12-M12-O12	7-I20-K20-M20-O20	7-J3-L3-N3-P3	7-J11-L11-N11-P11	7-J19-L19-N19-P19	7-J4-L4-N4-P4	7-J12-L12-N12-P12	7-J20-L20-N20-P20
	E	7-A5-C5-E5-G5	7-A13-C13-E13-G13	7-A21-C21-E21-G21	7-A6-C6-E6-G6	7-A14-C14-E14-G14	7-A22-C22-E22-G22	7-B5-D5-F5-H5	7-B13-D13-F13-H13	7-B21-D21-F21-H21	7-B6-D6-F6-H6	7-B14-D14-F14-H14	7-B22-D22-F22-H22
	F	7-I5-K5-M5-O5	7-I13-K13-M13-O13	7-I21-K21-M21-O21	7-I6-K6-M6-O6	7-I14-K14-M14-O14	7-I22-K22-M22-O22	7-J5-L5-N5-P5	7-J13-L13-N13-P13	7-J21-L21-N21-P21	7-J6-L6-N6-P6	7-J14-L14-N14-P14	7-J22-L22-N22-P22
	G	7-A7-C7-E7-G7	7-A15-C15-E15-G15	7-A23-C23-E23-G23	7-A8-C8-E8-G8	7-A16-C16-E16-G16	7-A24-C24-E24-G24	7-B7-D7-F7-H7	7-B15-D15-F15-H15	7-B23-D23-F23-H23	7-B8-D8-F8-H8	7-B16-D16-F16-H16	7-B24-D24-F24-H24
	H	7-I7-K7-M7-O7	7-I15-K15-M15-O15	7-I23-K23-M23-O23	7-I8-K8-M8-O8	7-I16-K16-M16-O16	7-I24-K24-M24-O24	7-J7-L7-N7-P7	7-J15-L15-N15-P15	7-J23-L23-N23-P23	7-J8-L8-N8-P8	7-J16-L16-N16-P16	7-J24-L24-N24-P24
P8		1	2	3	4	5	6	7	8	9	10	11	12
	A	8-A1-C1-E1-G1	8-A9-C9-E9-G9	8-A17-C17-E17-G17	8-A2-C2-E2-G2	8-A10-C10-E10-G10	8-A18-C18-E18-G18	8-B1-D1-F1-H1	8-B9-D9-F9-H9	8-B17-D17-F17-H17	8-B2-D2-F2-H2	8-B10-D10-F10-H10	8-B18-D18-F18-H18
	B	8-I1-K1-M1-O1	8-I9-K9-M9-O9	8-I17-K17-M17-O17	8-I2-K2-M2-O2	8-I10-K10-M10-O10	8-I18-K18-M18-O18	8-J1-L1-N1-P1	8-J9-L9-N9-P9	8-J17-L17-N17-P17	8-J2-L2-N2-P2	8-J10-L10-N10-P10	8-J18-L18-N18-P18
	C	8-A3-C3-E3-G3	8-A11-C11-E11-G11	8-A19-C19-E19-G19	8-A4-C4-E4-G4	8-A12-C12-E12-G12	8-A20-C20-E20-G20	8-B3-D3-F3-H3	8-B11-D11-F11-H11	8-B19-D19-F19-H19	8-B4-D4-F4-H4	8-B12-D12-F12-H12	8-B20-D20-F20-H20
	D	8-I3-K3-M3-O3	8-I11-K11-M11-O11	8-I19-K19-M19-O19	8-I4-K4-M4-O4	8-I12-K12-M12-O12	8-I20-K20-M20-O20	8-J3-L3-N3-P3	8-J11-L11-N11-P11	8-J19-L19-N19-P19	8-J4-L4-N4-P4	8-J12-L12-N12-P12	8-J20-L20-N20-P20
	E	8-A5-C5-E5-G5	8-A13-C13-E13-G13	8-A21-C21-E21-G21	8-A6-C6-E6-G6	8-A14-C14-E14-G14	8-A22-C22-E22-G22	8-B5-D5-F5-H5	8-B13-D13-F13-H13	8-B21-D21-F21-H21	8-B6-D6-F6-H6	8-B14-D14-F14-H14	8-B22-D22-F22-H22
