New technologies for high throughput genetic analysis of complex genomes

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Laura-Jayne Gardiner

July 2014

UNIVERSITY OF LIVERPOOL

Acknowledgements

I would firstly like to thank my parents and my husband Dave, who have provided me with constant support and encouragement.

I am particularly grateful for the years of advice, guidance and teasing provided by my main supervisor Anthony Hall and my secondary supervisor Neil Hall. The wisdom and ideas from the car sharing collaborative is greatly appreciated. Thank you to the many members of the CGR and the bioinformatics office for their help and for keeping me sane.

Finally thanks to Walt Disney, Topshop and to Jacobs Creek for keeping up my spirits throughout this project.

Table of Contents

Table of Contents ………………………………………………………………………………… **3**

List of Tables ……………………………………………………………………………….. **6**

List of Figures ……………………………………………………………………………….. **7**

List of Abbreviations ……………………………………………………………………….. **9**

Abstract ……………………………………………………………………………………… **10**

Chapter 1. Introduction ……………………………………………………………………….. **12**

1.1 Summary ……………………………………………………………………………….... **12**

1.2 Outline of sequencing technologies ……………………………………………………….. **13**

1.2.1 Roche 454 sequencing ……………………………………………………………….. 16

1.2.2 Applied Biosystem's SOLiD™ sequencing system ……………………………………… 18

1.2.3 Illumina sequencing ………………………………………………………………………. 21

1.2.4 Outlining recent developments in next generation sequencing …………………………… 22

1.3 Bioinformatical tools: Mapping and downstream analyses …………………………….. **24**

1.3.1 Mapping ……………………………………………………………………………….... 25

1.3.1.1 BWA …………………………………………………………………………………… 28

1.3.2 Analyses downstream of mapping ……………………………………………………... 30

1.3.3 Further bioinformatical tools ……………………………………………………………… 34

1.4 Arabidopsis ……………………………………………………………………………… **37**

1.5 Brachypodium ……………………………………………………………………………... **40**

1.6 Wheat ……………………………………………………………………………………… **41**

1.6.1 Evolution of wheat ……………………………………………………………………... 43

1.6.2 Gene enrichment in wheat ……………………………………………………………... 44

1.7 Identifying mutations responsible for traits in complex organisms …………………… **49**

1.7.1 Summary ………………………………………………………………………………….. 49

1.7.2 Methods for identifying traits ……………………………………………………………. 50

1.8 Wheat methylation studies ……………………………………………………………... **55**

1.9 Aims ………………………………………………………………………………………… **58**

Chapter 2. Mutant identification in the model plant *Arabidopsis thaliana* ………………… **60**

2.1 Introduction ……………………………………………………………………………... **60**

2.2 *Ebi-1* SNP identification ……………………………………………………………... **62**

2.3 Mutant identification in the *Arabidopsis* strain Ws-2 …………………………………… **64**

2.4 Further mutation identification in the *Arabidopsis* strain Ws-2 ……………………….. **68**

2.4.1 SHORE mapping pipeline ……………………………………………………………… 68

2.4.2 Development of a novel pipeline for mutant identification with use of BWA and SAMtools ……………………………………………………………………………….. 73

2.5 Mutant identification in a diploid using an artificial dataset …………………………… **77**

2.5.1 Development of an artificial diploid sequence dataset …………………………………… 78

2.5.2 Mutant identification in the artificial diploid dataset ………………………………….. 80

2.6 Mutant identification in a tetraploid using an artificial dataset ……………………….. 82

2.6.1 Development of an artificial tetraploid sequence dataset ………………………………… 82

2.6.2 Mutant identification in the artificial tetraploid dataset ………………………………….. 83

2.7 Mutant identification in a hexaploid using an artificial dataset ……………………….. 86

2.7.1 Development of an artificial hexaploid sequence dataset ………………………………… 86

2.7.2 Mutant identification in the artificial hexaploid dataset ………………………………….. 87

2.8 Conclusions …………………………………………………………………………… 90

Chapter 3. Validation of a wheat gene capture array ……………………………………….. 93

3.1 Introduction ………………………………………………………………………….. 93

3.2 Intron and exon modeling …………………………………………………………… 95

3.3 Ordering array probes ……………………………………………………………….. 96

3.4 Comparative enrichment study (exome array versus gene capture array) ……………. 100

3.4.1 Mapping analysis ……………………………………………………………………… 100

3.4.2 SNP analysis ………………………………………………………………………....... 103

3.5 Comprehensive analysis of the array targets ……………………………………………. 104

3.6 Mapping and SNP identification in four wheat varieties ………………………………. 106

3.7 Conclusions ………………………………………………………………………........... 115

Chapter 4. Mutant identification combined with target enrichment in wheat ……………. 118

4.1 Introduction ………………………………………………………………………........... 118

4.2 Mutant identification in the diploid wheat *T. monococcum* ……………………………. 121

4.2.1 Sample preparation and mapping ……………………………………………………… **121**

4.2.2 Initial homozygote allele frequency determination for mutant identification …………… **125**

4.2.3 Final haplotyping algorithm for mutant identification …………………………………… **127**

4.2.4 Conclusions from mutant identification in *T. monococcum* ……………………………… **130**

4.3 Identification of genes linked to stripe rust resistance in 3 hexaploid wheat mutants ... 135

4.3.1 Mapping and SNP identification pipelines ……………………………………………….. **135**

4.3.2 P2 and P3 datasets; Mutant Identification ………………………………………………… **137**

4.4 Conclusions …………………………………………………………………………… 142

Chapter 5. A comprehensive genome wide analysis of methylation patterns in wheat …… 145

5.1 Introduction ………………………………………………………………………........... 145

5.2 Design of the methylation array …………………………………………………………. 147

5.3 Identification of a reference list of homeologous SNPs …………………………………. 148

5.4 Enrichment performance and validation of the array ……………………………………... 151

5.4.1 Non-Bisulfite treated samples …………………………………………………………….. **151**

5.4.2 Setting thresholds for methylation ………………………………………………………... 155

5.4.3 Bisulfite treated samples ………………………………………………………………….. 155

5.5 Determination of the available dataset for analysis ……………………………………... 158

5.6 Identification of global methylation patterns …………………………………………… 160

5.6.1 Identification of genome specific methylation/non-methylation in wheat ……………….. **161**

5.6.2 Identification of genome independent methylation ……………………………………….. 166

5.7 Transposon and chloroplast methylation state assessment through the analysis of off target sequence . ………………………………………………………………………….. 167

5.8 Investigating temperature dependent differential methylation and gene expression …. 169

5.9 Validation of homeologous SNP calls and methylation status ………………………….. 176

5.10 Conclusions …………………………………………………………………………….. 179

6 Discussion ……………………………………………………………………………… 181

7 References ……………………………………………………………………………… 188

Files on disc

8 Appendix 1 (tables 1-9 in file: Appendix_1.docx) ……………………………………… 64

9 Command outline appendix (sections 1-5 in file: Command_outline_appendix.docx) ... 64

10 Allele-frequency-interval-determination.pl ……………………………………………. 74

11 Appendix 2 (table 1 and figure 1 in file: Appendix_2.docx) …………………………….. 95

12 3rd_base_script.pl ………………………..…………………………..………………... 108

13 ID_varietal_SNPs.pl ………………………..…………………………..………………... 108

14 Four_varieties_SNPs_beta_array_final.txt ……………………….…………………... 109

15 Appendix 3 (tables 1-6 in file: Appendix_3.docx) …………………………………….... 125

16 Homozygote_frequency_plus_plot.pl ………………………..…………………………. 132

17 Haplotyping_hex_wheat_plus_plot.pl ………………………..………………………… 134

18 Homeologous_SNP_list.txt ………………………..…………………………..………… 148

19 Are_SNP_reads_methylated.pl ………………………..……………………………….... 149

20 Assign_read_to_genome.pl ………………………..……………………………………... 149

21 Associate_sequencing_read_with_SNP.pl ………………………..…………………….. 149

22 Calc_%_read_meth2.pl ………………………..……………………………………….... 149

23 Appendix 4 (tables 1-8 and figure 1 in file: Appendix_4.docx) ………………………… 160

List of Tables

Table 1.1 A comparison of next-generation sequencing platforms ………………………….. 16

Table 2.1. Top candidate homozygous SNPs for the phenotype inducing SNP of an *Arabidopsis* Ws-2 mutant ………………………………………………………………….... 68

Table 2.2. Top candidate homozygous SNPs for the phenotype inducing SNP of an *Arabidopsis* Ws-2 mutant ………………………………………………………………….... 72

Table 2.3. Descriptive mapping statistics for an *Arabidopsis* Ws-2 mutant ………………….. 74

Table 2.4. Top candidate homozygous SNPs for the phenotype inducing SNP of an *Arabidopsis* Ws-2 mutant ………………………………………………………………….... 77

Table 3.1. Exome capture array versus gene capture array …………………………………… 102

Table 3.2. Exome capture array targets versus gene capture array targets …………………… 105

Table 3.3. Mapping statistics for four varieties of wheat …………………………………….. 106

Table 4.1 Mapping statistics for the 4 enriched wheat DNA samples in relation to the pseudo-chromosome reference sequence …………………………………………………….. 124

Table 4.2 Detailing the pseudo-chromosome regions that harbour potential deletions ……… 131

Table 4.3 Yr-7 iSelect SNPs mapping to wheat pseudo-chromosome 2 …………………….. 142

Table 5.1. Mapping Statistics for enriched wheat DNA samples (non-bisulfite treated) …….. 152

Table 5.2. Mapping Statistics for six enriched and bisulfite treated wheat DNA samples …… 156

Table 5.3. Overall Mapping Statistics in Bisulfite treated data ……………………………… 157

Table 5.4. Summary of orientation of the methylation sites that were analyzed in sample 12 (12°C) and sample 27 (27°C) ……………………………………………………………… 163

Table 5.5. Repeat composition of the 12°C sample's unmapped sequencing reads …………. 168

Table 5.6. Annotation of differentially methylated sites ……………………………………… 176

Table 5.7. Methylation site validation data …………………………………………………… 177

List of Figures

Figure 1.1 Basic outline of analysis of next-generation sequencing datasets to perform SNP calling ……………………………………………………………………………………….. 12

Figure 1.2. Steps to generate a mate-paired library for Roche 454 sequencing …………….. 18

Figure 1.3. An outline of SOLiD sequencing technology …………………………………... 20

Figure 1.4. An outline of the Illumina sequencing-by-synthesis technology ……………….. 22

Figure 1.5. An outline of assembly and mapping of sequencing reads ……………………... 25

Figure 1.6. Mapping and SNP calling pipeline ……………………………………………… 27

Figure 1.7. Example of a Circos plot: Alignment of wheat 454 reads, SNPs and genetic maps to the *Brachypodium* genome ………………………………………………………… 36

Figure 1.8. Image of the model diploid plant *Arabidopsis thaliana* ………………………… 37

Figure 1.9. Image of the model diploid plant *Brachypodium distachyon* …………………… 40

Figure 1.10. Evolution of hexaploid bread wheat …………………………………………… 44

Figure 1.11. Development of 2 wheat capture arrays in solution …………………………… 47

Figure 2.1. Positions of *ebi-1* SNPs along the *Arabidopsis* chromosomes 1-5 ……………... 64

Figure 2.2 Pipeline for mutant identification using SHOREmapping ……………………… 65

Figure 2.3. SHOREmap denovo output pdf file for an *Arabidopsis* Ws-2 mutant ………… 67

Figure 2.4. SHOREmap denovo output pdf file for an *Arabidopsis* Ws-2 mutant …………. 70

Figure 2.5. Allele frequency analysis of an *Arabidopsis* Ws-2 mutant ……………………... 75

Figure 2.6. Outline of sequence input into SAMtools wgsim ……………………………… 79

Figure 2.7. Allele frequency analysis of a simulated *Arabidopsis* diploid mutant …………. 81

Figure 2.8. Allele frequency analysis of a simulated *Arabidopsis* tetraploid mutant ………. 85

Figure 2.9. Allele frequency analysis of a simulated *Arabidopsis* hexaploid mutant ………. 89

Figure 3.1. Construction of pseudo-chromosomes from the gene capture array design-space contigs using *Brachypodium*-wheat markers ………………………………………………... 97

Figure 3.2. Using barley for the construction of pseudo-chromosomes from the gene capture array design-space contigs ……………………………………………………….. 99

Figure 3.3. Extract from a VarScan SNP call output at a position with more than one alternate allele ………………………………………………………………………….. 107

Figure 3.4. SNPs that have been identified in three varieties of wheat ……………………... 110

Figure 3.5. Circos plot outlining the three wheat varieties Rialto, Truman and Utmost ……. 111

Figure 3.6. Circos plot outlining the three wheat varieties Rialto, Truman and Utmost ……. 114

Figure 4.1 Construction of a RIL line ……………………………………………………. 120

Figure 4.2 Processing 4 sets of enriched sequencing data to identify a mapping interval containing the deletion that is inducing the phenotype of interest…………………………… 123

Figure 4.3 Frequencies of Bulk A and Bulk B homozygotes calculated along each pseudo-chromosome ……………………………………………………………………………… 126

Figure 4.4 Outline of mutant identification algorithm ……………………………………… 128

Figure 4.5 Homozygosity scores calculated for Bulk A and Bulk B datasets ………………. 129

Figure 4.6 Homozygosity frequencies/scores calculated for Bulk A and Bulk B datasets along each pseudo-chromosome ………………………………………………………….. 133

Figure 4.7 Processing 5 sets of enriched sequencing data to identify a mapping interval containing the gene that is inducing the phenotype of interest ……………………………… 136

Figure 4.8 Homozygosity scores calculated for the P2 and P3 bulk segregant datasets along each pseudo-chromosome ………………………………………………………………… 139

Figure 4.9 Homozygosity scores calculated from the P2 and P3 bulk segregant datasets for the iSelect Yr-7 linked SNP positions ……………………………………………………. 141

Figure 5.1. Theory behind association of methylation sites with the 3 wheat genomes …….. 149

Figure 5.2. Pipeline for association of methylation sites with the 3 wheat genomes ……….. 150

Figure 5.3. Average depth of coverage per bait probe plotted for the 12°C sample ………... 154

Figure 5.4. Determination of the subset of data that was available for detailed analysis in the 12°C sample ………………………………………………………………………... 159

Figure 5.5. Categorizing observed methylation averaged across the 12°C and 27°C sample datasets ………………………………………………………………………………. 162

Figure 5.6. Positional information for methylation sites ……………………………………. 165

Figure 5.7. Frequency plot of percentage per genome contribution to overall gene expression …………………………………………………………………………………… 170

Figure 5.8. Frequency plot of percentage per genome contribution to overall gene expression (subset) ……………………………………………………………………….. 172

Figure 5.9. Sanger sequencing output trace ………………………………………………… 178

List of Abbreviations

AFA, Adaptive Focused Acoustics

BAC, Bacterial Artificial Chromosome

BAM, Binary Alignment Map

CGR, Centre for Genomic Research

EMS, Ethyl Methanesulfonate

NIL, Near Isogenic Line

RIL, Recombinant Inbred Line

SAM, Sequence Alignment Map

SNP, Single Nucleotide Polymorphism

GAIIx, Illumina Genome Analyzer IIx

Abstract

High throughput sequencing can generate hundreds of millions of reads in a single day and is revolutionizing modern genetics. This project aimed to utilize next generation genetic approaches to analyze non-model but important agronomical plant species. A key feature of these species is their complexity. Mapping and SNP calling of these sequencing datasets is fundamental to many downstream analyses that have been implemented here including; mutant identification, comparative analyses between related organisms and epigenetic studies.

The first objective in this project involved developing accelerated mutant identification techniques using mapping-by-sequencing analyses that combine whole genome sequencing with genetic mapping. Such methods have largely required a complete reference sequence and are typically implemented on a mapping population with a common mutant phenotype of interest. Here mutant identification was demonstrated on the model diploid plant *Arabidopsis thaliana* as a proof of principle of the methodology. It was also demonstrated on a simulated hexaploid mutant that was developed using the *Arabidopsis* reference genome. In species such as wheat, no finished genome reference sequence is available and, due to its large genome size (17 Gb), re-sequencing at sufficient depth of coverage is not practical. Therefore a genomic target enrichment approach was validated and used here to capture the gene rich regions of hexaploid bread wheat, reducing the sequencing cost while still allowing analysis of the majority of wheat's genic sequence. A pseudo-chromosome based reference sequence was developed from this genic sequence with a long-range order of genes based on synteny of wheat with *Brachypodium distachyon.* Using the capture probe set for target enrichment followed by next generation sequencing; an early flowering locus was mapped in the diploid wheat *Triticum monococcum* and in hexaploid bread wheat *Triticum aestivum,* the stripe rust resistance gene was located. A bespoke pipeline and algorithm was developed for mutant loci identification and the pseudo-chromosome reference was implemented. This novel method will allow widespread application of sliding window mapping-by-sequencing analyses to datasets that are; enriched, lacking a finished reference genome or polyploid.

The second main objective of this project involved a study of methylation patterns in wheat utilizing sodium bisulfite treatment, combined with target enrichment. An enrichment system was specifically designed, developed, validated and implemented here to perform one of the first studies of methylation patterns in hexaploid bread wheat across the 3 genomes that used a genome-wide subset of genes and can thus be used to infer genome-wide methylation patterns and observations. This investigation confirmed that differential methylation exists

between the A, B and D genomes of wheat and that temperature is capable of altering methylation states.

Note

The work that is outlined in section 4.2 has been adapted for publication and is currently under review (reference 1 below). After publication the relevant pipelines that are detailed in section 4.2 will be made publicly available in Iplant. The work that is outlined in chapter 5 has also been adapted for publication and is currently in the final stages of editing prior to journal submission (reference 2 below).

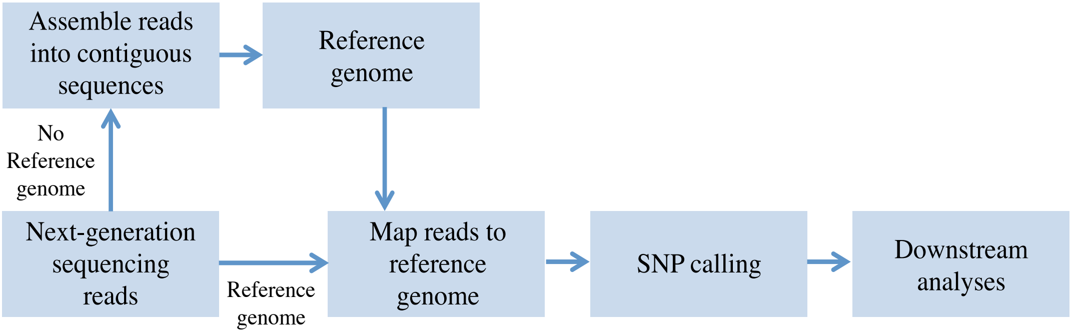
1. Gardiner L, Gawronski P, Olohan L, Schnurbusch T, Hall N, Hall A (2014) Using genic sequence capture in combination with a syntenic pseudo genome to map a deletion mutant in a complex wheat species. Under review-accepted with revision

2. Gardiner L, Olohan L, Price J, Quinton-Tulloch M, Hall N, Hall A (2014) A genome-wide epigenetic study of hexaploid wheat using target enrichment. Paper ready for submission

Chapter 1. Introduction

1.1 Summary

This project is primarily bioinformatically based and focuses on the development of strategies for the analysis of high throughput next-generation DNA re-sequencing datasets in the complex non-model organism wheat. Such analyses are a great challenge due to the large size of the sequencing datasets that are generated. This challenge is intensified when taking into account the polyploid nature of wheat (containing six sets of chromosomes) and wheat's poorly annotated and fragmented genome sequence. Here analyses will be developed to allow the use of these sequencing datasets for mutant identification and comparative analyses between varieties. Epigenetic studies will also be performed i.e. the study of heritable changes in gene activity that are not caused by DNA sequence change e.g. methylation. Such analyses will be outlined in this chapter and typically utilize SNP (single nucleotide polymorphism) calls that are made as a starting point and represent natural DNA sequence variation at a single nucleotide level. The basic profile for sequencing data analysis to determine SNP calls is summarized in figure 1.1, although within this project no sequence assembly has been required due to the availability of reference genome sequences, even if incomplete, for the genomes under analysis. Relevant software that is available to carry out each step shown in figure 1.1 will be outlined in this chapter along with an overview of current sequencing technologies.



Assemble reads into contiguous sequences
Reference genome
No Reference genome
Next-generation sequencing reads
Reference genome
Map reads to reference genome
SNP calling
Downstream analyses

Figure 1.1 Basic outline of analysis of next-generation sequencing datasets to perform SNP calling.

The genomes of interest within this project will also be introduced here; the model diploid plant *Arabidopsis thaliana* that is used to validate analyses prior to adaptation for a complex genome, the model grass *Brachypodium distachyon* that is primarily used to enable comparative analyses, the non-model diploid wheat strain *Triticum monococcum* and hexaploid bread wheat *Triticum aestivum*. Further to this, and to manage the large, complex genome of wheat, the design of two wheat capture array probe sets in solution will be outlined (design by Hall, A). Here these two arrays are validated and implemented in subsequent comparative analyses and mutant identification analyses. An additional in solution capture array, developed within this project, partly using the design and knowledge gained from the first two arrays, will also be outlined. This capture array was implemented within this study to document methylation patterns in hexaploid bread wheat.

This being a primarily bioinformatical project any sequencing that was required is carried out by the CGR at Liverpool University unless otherwise stated.

1.2 Outline of sequencing technologies

Derived in 1977, Sanger sequencing was the original method of choice for rapid determination of DNA sequence for around 30 years, utilizing chain termination. Chain termination involves the synthesis of new DNA strands on a single stranded template and the random incorporation of chain-terminating nucleotides to produce a set of different sized DNA fragments. The last base in a DNA fragment can be identified using a unique label and sizing of these fragments using electrophoresis gains positional information allowing us to read the DNA sequence (Sanger *et al*., 1977). The development of Sanger sequencing as a technique was aided by subsequent laboratory automation and parallelization enabling, currently, the generation of over 2Mb of sequencing data in a single day with read lengths up to ~900bp (Schuster, 2008).

Sanger sequencing was utilized to publish a complete finished genome sequence for the first higher plant, *Arabidopsis,* in 2000 (section 1.4). This analysis involved cloning *Arabidopsis* fragments into a bacterial host using large-insert bacterial artificial chromosome (BAC) libraries. A scaffold of these BAC clones was determined to cover the genome. BACs that formed this scaffold or 'tiling path' were then sheared and sequenced (The Arabidopsis Genome Initiative, 2000). Due to the creation of this tiling path or low-resolution genome map followed by shotgun clone-by-clone sequencing this methodology is slower than if the whole genome was shotgun sequenced directly but relies less heavily on computational genome assembly. To date computing power has steadily become cheaper, and with highly developed software available for sequence analysis, the feasibility for whole genome shotgun sequencing has increased. Sanger sequencing was also utilized to publish the *Brachypodium* genome sequence in 2010, here whole genome shotgun sequence data was generated and assembled then validated using existing genetic maps, physical maps and BAC sequences (The International Brachypodium Initiative, 2010). The *Brachypodium* sequencing initiative is an example of the feasibility of whole genome shotgun sequence for *de novo* genome sequencing. With the advent of next generation sequencing, that significantly streamlines the sequencing process, combined with constantly improving sequencing read lengths, the necessity for a detailed physical or genetic map to complement whole genome shotgun sequencing for *de novo* genome assembly lessens.