	F	8-I5-K5-M5-O5	8-I13-K13-M13-O13	8-I21-K21-M21-O21	8-I6-K6-M6-O6	8-I14-K14-M14-O14	8-I22-K22-M22-O22	8-J5-L5-N5-P5	8-J13-L13-N13-P13	8-J21-L21-N21-P21	8-J6-L6-N6-P6	8-J14-L14-N14-P14	8-J22-L22-N22-P22
	G	8-A7-C7-E7-G7	8-A15-C15-E15-G15	8-A23-C23-E23-G23	8-A8-C8-E8-G8	8-A16-C16-E16-G16	8-A24-C24-E24-G24	8-B7-D7-F7-H7	8-B15-D15-F15-H15	8-B23-D23-F23-H23	8-B8-D8-F8-H8	8-B16-D16-F16-H16	8-B24-D24-F24-H24
	H	8-I7-K7-M7-O7	8-I15-K15-M15-O15	8-I23-K23-M23-O23	8-I8-K8-M8-O8	8-I16-K16-M16-O16	8-I24-K24-M24-O24	8-J7-L7-N7-P7	8-J15-L15-N15-P15	8-J23-L23-N23-P23	8-J8-L8-N8-P8	8-J16-L16-N16-P16	8-J24-L24-N24-P24
P9		1	2	3	4	5	6	7	8	9	10	11	12
	A	9-A1-C1-E1-G1	9-A9-C9-E9-G9	9-A17-C17-E17-G17	9-A2-C2-E2-G2	9-A10-C10-E10-G10	9-A18-C18-E18-G18	9-B1-D1-F1-H1	9-B9-D9-F9-H9	9-B17-D17-F17-H17	9-B2-D2-F2-H2	9-B10-D10-F10-H10	9-B18-D18-F18-H18
	B	9-I1-K1-M1-O1	9-I9-K9-M9-O9	9-I17-K17-M17-O17	9-I2-K2-M2-O2	9-I10-K10-M10-O10	9-I18-K18-M18-O18	9-J1-L1-N1-P1	9-J9-L9-N9-P9	9-J17-L17-N17-P17	9-J2-L2-N2-P2	9-J10-L10-N10-P10	9-J18-L18-N18-P18
	C	9-A3-C3-E3-G3	9-A11-C11-E11-G11	9-A19-C19-E19-G19	9-A4-C4-E4-G4	9-A12-C12-E12-G12	9-A20-C20-E20-G20	9-B3-D3-F3-H3	9-B11-D11-F11-H11	9-B19-D19-F19-H19	9-B4-D4-F4-H4	9-B12-D12-F12-H12	9-B20-D20-F20-H20
	D	9-I3-K3-M3-O3	9-I11-K11-M11-O11	9-I19-K19-M19-O19	9-I4-K4-M4-O4	9-I12-K12-M12-O12	9-I20-K20-M20-O20	9-J3-L3-N3-P3	9-J11-L11-N11-P11	9-J19-L19-N19-P19	9-J4-L4-N4-P4	9-J12-L12-N12-P12	9-J20-L20-N20-P20
	E	9-A5-C5-E5-G5	9-A13-C13-E13-G13	9-A21-C21-E21-G21	9-A6-C6-E6-G6	9-A14-C14-E14-G14	9-A22-C22-E22-G22	9-B5-D5-F5-H5	9-B13-D13-F13-H13	9-B21-D21-F21-H21	9-B6-D6-F6-H6	9-B14-D14-F14-H14	9-B22-D22-F22-H22
	F	9-I5-K5-M5-O5	9-I13-K13-M13-O13	9-I21-K21-M21-O21	9-I6-K6-M6-O6	9-I14-K14-M14-O14	9-I22-K22-M22-O22	9-J5-L5-N5-P5	9-J13-L13-N13-P13	9-J21-L21-N21-P21	9-J6-L6-N6-P6	9-J14-L14-N14-P14	9-J22-L22-N22-P22
	G	9-A7-C7-E7-G7	9-A15-C15-E15-G15	9-A23-C23-E23-G23	9-A8-C8-E8-G8	9-A16-C16-E16-G16	9-A24-C24-E24-G24	9-B7-D7-F7-H7	9-B15-D15-F15-H15	9-B23-D23-F23-H23	9-B8-D8-F8-H8	9-B16-D16-F16-H16	9-B24-D24-F24-H24
	H	9-I7-K7-M7-O7	9-I15-K15-M15-O15	9-I23-K23-M23-O23	9-I8-K8-M8-O8	9-I16-K16-M16-O16	9-I24-K24-M24-O24	9-J7-L7-N7-P7	9-J15-L15-N15-P15	9-J23-L23-N23-P23	9-J8-L8-N8-P8	9-J16-L16-N16-P16	9-J24-L24-N24-P24
P10		1	2	3	4	5	6	7	8	9	10	11	12
	A	10-A1-C1-E1-G1	10-A9-C9-E9-G9	10-A17-C17-E17-G17	10-A2-C2-E2-G2	10-A10-C10-E10-G10	10-A18-C18-E18-G18	10-B1-D1-F1-H1	10-B9-D9-F9-H9	10-B17-D17-F17-H17	10-B2-D2-F2-H2	10-B10-D10-F10-H10	10-B18-D18-F18-H18

	B	10-I1-K1-M1-O1	10-I9-K9-M9-O9	10-I17-K17-M17-O17	10-I2-K2-M2-O2	10-I10-K10-M10-O10	10-I18-K18-M18-O18	10-J1-L1-N1-P1	10-J9-L9-N9-P9	10-J17-L17-N17-P17	10-J2-L2-N2-P2	10-J10-L10-N10-P10	10-J18-L18-N18-P18
	C	10-A3-C3-E3-G3	10-A11-C11-E11-G11	10-A19-C19-E19-G19	10-A4-C4-E4-G4	10-A12-C12-E12-G12	10-A20-C20-E20-G20	10-B3-D3-F3-H3	10-B11-D11-F11-H11	10-B19-D19-F19-H19	10-B4-D4-F4-H4	10-B12-D12-F12-H12	10-B20-D20-F20-H20
	D	10-I3-K3-M3-O3	10-I11-K11-M11-O11	10-I19-K19-M19-O19	10-I4-K4-M4-O4	10-I12-K12-M12-O12	10-I20-K20-M20-O20	10-J3-L3-N3-P3	10-J11-L11-N11-P11	10-J19-L19-N19-P19	10-J4-L4-N4-P4	10-J12-L12-N12-P12	10-J20-L20-N20-P20
	E	10-A5-C5-E5-G5	10-A13-C13-E13-G13	10-A21-C21-E21-G21	10-A6-C6-E6-G6	10-A14-C14-E14-G14	10-A22-C22-E22-G22	10-B5-D5-F5-H5	10-B13-D13-F13-H13	10-B21-D21-F21-H21	10-B6-D6-F6-H6	10-B14-D14-F14-H14	10-B22-D22-F22-H22
	F	10-I5-K5-M5-O5	10-I13-K13-M13-O13	10-I21-K21-M21-O21	10-I6-K6-M6-O6	10-I14-K14-M14-O14	10-I22-K22-M22-O22	10-J5-L5-N5-P5	10-J13-L13-N13-P13	10-J21-L21-N21-P21	10-J6-L6-N6-P6	10-J14-L14-N14-P14	10-J22-L22-N22-P22
	G	10-A7-C7-E7-G7	10-A15-C15-E15-G15	10-A23-C23-E23-G23	10-A8-C8-E8-G8	10-A16-C16-E16-G16	10-A24-C24-E24-G24	10-B7-D7-F7-H7	10-B15-D15-F15-H15	10-B23-D23-F23-H23	10-B8-D8-F8-H8	10-B16-D16-F16-H16	10-B24-D24-F24-H24