High-throughput massively parallel sequencing appeared in 2007 and is referred to as next-generation sequencing. It has the ability to generate hundreds of millions of reads in a single day, revolutionizing modern genetics and allowing a dramatic increase in sequencing output, which is constantly improving, while also lowering sequencing costs. In such a fast moving, competitive industry sequencing technologies are constantly being developed and although technologies tend to be adaptations rather than total overhauls, many of the instruments that were used at the beginning of this project (2010) are now largely redundant and have been replaced by newer models generating longer, higher quality reads and larger volumes of data. Here the current next-generation sequencing technologies will be outlined with the main focus on those that were utilized throughout this project.

Next-generation sequencing typically follows two paths to sequence generation, determined during library preparation of the DNA sample to be run on the desired instrument. The first path, fragment library production, involves simply fragmenting the DNA to pre-defined lengths and sequencing the fragments directly. Sequencing using fragment libraries typically requires less input DNA, it is appropriate for sequence lengths ≤ 300 bp, it has a simpler library construction workflow and it results in higher recovery of unique molecules (Applied Biosystems, 2012). The second method is mate-paired library production. Here the DNA is again fragmented and DNA fragments that are “mates” originate from the 2 ends of one DNA fragment. The distance between “mates” can vary greatly depending on initial read length 500bp-6kb (if the distance is 500bp or less the term paired-end has been employed by certain companies). In addition to the sequence information this method informs of the physical distance between the 2 reads in the genome. For example, if the DNA was fragmented to produce ~500bp fragments then we know that the mate-pairs will map approximately 500bp apart in the genome. This information can help to resolve larger structural rearrangements (insertions, deletions, inversions), as well as helping to assemble across repetitive regions. Sequencing using mate-paired libraries involves more input DNA but tends to result in more even coverage of the genome.

For fragmentation of DNA the Covaris 'Adaptive Focused Acoustics' technology is largely employed. Double stranded DNA is fragmented on exposure to the energy of AFA. AFA can be readily controlled so that the output fragment size after DNA shearing can be precisely selected in the range 100bp-5kb. Covaris technology shows no G/C bias (problematic in enzymatic shearing), has reproducible results, is isothermal (leading to high recovery/fidelity of DNA) and is fast and easy to use (Covaris, The sample prep advantage, 2011).

At the onset of this project in 2010 the most common sequencing technologies included the Roche 454 GS FLX Titanium Series, the Applied Biosystems SOLiD™ 3 Plus Systems and the Illumina Genome Analyzer IIx (GAIIx). A summary of these next-generation sequencing tools and their outputs can be found in table 1.1. The Illumina® (Genome Analyzer^{TM}) and Applied Biosystems^{TM} (SOLiD^{TM}) (Cullum *et al.*, 2010) technologies produce huge amounts of highly accurate (>99%) sequence data (up to ~100 Gb) however, they produce short read sequences that are typically ~35-100bp in length which would make assembly of *de novo* genomes difficult and labour intensive. As such, the sequence output from these technologies is more suitable for whole genome re-sequencing and high-throughput applications e.g. sequencing closely related species, sequencing enriched datasets and transcriptome sequencing. The high coverage that is generated allows confident SNP and mutation detection within populations.

The instruments produced by Roche/454 and Solexa established the next generation sequencing campaign offering millions of reads that are now typically greater than ~500bp in length. Although the sequence data output of these technologies is lower, the increased read length and high accuracy (> 99.997%) allows data to be used primarily for *de novo* sequencing of the genome of interest and medium-throughput applications (Mane *et al.*, 2011).

Sequencing technology	Output (Gb)	Run time (days)	Max read length (bp)	Accuracy	Cost per Mbp ($)
Illumina					
Genome Analyzer IIx	95	14	150	>99%	2
HiSeq 2000	600	11	100	>99%	0.07
HiSeq 2500	1000	6	200	>99%	0.05
HiSeq X ten	1800	3	150	>99%	0.01
MiSeq	15	3	300	>99%	0.10
Applied Biosystems					
SOLiD™ 3 Plus System	60	14	50	>99.9%	-
SOLiD™ 4 Plus System	100	16	50	>99.9%	0.13
SOLiD™ 5500	90	7	75	>99.99%	0.10
SOLiD™ 5500xl	180	7	75	>99.99%	0.10
Life Technologies					
Ion Torrent (PGM)	2	7hr	400	>99%	0.38
Ion Proton (P1)	10	4hr	200	>99%	0.15
Pacific Biosciences					
PacBio RS II	6.4*	3hr	>30000	85%	1
Roche					
Roche 454 GS FLX	0.7	1	1000	>99.997%	10

Table 1.1 A comparison of next-generation sequencing platforms. An outline of the mainstream next-generation sequencing tools; past and present. Figures represent the upper limit/best-case scenario in each category and are taken from the manufacturers specification sheets for each sequencing technology. (*Estimated using the maximum 16 SMRT cells)

1.2.1 Roche 454 sequencing

The Roche 454 GS FLX titanium series implements pyrosequencing and can generate high accuracy reads with lengths up to 1000bp (mode length ~700bp). It is known for its longer read lengths and short run times but higher costs per base pair of sequencing (see table 1.1). 454 pyrosequencing is based on sequencing-by-synthesis; hundreds of thousands of beads, each carrying copies of a unique single-stranded DNA molecule, are sequenced in parallel. They are added to a PicoTiterPlate for sequencing. This plate contains over a million wells; each can hold 1 capture bead. Nucleotide addition to the plate, if the nucleotide is complementary to the template strand, results in the polymerase extending the existing DNA strand by adding nucleotides i.e. controlled non-competitive base extension. This addition results in a light signal that is recorded by the instrument (Rothberg and Leamon, 2008). For this project 454 data is used primarily for *de novo* sequencing of the genome of interest i.e. it can be used to produce a reference sequence that is required for further analyses if this information is not already available e.g. in the case of whole genome shotgun sequencing of the wheat genome (Brenchley *et al.,* 2012).

There is a tendency within 454 sequencing data to see a high error rate in homopolymer regions (regions with three or more identical bases that are consecutive). This is caused as the light signal that is produced during nucleotide addition to the growing DNA strand is proportional to the number of identical bases that are incorporated i.e. the length of the homopolymer. Problems occur when light intensities do not faithfully reflect the homopolymer length or if the exact light intensity is wrongly determined and a subsequent homopolymer length error occurs e.g. AAAA recorded when the correct sequence is AAA. The error rate increases as the homopolymer length increases (Quince *et al.,* 2009).

To implement 454 sequencing initially a DNA library must be prepared from a DNA sample. To produce a fragment library DNA must be sheared to 400-600bp sized fragments. Adaptors can then be attached to both ends of each fragment after repair. Otherwise to produce a mate-paired library an exemplary protocol is as follows (figure 1.2); DNA is fragmented to an average length of 2.5 Kb. Fragmented genomic DNA is end-repaired with unlabeled nucleotides and biotin-labeled circularization adaptors are ligated onto both ends. The fragments are circularized and the resultant circular DNA can be again fragmented creating DNA fragments that have the adaptor DNA in the middle and genomic DNA that was once approximately 2.5 kb apart on each end. These desired fragments are purified from the rest of the genomic DNA using streptavidin beads to capture biotin labels (Roche, 454 Life Sciences, 2006; Berglund *et al.,* 2011). Additional library adaptors can finally be attached to both ends of each fragment.

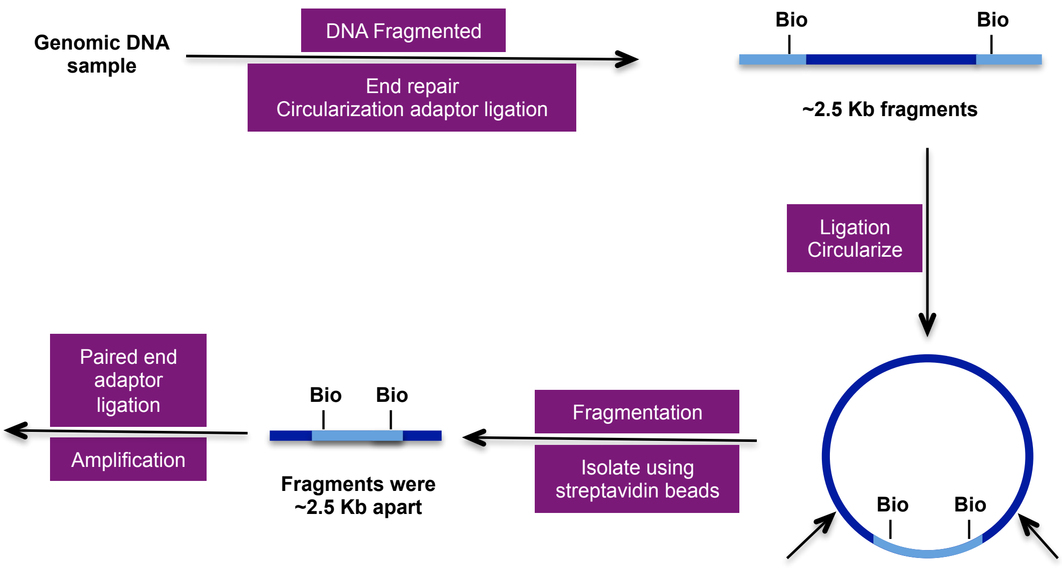


Figure 1.2. Steps to generate a mate-paired library for Roche 454 sequencing (Roche, 454 Life Sciences, 2006; Berglund *et al.*, 2011).

The double stranded DNA fragments with library adaptors at each end are separated into single strands and an emulsion PCR is carried out to amplify them. Emulsion PCR requires the DNA library fragments along with capture beads and PCR reagents in a water mixture, to be mixed with synthetic oil causing ‘micro-reactors’ to form as the water mixture forms droplets around the beads. Each ‘micro-reactor’ will typically contain only one DNA fragment; as such one DNA fragment in each droplet will be amplified in the PCR reaction into millions of copies of DNA that are immobilized on the capture beads. The beads are screened from the oil and cleaned and those containing amplified DNA from one initial DNA fragment only are used for sequencing (Roche, 454 Life Sciences, 2007).

1.2.2 Applied Biosystem’s SOLiD™ sequencing system

The Applied Biosystems SOLiD™ 3 Plus System can generate accurate sequencing reads that are much shorter than those of the 454 sequencer, typically 50-75bp in length, however at a far greater output per run (up to 60Gb in the 3 plus system and upwards of 100Gb in more recent models) and with a lower cost per base pair of sequencing (table 1.1). As a result this sequencer is effectively used for whole genome re-sequencing within this project. SOLiD™ sequencing like 454 sequencing utilizes emulsion PCR to amplify the prepared DNA library (fragment or mate-paired) although it differs from 454 sequencing by implementing sequencing-by-ligation and by de-coding DNA using a unique ‘colour-space’ methodology.

SOLiD sequencing technology (figure 1.3) uses ‘colour-space’ methodology that is based on ligation of dye-labeled oligonucleotides to the DNA, to assay two nucleotides at a time. It allows effective sequencing since most bases are interrogated by two different probes that have ligated independently. The dual interrogation i.e. overlapping dimers, allows for error correcting and effective sequence discrimination. SOLiD™ sequencing therefore implements controlled ligation and competitive oligonucleotide extension and boasts up to 99.99% sequencing accuracy due to its dual interrogation system (Applied Biosystems, 2011).

Prior to the emulsion PCR, DNA libraries must be prepared from samples so that they can be run on the SOLiD sequencer. To produce a fragment library in this case DNA must be sheared to ~100-110bp sized fragments on average. After end-repair of the DNA, adaptors can then be attached to both ends of each fragment. Size selection allows only the correctly sized DNA fragments to be taken forward for analysis. The fragments are then nick translated as there is a 5' phosphate on the end repaired template so the "nick" left over from ligating primers without a 5'-phosphate is translated towards the primer terminus (moves nick along from adaptor-DNA join to allow cutting at specific site including more DNA fragment). Finally the library is PCR amplified (Applied Biosystems, 2012).

To produce a mate-paired library two protocols can be followed. The first to produce typically 2 × 25 bp mate-paired libraries, the protocol is very similar to that in figure 1.1. Adaptors are ligated to sheared DNA and this DNA is circularized with biotinylated internal adaptors. The circularized DNA is cleaved ~25bp away from recognition sites in the adaptor, creating small DNA fragments that have the adaptor DNA in the middle and 25 nucleotides of genomic DNA that were once approximately 600bp-6kb apart on each end (depending on fragment size). P1 and P2 Adaptors are then ligated to the ends of the mate-paired library for subsequent amplification by PCR. The second protocol produces 2 x 50 bp mate-paired libraries. DNA fragments are ligated to LMP CAP Adaptors and then circularized with internal adaptors. The LMP CAP Adaptor is without a 5'-phosphate so nick translation is carried out on a defined length of DNA that is controlled by reaction temperature and time. The DNA is then cut at the position opposite to the nick to release the DNA mate pair. Again P1 and P2 Adaptors are then ligated to the ends of the mate-paired library for subsequent amplification by PCR (Applied Biosystems, 2012). After emulsion PCR amplification of the SOLiD™ sequencing DNA library, beads with a single amplified DNA fragment are separated from undesired beads using bead enrichment with P2-coated Polystrene beads that are used to collect P2-positive beads. These desired beads are then deposited on a slide for sequencing.

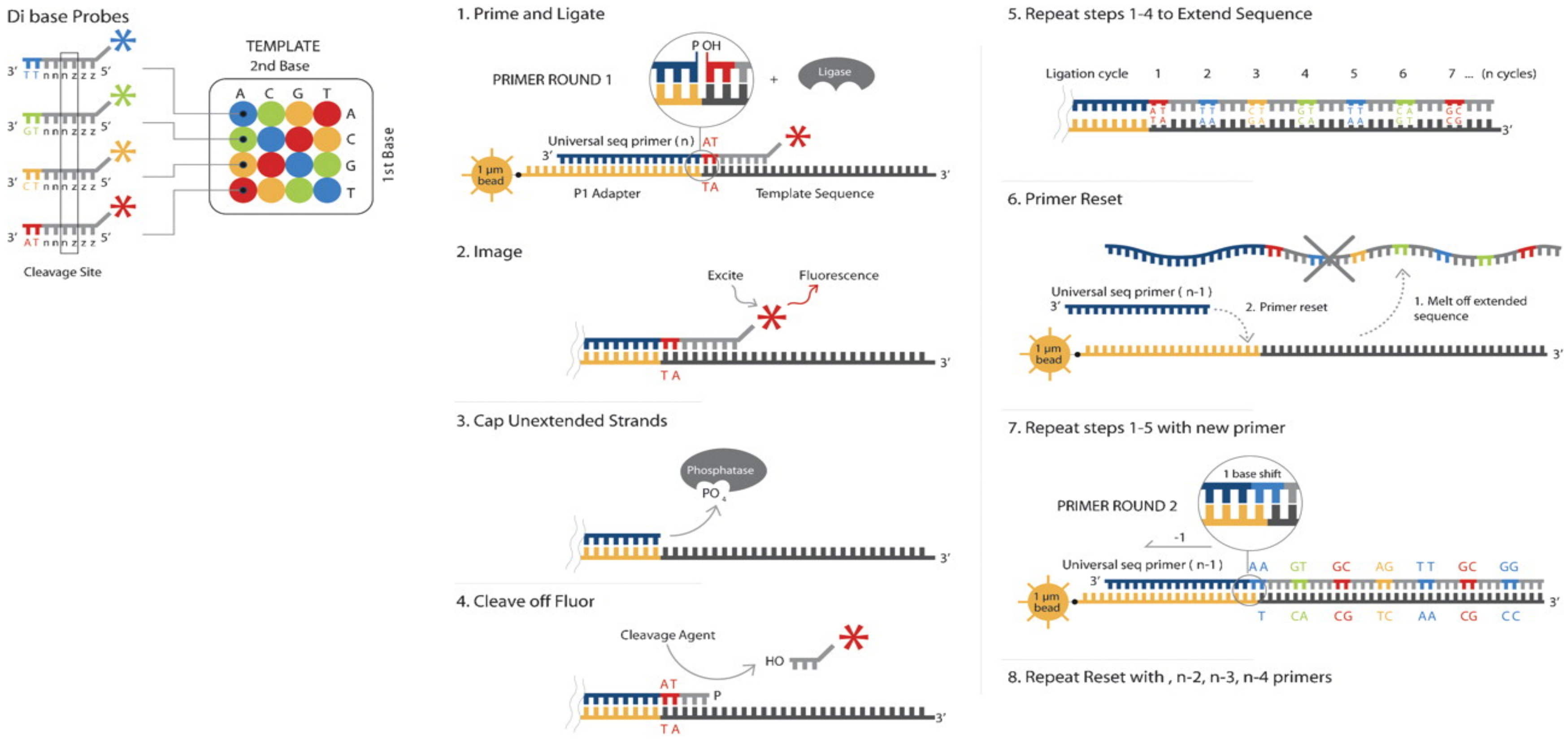


Figure 1.3. An outline of SOLiD sequencing technology. Taken from: Valouev, A. *et al*. (2008). 4 fluorescently labeled di-base probes that encode for 16 possible two-base combinations (4 dimers each) compete for ligation to the sequencing primer **(1)**. The first 2 bases in a ligation reaction are ultimately interrogated. Following a series of ligation cycles (ligation, detection and cleavage) **(2,3,4,5)** the product of extension is removed and a primer complementary to the n-1 position (if the position of the previous primer was n) is used for the next ligation cycle **(6,7)**. This process is repeated 3 more times **(8)**

1.2.3 Illumina sequencing

The Illumina Genome Analyzer IIx is comparable with the Applied Biosystems SOLiD™ System generating accurate sequencing reads that are much shorter than those of the 454 sequencer; typically 100bp in length, however at a far greater output per run (up to 100Gb and almost 20x this in more recent models) and with a lower cost per base pair of sequencing (table 1.1). The Illumina technology also rivals the SOLiD™ System in terms of output and read length. The Illumina Genome Analyzer IIx was effectively used within this project for whole genome re-sequencing. Illumina's technology uses a proprietary reversible terminator-based method for sequencing-by-synthesis and utilizes bridge amplification on a flow cell (figure 1.4) in preference to emulsion PCR.

During Illumina sequencing primers are annealed to DNA strands that are bound to the flow cell and sequencing is enabled through detection of single bases as they are incorporated into growing complementary DNA strands. A fluorescently labeled terminator is imaged as each dNTP is added and then cleaved to allow incorporation of the next base. Since all four reversible terminator-bound dNTPs are present during each sequencing cycle, natural competition minimizes incorporation bias resulting in base-by-base sequencing (Illumina, 2011). This method of sequencing therefore utilizes controlled replication and competitive base extension.

To produce a fragment library for the Illumina Genome Analyzer, DNA is sheared to less than ~800bp. Size selection allows only the correctly sized DNA fragments to be taken forward for analysis. Fragment ends are polished and A-tailed to allow adaptor ligation and finally the library can be PCR amplified. During this PCR further adaptor sequence is added to the fragmented DNA to enable hybridization to oligonucleotides on a flowcell surface. This allows enrichment of fully ligated templates (Kozarewa *et al.,* 2009). To produce up to 2 x 100 bp mate-paired libraries DNA is fragmented to 2-5Kb in size. Fragments are end-repaired with biotin labeled nucleotides and then circularized. The circularized DNA is fragmented and biotinylated fragments can be affinity purified (ends of original DNA fragment ligated together). These biotinylated fragments are end-repaired and adaptors are ligated to each end before PCR amplification where further adaptor sequence is added to enable hybridization to oligonucleotides on the flowcell surface (Berglund *et al.,* 2011).

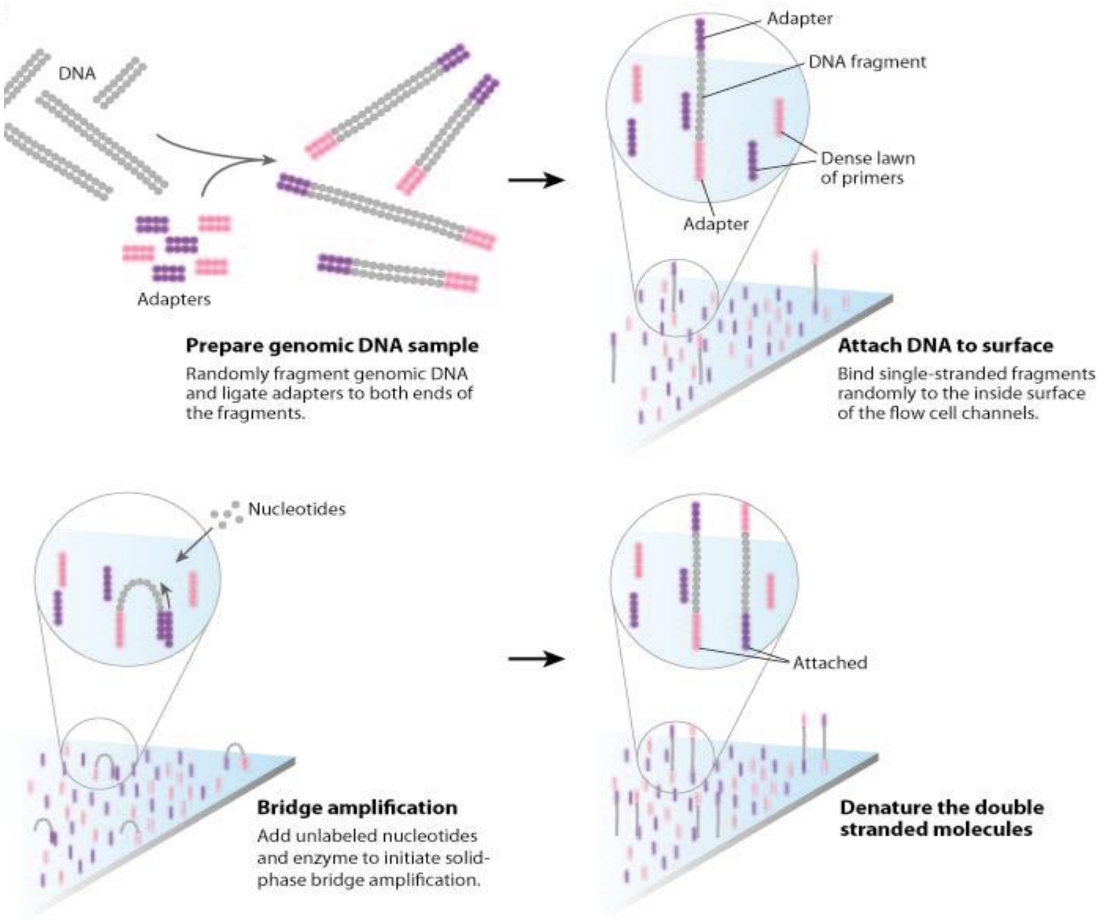

DNA
Adapters
Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
Adapter
DNA fragment
Dense lawn of primers
Adapter
Attach DNA to surface
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
Nucleotides
Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
Attached
Denature the double stranded molecules

Figure 1.4. An outline of the Illumina sequencing-by-synthesis technology. Taken from: Mardis, 2008. Single stranded DNA fragments from mate-paired or fragment libraries are hybridized to the 'lawn of primers' on the flow cell (via complementary oligonucleotide sequences). Flow cell Bridge amplification occurs when the DNA flips over to hybridize adjacent primers forming a bridge. This hybridized primer is extended by polymerases until the bridge is double stranded. It is subsequently denatured leaving 2 single stranded fragments, each covalently bound to a flow cell primer. This cycle is repeated as necessary to fill the flow cell with library fragments and later reverse strands are cleaved and washed away. This generates a cluster on the flow cell with forward strands only that is ready for sequencing (Illumina, 2011).