	H	10-I7-K7-M7-O7	10-I15-K15-M15-O15	10-I23-K23-M23-O23	10-I8-K8-M8-O8	10-I16-K16-M16-O16	10-I24-K24-M24-O24	10-J7-L7-N7-P7	10-J15-L15-N15-P15	10-J23-L23-N23-P23	10-J8-L8-N8-P8	10-J16-L16-N16-P16	10-J24-L24-N24-P24
P11		1	2	3	4	5	6	7	8	9	10	11	12
	A	11-A1-C1-E1-G1	11-A9-C9-E9-G9	11-A17-C17-E17-G17	11-A2-C2-E2-G2	11-A10-C10-E10-G10	11-A18-C18-E18-G18	11-B1-D1-F1-H1	11-B9-D9-F9-H9	11-B17-D17-F17-H17	11-B2-D2-F2-H2	11-B10-D10-F10-H10	11-B18-D18-F18-H18
	B	11-I1-K1-M1-O1	11-I9-K9-M9-O9	11-I17-K17-M17-O17	11-I2-K2-M2-O2	11-I10-K10-M10-O10	11-I18-K18-M18-O18	11-J1-L1-N1-P1	11-J9-L9-N9-P9	11-J17-L17-N17-P17	11-J2-L2-N2-P2	11-J10-L10-N10-P10	11-J18-L18-N18-P18
	C	11-A3-C3-E3-G3	11-A11-C11-E11-G11	11-A19-C19-E19-G19	11-A4-C4-E4-G4	11-A12-C12-E12-G12	11-A20-C20-E20-G20	11-B3-D3-F3-H3	11-B11-D11-F11-H11	11-B19-D19-F19-H19	11-B4-D4-F4-H4	11-B12-D12-F12-H12	11-B20-D20-F20-H20
	D	11-I3-K3-M3-O3	11-I11-K11-M11-O11	11-I19-K19-M19-O19	11-I4-K4-M4-O4	11-I12-K12-M12-O12	11-I20-K20-M20-O20	11-J3-L3-N3-P3	11-J11-L11-N11-P11	11-J19-L19-N19-P19	11-J4-L4-N4-P4	11-J12-L12-N12-P12	11-J20-L20-N20-P20
	E	11-A5-C5-E5-G5	11-A13-C13-E13-G13	11-A21-C21-E21-G21	11-A6-C6-E6-G6	11-A14-C14-E14-G14	11-A22-C22-E22-G22	11-B5-D5-F5-H5	11-B13-D13-F13-H13	11-B21-D21-F21-H21	11-B6-D6-F6-H6	11-B14-D14-F14-H14	11-B22-D22-F22-H22
	F	11-I5-K5-M5-O5	11-I13-K13-M13-O13	11-I21-K21-M21-O21	11-I6-K6-M6-O6	11-I14-K14-M14-O14	11-I22-K22-M22-O22	11-J5-L5-N5-P5	11-J13-L13-N13-P13	11-J21-L21-N21-P21	11-J6-L6-N6-P6	11-J14-L14-N14-P14	11-J22-L22-N22-P22
	G	11-A7-C7-E7-G7	11-A15-C15-E15-G15	11-A23-C23-E23-G23	11-A8-C8-E8-G8	11-A16-C16-E16-G16	11-A24-C24-E24-G24	11-B7-D7-F7-H7	11-B15-D15-F15-H15	11-B23-D23-F23-H23	11-B8-D8-F8-H8	11-B16-D16-F16-H16	11-B24-D24-F24-H24