1.2.4 Outlining recent developments in next generation sequencing

The Illumina® (Genome Analyzer^{TM}) and Applied Biosystems^{TM} (SOLiD^{TM}) (Cullum *et al.*, 2010) technologies that were heavily used at the onset of this project quickly became redundant and in 2014 we see a movement away from SOLiD sequencing altogether with the replacement of the Illumina Genome Analyzer with the current firm favorites for whole genome re-sequencing; the HiSeq 2000 and 2500. With essentially the same sequencing chemistry as its predecessor, but with improved optics and improved sequencing-by-synthesis chemistry, the HiSeq can produce 150bp sequencing reads at a maximum capacity of 1000Gb of data in a single run (table 1.1) (Illumina, 2014). While runs can be lengthy (over 6 days), the HiSeq is the main workhorse of sequencers with its unrivalled high sequencing output and consistently decreasing costs, and many of the sequencing datasets that have been generated more recently for this study have employed its use. The release of the most recent HiSeq X ten in early 2014 promises still higher outputs per run (1800Gb) and decreases the cost of sequencing again. It reportedly, at maximum running capacity, is able to achieve the prized $1000 human genome sequence at 30x (Illumina, 2014(b)).

The 454 sequencer has been utilized for *de novo* sequence assembly requiring long sequencing reads. Throughout 2013 there has been a looming redundancy for the Roche 454 sequencer with greater accessibility of the new PacBio RS technology. The PacBio RS II was released in April offering higher throughput (~400Mb per SMRT cell and up to 16 SMRT cells per run) and longer sequencing lengths (up to 30Kbp and ~8.5Kbp on average) than its predecessor. The PacBio RS II is a Single Molecule, Real-Time (SMRT®) DNA Sequencing System that uses SMRT cells. Each SMRT cell contains thousands of zero-mode waveguides (ZMWs) that act as light microscopes each containing an immobilized DNA template and polymerase complex at the bottom. Fluorescent dye-labeled nucleotides diffuse into the ZMW chamber and nucleotides held by the polymerase prior to incorporation into a complementary strand emit a signal that identifies the base being incorporated. These ZMWs provide a window to watch the DNA polymerases perform sequencing by synthesis and to record light pulses that are emitted by nucleotide incorporation to read the DNA sequence directly (Pacific Biosciences, 2014). Although the 454 sequencer was used within this study it is predicted that future *de novo* sequencing efforts are more likely to employ use of the more cost effective PacBio RS II. Although its first pass error rate is higher, these errors are distributed randomly. As such, its generation of a consensus sequence increases accuracy to ~99.999% and its extremely long read lengths are an invaluable advance to *de novo* sequencing efforts (Pacific Biosciences, 2014 (b)).

When considering lower throughput sequencers, Illumina’s MiSeq (based on the same sequencing chemistry as the HiSeq) is generally favored over the offering under the umbrella of Life Technologies (responsible for SOLiD sequencing), the Ion Torrent PGM. The Ion Torrent performs DNA sequencing based on flooding a microwell containing the DNA strand to be sequenced with a single dNTP; if complementary to the template DNA nucleotide the dNTP will be incorporated into the growing complement strand releasing

Hydrogen ions that can be detected by a sensor. Although the MiSeq has lengthier runs (minimum of ~24 hours compared to ~7 hours) it currently has a higher data output (up to 15Gb compared to 2Gb) and average read length (~250bp compared to ~200bp) and it is generally considered to be the simpler of the 2 platforms to use and for sample preparation (Illumina, 2014). The release of the Life technologies Ion Proton may rival the MiSeq in the future with increasing data output of 10Gb and even up to 60Gb with continuing development (Life technologies, 2014). These technologies have not been utilized within this study due to the large complex genomes that have mainly been under analysis and the high depth of coverage required for analyses.

In 2012 Oxford Nanopore technologies promised a sequencing device, the GridION that is based on nanopore technology. This has yet to come to market although a miniature prototype the MinION has been released to a select few scientists for trial under embargo. The technology involves threading a single stranded DNA molecule through a protein pore while conductivity is measured allowing sequence discrimination. The fast natural rate of DNA passage through the pore has been an issue, as sequence could not be read quickly enough. Various solutions involving molecular motors or processive enzymes have been investigated to slow DNA passage (Clarke *et al.,* 2009). Ultimately it is claimed that the nanopore sequencing technology will be easy to parallelize, inexpensive, able to create very long reads and offers a label-free approach without the need for sample amplification. This would be likely to make Oxford Nanopore technologies key to the future of sequencing although this has yet to be confirmed with widespread product release or published results.

1.3 Bioinformatical tools: Mapping and downstream analyses

Following on from sequencing a genome of interest the sequencing reads can commonly be mapped to a reference genome or assembled into a contiguous sequence (contig) (see figure 1.5). If there was no previous genomic information for the genome we would have to assemble the sequencing reads into contiguous sequences i.e. a reference genome or transcriptome and as such longer sequencing reads are desirable. Software such as Newbler (developed for Roche 454 or Sanger sequencing reads) or Geneious (used for Sanger, 454, Ion Torrent, PacBio, or Illumina sequencing reads), to name only a few, can be used for this assembly. Sequencing data can then be mapped to this assembled ‘reference sequence’ to determine the natural variation within a sample for downstream analyses. More recently, tools such as Cortex can function, typically in the absence of a reference genome, to assemble even short sequencing reads for one or more species and then go on to detect natural variation within the species directly. This streamlining of sequence assembly and natural variation identification paves the way for accessible analysis of species that do not have a reference sequence (Iqbal *et al.,* 2012).

Throughout this project the two main genomes of interest *Arabidopsis* and bread wheat, have assembled genome sequences available. In the case of bread wheat this reference sequence is incomplete and has not yet been organized into structured chromosomes but is still adequate for mapping analyses. If the genome's sequence is already available then it is commonplace to map our sequencing reads against that 'reference' genome using a mapping software tool and as such short sequencing reads are adequate. This is desirable as then natural variation can be called from the mapping output directly in relation to the reference genome where there are differences between the two.

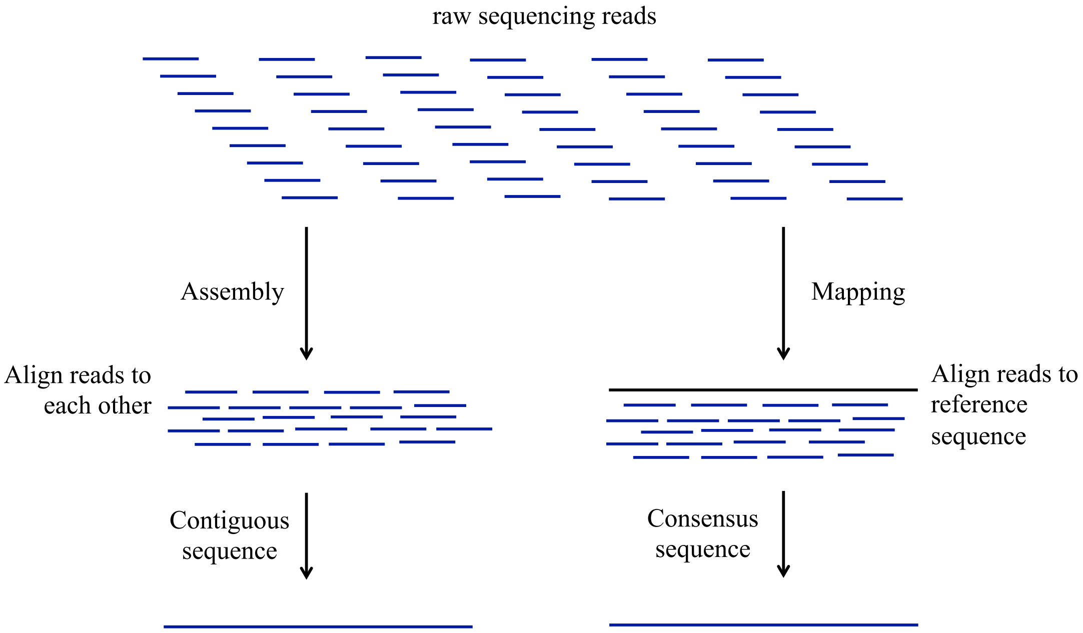


Figure 1.5. An outline of assembly and mapping of sequencing reads. (left) Align sequencing reads to each other in an assembly to generate a contiguous sequence. **(right)** Mapping of sequencing reads to a reference genome to generate a consensus sequence.

1.3.1 Mapping

A typical mapping software has to align relatively short sequencing reads to a reference sequence. It must take into account if the sequencing data is a fragment library or a paired-end library, if the sequence data is in colour-space, the read length and the read quality. Many reads will not align perfectly to the reference sequence due to errors in the sequencing reads and also the potential for natural variation between the reference and the sequence data. This can introduce insertions/deletions and/or SNPs between the reference and the sequencing data and a mapper must typically accept a defined number of mismatches between a sequencing read and the reference before it will define a mapping as unsuccessful.

The two most commonly used mapping programs are detailed as follows and are ideal for the short nucleotide sequences generated by next generation sequencing;
Burrows-Wheeler Aligner (BWA) aligns relatively short nucleotide sequences against a long reference sequence. All of its algorithms perform gapped alignment allowing insertion/deletion detection (Li and Durbin, 2009).
Bowtie aligns short DNA sequences to large genomes. It is fast and relatively memory-efficient. Bowtie extends previous Burrows-Wheeler techniques with a novel quality-aware backtracking algorithm permitting mismatches (Langmead *et al.,* 2009).

The two mapping softwares are highly comparable in terms of memory usage, numbers of reads mapped and number of false positive SNPs called from mappings. However, BWA allows easy and accurate identification of desired uniquely mapping reads and assigns lower quality values to the undesirable non-uniquely mapping reads. BWA also allows for insertions and deletions whereas, at the onset of this project, Bowtie may not have mapped a read if there was an insertion or a deletion within it (Heng, Li, 2010). With the recent introduction of Bowtie2, that now allows gapped alignment, the two tools have become even more comparable. Bowtie2 is thought to run more quickly at the risk of failing to align some reads with a valid match, while BWA runs more slowly and is specifically designed not to miss any potential mappings. There are strengths and weaknesses to both tools and as such there is a degree of consideration of the desired mapping outcome and personal preference involved in choosing between them (Ruffalo *et al.,* 2011).

This reported high degree of comparability between BWA and Bowtie2 was tested here in a direct comparison of the two tools to map an Illumina *Arabidopsis* sequencing dataset. The dataset was run through both tools specifying the paired-end nature of the sequencing reads and otherwise using default parameters. It was noted that there was less than 1% difference in the percentage of the reference sequence that was mapped to (98% by BWA and 98.8% by Bowtie2) and that in terms of the average depth of coverage BWA generated ~35x which was marginally higher than the ~31x that was generated by Bowtie2. As predicted the differences in mapping ability between the two softwares on the surface appear to be relatively small. BWA was selected for mapping analyses during this project, rather than the main alternative Bowtie and its mapping pipeline is outlined in blue in figure 1.6.

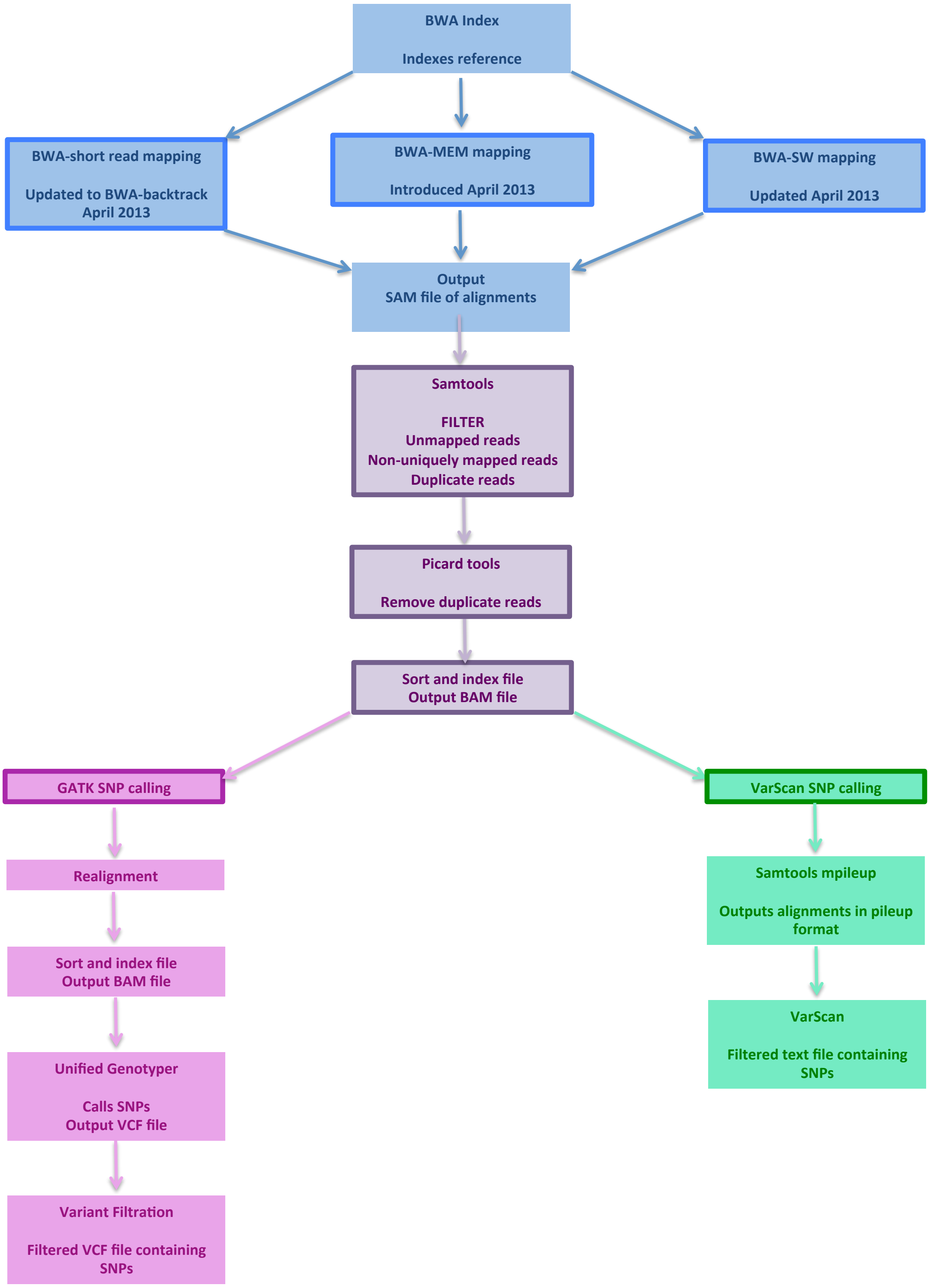


Figure 1.6. Mapping and SNP calling pipeline. (blue) The various pipelines for mapping of sequencing reads using BWA **(purple)** downstream processing of mapping output using SAMtools and picard tools **(pink)** Utilization of GATK to identify and filter SNPs from the mapping output **(green)** Utilization of VarScan to identify and filter SNPs from the mapping output

1.3.1.1 BWA

At the onset of this project BWA initially implemented two algorithms (BWA-short and BWA-SW). The former was intended for query sequences shorter than 200bp and the latter was used for longer sequences (up to 100kbp) (Li and Durbin, 2009). In software versions that were released after Easter 2013, which were developed to cope with the general increases that were being seen in sequencing read length, three algorithms are now used within BWA; BWA-backtrack, BWA-MEM and BWA-SW (Li and Durbin, 2010). BWA-backtrack has replaced BWA-short and although many parameters and commands are conserved between the two it is now advised for use on sequencing reads that are below 100bp in length. BWA-MEM is a newly developed algorithm and, as of 2014, is recommended for mapping all high quality queries over 70bp. BWA-SW, like BWA-MEM, tolerates longer sequences (up to 1Mbp) and was originally the mapper of choice for longer sequencing reads. With the addition of BWA-MEM, BWA-SW is usually recommended if these longer sequences have frequent alignment gaps as it may have better sensitivity. With the bulk of the mapping in this project being performed before Easter 2013 on sequencing reads of 100bp or less mapping analyses have been largely performed using BWA-short.

BWA works by first building an index of the reference sequence. For the read to be aligned to the reference BWA uses a seed from the sequencing read e.g. in BWA-short typically 32bp and with this seed it performs fast look-ups of sections of the reference in the index to identify the reference sequence it can most closely match the seed to. It, by default allows a maximum of 2 differences between the seed and the reference. Once it has this initial seed match in the reference it tries to extend the match along the rest of the sequencing read using its maximum number of differences threshold to ensure that only accurate matches are reported (by default the maximum number of differences between the entire sequencing read and the reference is 4%).

For each successful read alignment to the reference sequence BWA calculates a mapping quality score, this is increased the more closely a read matches the reference sequence and also if the read only maps to one location, as it ideally should. Quality scores can be used to distinguish reads mapping to only one location (uniquely mapped reads) and if mapped to multiple locations can be used to distinguish the most likely correct mapping position for a read.

BWA can, like most other mapping tools, map both mate-paired and fragment sequencing datasets. There is little difference between the 2 except that for the mate-paired dataset, after alignment to the reference, additional checks are performed to identify mate-pairs and their mapping orientation in relation to one another. It estimates the mate-pair insert size on average from all read pairs with both pairs mapped to high quality (20). This insert size can then be used to estimate if any individual mate-pairs that have been aligned to the reference are the correct distance apart and are as such likely to be mapped correctly. Reads pairs are marked accordingly so that they can be discriminated from one another; if they are both mapped and in the correct orientation with the correct insert size separating them, if they are both mapped in the incorrect orientation and finally if only one of a pair is mapped. Quality scores are adjusted according to how well the read maps in addition to information regarding mapping of its mate-pair. It is typical to eliminate mappings from downstream analyses where both mate-paired reads are mapped to different chromosomes whereas mate-pairs mapped with an incorrect orientation on the same chromosome e.g. a larger/smaller insert size could be indicative of insertions/deletions (Li and Durbin, 2009).

Mapping analyses using mate-paired reads are of great value. Although they require a more complex sequencing preparation and higher DNA input, when the resultant sequencing reads are mapped to a well-defined reference sequence they can help to resolve larger structural rearrangements (insertions, deletions, inversions), while providing fairly even coverage of the genome. This does come at a price of loss of depth of coverage partly due to the removal of mate-pairs mapping in the incorrect orientation (likely due to an incorrect mapping of one of the pair) or only one of a pair mapping. Such a loss of coverage can be minor in a well-defined genome and if structural rearrangements are not of primary concern mate pairs where only one of a pair map can be included in downstream analyses to boost coverage. It is possible to map mate-paired data as fragment data and this method is favored if the reference genome is in shorter assembled contigs rather than long chromosomes i.e. more fragmented. This is the case of the wheat reference sequence that is implemented within this project. With a set of shorter reference contigs we are highly unlikely to be able to resolve larger structural rearrangements anyway and many of the mates in mate-pairs are likely to correctly map to different reference sequences and be flagged as having an incorrect orientation. Therefore the benefit of mate-paired mapping is diminished and the quicker fragment mapping generating generally higher coverage is a sensible option. Throughout this project mate-paired reads are still generated for wheat regardless of the un-suitability for the current mapping reference. This is sensible since the ability to map mate-paired reads as fragment reads does not hinder this current investigation but data can also be re-mapped as mate-paired and structural re-arrangements investigated at a later stage when full chromosome sequences are released.

1.3.2 Analyses downstream of mapping

The standard output from mapping software is a SAM file (Sequence Alignment Map); this is a generic format for storing large nucleotide sequence alignments i.e. a text file containing sequence alignment data. It will typically display the following columns per line in a tab-delimited format:

1. Sequencing read ID
2. Flag. A bitwise number to indicate the sequencing read mapping status i.e. unmapped, mapped, mapping orientation, secondary alignment etc.
3. Reference sequence name that sequencing read has mapped to
4. Leftmost mapping position of sequencing read
5. Mapping quality
6. Cigar string. Details the alignment match of the read to the reference including any insertions, deletions, skipped regions and mismatched regions
7. Reference sequence name that mate-pair of the sequencing read has mapped to
8. Leftmost mapping position of mate-pair of the sequencing read
9. Sequencing read length
10. Mapped sequencing read
11. Quality string that is associated with the sequencing read

The most commonly used downstream processing tools use the SAM file output from mapping analyses and are detailed as follows and in figure 1.6 in purple:

SAMtools

SAMtools is a set of programs that manipulate alignments outputted from mapping analyses. Most programs within SAMtools take input in the BAM format (the binary version of a SAM file) and as such it can import from and export to the SAM format. It can carry out sorting, merging and indexing and allows retrieval of reads in specific regions or passing specific filters swiftly. SAMtools mpileup can be used to generate raw SNP and indel calls. It can also generate an easy to read pileup file detailing, for every mapped position, all of the alleles mapping to it and their qualities. This allows easy calculations of percentage of reads with a specific allele (Li *et al.*, 2009). Tools are also available that filter SNP/indel calls directly from a pileup file e.g. VarScan.

Picard tools

Picard tools are an additional set of tools from the makers of SAMtools that were also developed for the manipulation of SAM and BAM files. Tools that are included cover a multitude of functions but the main tool that is utilized within this project is known as MarkDuplicates. This tool allows the identification of and removal of duplicate sequencing reads. Such reads are a by-product of PCR amplification of a DNA sample before sequencing and inclusion of these reads during SNP calling can lead to incorrect representation of allele frequencies (Sourceforge.net, 2009).

Downstream processing tools such as SAMtools are used for processing of mapping outputs before SNP calling can be carried out, regardless of the SNP caller that is used. There are a wide range of SNP callers available and the choice of which to use depends heavily on the dataset under analysis i.e. bulk segregant, single sample, diploid, polyploid etc. and the type of SNP calling required; a greedy approach including as many SNPs as possible that is likely to include false positive results or a more stringent approach that will exclude many of the false positives but potentially also some real SNPs. Two well known SNP callers are outlined here that are complimentary with regard to the datasets that they are recommended for and the stringencies of their SNP calling algorithms. Both of these SNP callers are used within this project:

GATK

The GATK (Genome Analysis Tool Kit) (McKenna *et al.,* 2010) is a structured software library and a suite of tools for working with re-sequencing projects. These tools typically take a sorted, indexed BAM file (SAMtools compatible) as input and include; depth of coverage analyzers, a SNP/indel caller, a SNP filterer and a quality score re-calibrator plus local realigner to improve mapping analyses (DePristo *et al*., 2011). The standard GATK pipeline for SNP calling within this project is outlined in pink in figure 1.6. The GATK is well known for its stringent SNP calls, strict file formatting and command requirements and it has been extensively tested in diploid datasets. At the onset of this project it was designed for use in single sample diploid organisms as such it has been used within this study solely for this purpose when stringent calls are required. By 2013 GATK had the capability to handle polyploid or pooled datasets i.e. report multiple alternate alleles, although it employs the same stringent model as for diploid organisms for determination of allele counts. The output of this 'ploidy' pipeline can be difficult to manage and interpret.