	H	11-I7-K7-M7-O7	11-I15-K15-M15-O15	11-I23-K23-M23-O23	11-I8-K8-M8-O8	11-I16-K16-M16-O16	11-I24-K24-M24-O24	11-J7-L7-N7-P7	11-J15-L15-N15-P15	11-J23-L23-N23-P23	11-J8-L8-N8-P8	11-J16-L16-N16-P16	11-J24-L24-N24-P24
P12		1	2	3	4	5	6	7	8	9	10	11	12
	A	12-A1-C1-E1-G1	12-A9-C9-E9-G9	12-A17-C17-E17-G17	12-A2-C2-E2-G2	12-A10-C10-E10-G10	12-A18-C18-E18-G18	12-B1-D1-F1-H1	12-B9-D9-F9-H9	12-B17-D17-F17-H17	12-B2-D2-F2-H2	12-B10-D10-F10-H10	12-B18-D18-F18-H18
	B	12-I1-K1-M1-O1	12-I9-K9-M9-O9	12-I17-K17-M17-O17	12-I2-K2-M2-O2	12-I10-K10-M10-O10	12-I18-K18-M18-O18	12-J1-L1-N1-P1	12-J9-L9-N9-P9	12-J17-L17-N17-P17	12-J2-L2-N2-P2	12-J10-L10-N10-P10	12-J18-L18-N18-P18
	C	12-A3-C3-E3-G3	12-A11-C11-E11-G11	12-A19-C19-E19-G19	12-A4-C4-E4-G4	12-A12-C12-E12-G12	12-A20-C20-E20-G20	12-B3-D3-F3-H3	12-B11-D11-F11-H11	12-B19-D19-F19-H19	12-B4-D4-F4-H4	12-B12-D12-F12-H12	12-B20-D20-F20-H20
	D	12-I3-K3-M3-O3	12-I11-K11-M11-O11	12-I19-K19-M19-O19	12-I4-K4-M4-O4	12-I12-K12-M12-O12	12-I20-K20-M20-O20	12-J3-L3-N3-P3	12-J11-L11-N11-P11	12-J19-L19-N19-P19	12-J4-L4-N4-P4	12-J12-L12-N12-P12	12-J20-L20-N20-P20
	E	12-A5-C5-E5-G5	12-A13-C13-E13-G13	12-A21-C21-E21-G21	12-A6-C6-E6-G6	12-A14-C14-E14-G14	12-A22-C22-E22-G22	12-B5-D5-F5-H5	12-B13-D13-F13-H13	12-B21-D21-F21-H21	12-B6-D6-F6-H6	12-B14-D14-F14-H14	12-B22-D22-F22-H22
	F	12-I5-K5-M5-O5	12-I13-K13-M13-O13	12-I21-K21-M21-O21	12-I6-K6-M6-O6	12-I14-K14-M14-O14	12-I22-K22-M22-O22	12-J5-L5-N5-P5	12-J13-L13-N13-P13	12-J21-L21-N21-P21	12-J6-L6-N6-P6	12-J14-L14-N14-P14	12-J22-L22-N22-P22
	G	12-A7-C7-E7-G7	12-A15-C15-E15-G15	12-A23-C23-E23-G23	12-A8-C8-E8-G8	12-A16-C16-E16-G16	12-A24-C24-E24-G24	12-B7-D7-F7-H7	12-B15-D15-F15-H15	12-B23-D23-F23-H23	12-B8-D8-F8-H8	12-B16-D16-F16-H16	12-B24-D24-F24-H24
	H	12-I7-K7-M7-O7	12-I15-K15-M15-O15	12-I23-K23-M23-O23	12-I8-K8-M8-O8	12-I16-K16-M16-O16	12-I24-K24-M24-O24	12-J7-L7-N7-P7	12-J15-L15-N15-P15	12-J23-L23-N23-P23	12-J8-L8-N8-P8	12-J16-L16-N16-P16	12-J24-L24-N24-P24