VarScan

The VarScan SNP caller will take a SAMtools mpileup file as input and outputs SNP calls. The standard VarScan pipeline for SNP calling within this project is outlined in green in figure 1.6. With more simple statistics it proves to be more effective in extreme read depth, pooled samples, and contaminated or impure samples and can identify low frequency alternate alleles (Koboldt *et al.,* 2009; Koboldt *et al.,* 2012). Low frequency allele identification is useful within wheat as we would typically expect to see a homeologous homozygous SNP in one of its 3 genomes ~1/3 of the sequencing reads and a heterozygote in ~1/6 of the sequencing reads. VarScan can also output multiple alternate alleles in one position that has allowed its use on polyploid and pooled datasets within this project. Although parameters can be employed to increase stringency, without use of such parameters a less stringent SNP call can be performed using this tool, and it has been employed within this project for this purpose where necessary.

Several pipelines have been introduced that allow simultaneous mapping, SNP calling and mutant identification or mapping-by-sequencing analyses. These are all intended for use in diploid organisms and include, SHORE/SHOREmap, MAQGene, MutMap, NGM-Next-generation EMS mutation mapping and CloudMap (Schneeberger *et al.,* 2009; Doitsidou *et al.,* 2010; Abe *et al.,* 2012; Austin *et al.,* 2011; Minevich *et al.,* 2012). SHOREmap was one of the earliest examples of such a pipeline and is thought to be the most widely used. As such it will be implemented within this study where appropriate and has formed the basis for many derivative mapping-by-sequencing methodologies including MutMap and MAQGene. SHOREmap uses a combination of sequencing a phenotyped F2 mapping population, a technique known as bulk segregation analysis (Michelmore *et al*., 1991), and whole genome re-sequencing to generate approximate mapping intervals that should contain the phenotype inducing mutation.

SHORE links to programs such as BWA and Bowtie to carry out a mapping analysis. It then uses its own SNP calling program ‘SHORE consensus’ to define SNPs. This program sequentially scans an alignment to analyze base calls, base qualities, coverage, repetitiveness and alignment quality. This information is then used to predict differences between the mapped data and the reference sequence. The follow on tool from SHORE, SHOREmap uses its 'denovo' tool to define SNP markers as positions with support for two alleles i.e. potential heterozygotes and to analyze marker scarcity per user defined window (defined as the position-wise average distance to the closest identified marker divided by local marker density). A peak will be identified in a region of low marker density i.e. high homozygosity as SHOREmap assumes that the homozygous character of the region harbouring the causal mutation reduces the density of markers. SHOREmap’s 'annotate' tool is then used on the 'denovo' output to rank base pair substitutions according to their distance to the allele distribution peak. The causal mutation will typically be ranked as the top candidate. This program has allowed successful identification of phenotype inducing mutations in the diploid organism *Arabidopsis* and although ideally requiring a full contiguous chromosomal reference sequence and being based on diploid organisms, its methodologies, in theory, could potentially be applied to polyploid plants (Schneeberger *et al.,* 2009).

More recently, at the end of this project, SHOREmap was developed to be more comparable to CloudMap and will process re-sequencing data for not only a bulk-segregant F2 population but also for the parental lines that were used to produce the population. SHOREmap's software 'SHORE outcross' can take an outcrossed sample as input and can use the support of mutant parental homozygous SNP calls to distinguish high quality markers that have been identified in the F2 bulk segregant population. In the case of a backcrossed sample 'SHORE backcross' can eliminate SNPs that are shared between the F2 and the wild type parental line, reducing the number of 'markers' that are used in the analysis, focusing on only those that are likely to be relevant e.g. a mutant specific SNP markers list that is likely to contain the mutant phenotype inducing SNP. The homozygous character of the region harbouring the causal mutation will still allow 'a rough' visualization of the region. However, the reduced number of markers created the need for an alternative way of plotting the data (rather than distance between heterozygote markers) and the tendency now is to plot SNP markers individually, according to their allele frequency, to locate a region with increased homozygosity. Window averaged allele frequencies can also be calculated if desired. Such an analysis was demonstrated in *Arabidopsis* with additional selection for EMS-induced variation to define a region of interest as the lower arm of chromosome 3. Since this region only contained three homozygous EMS mutations limited further work was required to identify the causal mutation (Hartwig *et al.,* 2012).

At the onset of this project SNP callers in general were mainly intended for use on a dataset that is derived from a diploid organism. As such, only one alternate base would be identified per SNP position i.e. reference and alternate base. The complex nature of wheat means that it can harbour two or even three alternate bases at low frequencies, which poses a problem for SNP calling and it is only now, at the end of my thesis, that such polyploid application of SNP callers has been fully developed (SHORE/SHOREmap and all of the other detailed mapping-by-sequencing pipelines detailed above are still intended for use on a diploid only). For most analyses within this project it was noted that the typically rare occurrence (< 5%) of a second or third alternate allele in the sequence data, at sufficient depth of coverage to call a SNP allele, has made elimination of such SNP calls from the dataset altogether the safer and less time consuming option, having little or no impact on the analysis. For positions with only 1 alternate allele a diploid SNP caller is equally effective and will output a SNP call identically to how a polyploid SNP caller would. In samples requiring efficient identification of multiple alternate alleles with the use of SAMtools mpileup/VarScan and a bespoke parsing pipeline this has also been made possible even prior to specific polyploid SNP caller development.

1.3.3 Further bioinformatical tools

Evolver

It is known that over time a given species will accumulate mutations within its DNA sequence at a calculable rate. If this rate is reliable it can be used as a 'molecular clock' to estimate the divergence between sequences (Ho, 2008). We can use information on the expected amount of sequence change per site from an ancestor to a current species to create an artificially diverged current species sequence from an ancestral sequence. Evolver is a whole genome nucleotide sequence evolution simulator that was implemented within this study for this purpose. Evolver simulates the long-term average effects of mutation/selection in a species giving the user two options; inter-chromosomal changes (e.g. chromosome fission and fusion) and intra-chromosomal changes (e.g. substitutions and insertions/deletions) (Edgar *et al.,* 2012). Within this project only intra chromosomal changes were necessary to give an approximate outline of how a genome would neutrally evolve over time with regard to SNP accumulation at random. Evolver can take genome annotations such as gene locations, non-gene conserved elements, tandem arrays/microsatellites and CpG islands as input to allow informed evolution of the input genome. Although for the neutral intra chromosomal evolution that is utilized within this project none of the above are required.

Evolver takes an ancestral genome as input along with an indication of the length of time over which evolution is required and it outputs an evolved genome along with statistics of the evolutionary events that it has undergone i.e. number of substitutions made. The length of time over which to evolve the input sequence is specified as the 'branch length' i.e. the length of the branch separating the ancestor and evolved species in its evolutionary phylogenetic tree or the expected amount of sequence change per site. For example in wheat this average number of substitutions per site is hypothesized by Gu *et al.* to be 0.0093 for genome A and 0.0056 for genome B when compared to their donor *T. turgidum* and 0.0137 for genome D when compared to its ancestor Ae. *Tauschii* (Gu *et al.,* 2006). *T. turgidum* (AABB) is the tetraploid wheat that hybridized with the diploid wild grass *Ae. tauschii* (DD) ~8000 years ago to form hexaploid bread wheat and as such, these two species are the genome donors for hexaploid wheat.

Within this project the program evolver was used to create an evolved *Arabidopsis* genome (called *Arabidopsis* A genome within this study) that was as different from the *Arabidopsis* reference strain Columbia (Col-0) than the wheat A genome is from its donor T. *turgidum*. To output an evolved sequence evolver took the *Arabidopsis* Columbia reference sequence as input along with a ‘branch length’ of 0.0093. This branch length or expected amount of sequence change per site was the value for genome A in wheat (Gu *et al.,* 2006) and creates a random substitution approximately every 108 bases (Gu *et al.,* 2004). The two sequences; the *Arabidopsis* reference strain Col-0 and the evolved *Arabidopsis* A genome could then be used as the basis for a sequencing dataset for a simulation mutant identification analysis (see section 2.5).

Circos

The tool Circos was also implemented in this project; it is primarily a visualization tool to allow easy analysis and comparison between genomes and also within a genome. This is mainly due to the fact that it displays data in an attractive, publication quality, circular layout. Brenchley *et al.* utilized a Circos plot, shown in figure 1.7, to effectively demonstrate SNP density in the wheat reference strain Chinese Spring (tracks 2-4) and gene conservation between wheat and *Brachypodium* (track 1) using heat maps (Brenchley *et al.,* 2012**)**. These plots allowed easy identification of SNP or syntenic hotspot/dropout regions at a glance. Figure 1.7 additionally plots syntenic regions between the wheat genome sequence and the *Brachypodium* genome onto the *Brachypodium* chromosomes to allow visualization.

The circular effect of a Circos plot can allow highlighting of relationships between pairs of positions using ribbons (encode position, size and orientation of related elements). Circos can display data as scatter, line or histogram plots, heat maps, tiles, connectors and text (Krzywinski *et al.,* 2009). Within this project Circos will be used primarily to demonstrate SNP distribution across a genome and SNP conservation between genomes. Its heat map ideograms allow clear visualization of SNP frequency and histogram or line plots can also be used effectively to outline the depth of sequencing coverage on a chromosome-by-chromosome basis across the genome.

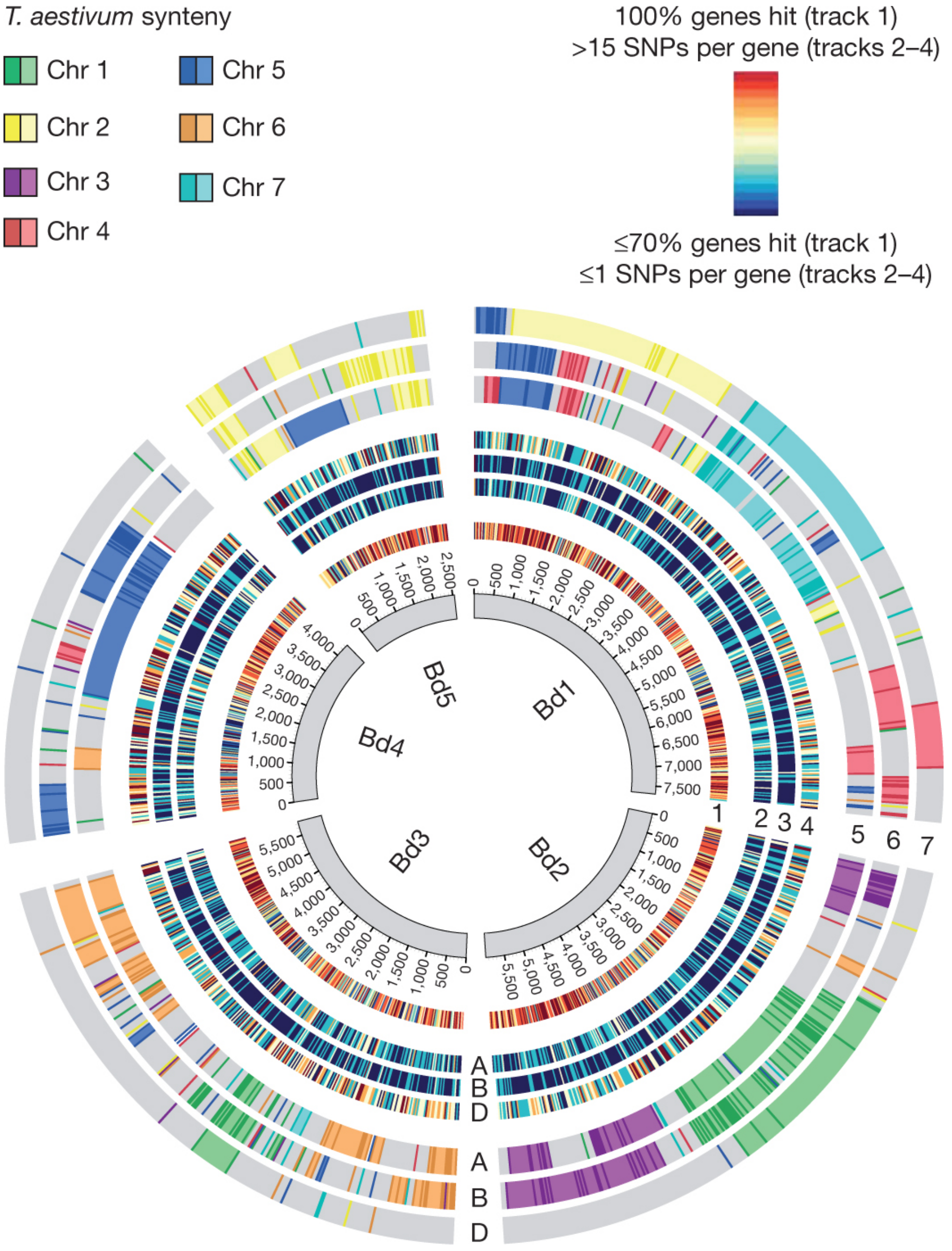


Figure 1.7. Example of a Circos plot: Alignment of wheat 454 reads, SNPs and genetic maps to the *Brachypodium* genome. Taken from Brenchley *et al.,* 2012. The inner circle represents the 5 *Brachypodium* chromosomes. Track 1 shows wheat and *Brachypodium* gene conservation per window of 20 genes in wheat. Tracks 2-4 show SNP density in wheat (average SNP number per gene in a window of 20 genes) in the A (track 2), B (track 3) and D (track 4) genomes. Tracks 5-7 show wheat-*Brachypodium* syntenic regions for the A (Track 5), B (track 6) and D (track 7) genomes. Genetic markers for each chromosome are shown in darker colours.

1.4 Arabidopsis

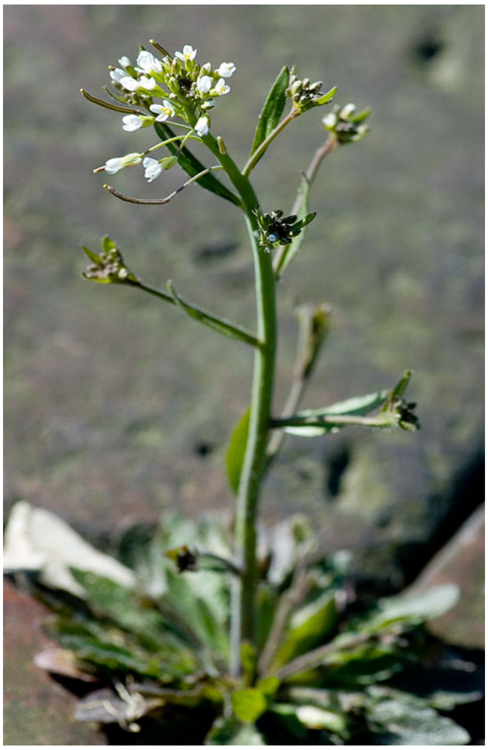


Figure 1.8. Image of the model diploid plant *Arabidopsis thaliana.* Taken from: Delaware Wildflowers at http://delawarewildflowers.org

Arabidopsis thaliana is a dicotyledonous species, a member of the Brassicaceae or mustard family. It is a model system and has been a focal point for laboratory studies of the cellular and molecular biology of flowering plants due to its possession of several desirable traits:

-It requires only light, air, water and low-level nutrients to complete its life cycle

-It has a rapid life cycle (approximately 6 weeks from germination to mature seed)

-It has limited space requirements and is easily grown in a greenhouse/indoor growth chamber

-It has prolific seed production

-It has relatively small genome (125 Mb)

-There are extensive genetic and physical maps of all of its 5 chromosomes (The Arabidopsis Genome Initiative, 2000)

-It has a large number of mutant lines and genomic resources many of which are available from Stock Centers

-It has a multinational research community of academic, government and industry laboratories

-It has a genome that can be manipulated through genetic engineering with great speed and ease (The National Science Foundation, 2009).

Finally, *Arabidopsis* has a dedicated website, The *Arabidopsis* Information Resource (TAIR), that maintains and regularly updates a database of it's genetic and molecular biology data. This includes the complete genome sequence along with gene structure, gene product information, metabolism information, gene expression data, DNA/seed stocks, genome maps, genetic/physical markers, publications and information about the *Arabidopsis* research community (TAIR, 2011). *Arabidopsis* was the first higher plant and only the third multicellular organism for which a complete finished genome sequence was published and is, as detailed above, an ideal model organism for biological research (Bevan and Walsh, 2005). Model systems have assisted understanding of biological processes at genetic, molecular and systems levels.

The reference *Arabidopsis* genome sequence that was produced in 2000 is from the Col-0 accession and was published by The *Arabidopsis* Genome Initiative. This international sequencing project implemented large-insert bacterial artificial chromosome (BAC), phage (P1) and transformation-competent artificial chromosome (TAC) libraries i.e. libraries of constructs in which fragments of DNA up to ~300Kb, in this case *Arabidopsis* fragments*,* can be cloned into bacteria where they are amplified (Xu, 2010). Several methods were used to identify overlapping BACs and positional information for clones in order to assemble physical maps of the *Arabidopsis* genome; the first is restriction fragment analysis where clones are treated with restriction enzymes and the resultant fragment sizes are analyzed. The degree of shared fragments between clones can allow definition of the degree of overlap between them to allow assembly into contigs (Marra *et al.,* 1999). The second method involves the hybridization of a clone to a southern blot containing digested DNA from the other clones. If hybridization occurs clonal overlap has been found (Bent *et al.,* 1998). Finally the third method utilizes sequence-tagged sites (STS) that are DNA sequences that have a single occurrence in the genome at a known location and therefore act as markers. STS's can be easily detected using PCR and if present in a clone sequence allow easy positional anchoring of it to the genome (Mozo *et al.,* 1999; Sato *et al.,* 1997).

The physical maps of the *Arabidopsis* genome were used with genetic maps to generate sequencing tiling paths; end sequence of BAC clones was used to assist integration of contigs and the majority of the genome could be represented by the final set of 10 contigs, representing chromosome arms, that were assembled largely from BAC and P1 clones. Subsequent sequencing of fragmented clones and BAC end sequences (supplemented by genetic mapping in centromeric regions (Copenhaver *et al.,* 1999)) allowed contig assembly and genetic markers allowed sequence verification. The sequenced region totaled ~115.4 Mb with ~10Mb un-sequenced centromeric and rDNA repeat regions (The Arabidopsis Genome Initiative, 2000). The *Arabidopsis* reference sequence is under continuous development and now covers over 119 Mb of the 125 Mb genome and includes 27,416 protein coding genes and 41,671 gene models (TAIR, 2011).

Over 750 natural accessions of *Arabidopsis* have been found. These accessions are quite variable in terms of form and development (e.g. leaf shape, hairiness) and physiology (e.g. flowering time, disease resistance). The commonly used background lines include Landsberg, Columbia, and Wassilewskija (TAIR, 2011).

Within this project the *Arabidopsis* genome is largely used to test and validate pipelines and methodologies prior to their application to the complex and large genome of wheat. "Large genomic resources like EST and full length cDNA databases, large collections of characterized mutants from genetic screens and insertion mutants as well as a huge set of expression data covering numerous environmental conditions and developmental stages make *Arabidopsis* also an excellent source for functional and comparative genomics" it offers the chance to test hypotheses quickly and efficiently before they are implemented on more complex genomes (Spannagla *et al*., 2011).

1.5 Brachypodium

Figure 1.9. Image of the model diploid plant *Brachypodium distachyon.* Taken from: James and the giant corn at http://www.jamesandthegiantcorn.com/2010/02/11/why-to-celebrate-the-publication-of-the-brachypodium-genome/

Brachypodium distachyon is a member of the grass subfamily Pooideae. It is a wild grass and it has become a widely recognized model plant for important crops such as wheat and barley due to its desirable traits, many of which are shared with Arabidopsis:

-It has a fully sequenced genome

-It has a small, compact, diploid genome (~272 Mbp)

-A short life cycle and few growth requirements

-It is in the same subfamily as economically important cereal grain species such as wheat and Barley (Kersey *et al.,* 2013)

-Easily experimentally manipulated

-Extensive genetic and physical maps of all 5 chromosomes

-Large number of mutant lines and genomic resources

-Multinational research community of academic, government and industry laboratories (International *Brachypodium* Initiative)

Finally, *Brachypodium,* like *Arabidopsis,* has a dedicated website, Brachypodium.org, that maintains and regularly updates a database of it's genetic and molecular biology data. This includes the complete genome sequence along with annotations including gene predictions, SNPs, structural variants and indels.

The *Brachypodium* genome was sequenced in full by the International *Brachypodium* Initiative and published in 2010. This sequencing initiative implemented similar techniques to those used to publish the *Arabidopsis* genome utilizing genetic maps, physical maps and sequenced BACs to confirm sequencing assemblies. However on this occasion the diploid inbred line of Bd21 was sequenced using whole-genome shotgun sequencing. Sequencing reads were assembled into scaffolds and comparison with genetic maps, physical maps and sequenced BACs detected false joins, created further joins and validated the assembly to yield 5 pseudomolecules spanning 272 Mb (The International Brachypodium Initiative, 2010).

Brachypodium has been used within this project for the unique purpose of anchoring wheat sequence to specific locations in the wheat genome. Due to its high synteny with the wheat genome, *Brachypodium*-wheat markers can be used for such comparative genomic organization of wheat using a genome for which a complete and well-annotated reference genomic sequence is available. Rice (*Oryza sativa* L.), a model, well-characterized species, has been previously used in comparative wheat analyses for molecular mapping and gene isolation (Liu and Anderson, 2003). Synteny and gene homology and order between rice and the other cereal genomes, e.g. wheat, is extensive (Goff *et al.,* 2002) but numerous studies show that co-linearity between the two species can frequently break down due to translocations, deletions and gene duplications (Bennetzen and Ma, 2003). The *Brachypodium* genome has become a popular option to rice due to its completed and annotated genomic sequence and data suggesting that better co-linearity exists between it and wheat than between rice and wheat (Caol *et al.,* 2012).

1.6 Wheat

Wheat is the dominant cereal crop grown in temperate countries and, with a global output of 681 million tonnes in 2011 (United States Department of Agriculture, 2012), is one of the most important crops for human and livestock feed (Shewry, 2009). The world is facing a potential crisis in terms of food security. With a population that is projected to reach 9 billion by 2050 the challenge is to produce and supply enough safe and nutritious food in a sustainable way (Foresight, 2011), making it a top priority to increase wheat yields (Allen *et al*., 2011). It is estimated that, in Europe, wheat production must double to keep pace with demand and to maintain stable prices. This increasing demand for wheat is challenged by a shortage of high quality agricultural land, increased fertilizer costs, disease, resource limitations and environmental issues that dramatically reduce optimal yields.

The common bread wheat genome is allohexaploid and is one of the largest higher plant genomes at ~17Gb in size (128 times larger than *Arabidopsis* and 5 times larger than *Homo sapiens*) (Dubcovsky and Dvorak, 2007). It's large genome size, polyploid complexity and high repetitive sequence content (~80-90%) poses challenges to the researcher including expense and difficulty in sequencing as well as in isolation and cloning of mutant loci (The National Science Foundation, 2009; Smith and Flavell, 1975). The recent sequencing and gene identification analysis of the wheat genome using 454 pyrosequencing (Brenchley *et al.,* 2012) has increased research prospects and opened up the cost effective possibility of utilizing targeted genome capture re-sequencing arrays to analyze wheat genes specifically.

Within this project several common UK hexaploid wheat varieties are analyzed including; Truman, a red winter wheat variety released by the University of Missouri Agricultural Experiment Station. It stands well in most environments producing high yields (Wisconsin Crop Improvement Association, 2011). Rialto, a winter wheat variety with good yields, relatively long straw (Hoad *et al.,* 2006) and good bread-making quality (Nickerson Ltd, 2007). Utmost, a Canadian Western red spring wheat variety that was developed by the wheat breeder Pierre Hucl and is known to have good yields and to contain the Sm1 gene to confer orange wheat blossom midge tolerance (Canadian Food Inspection Agency, 2013). Finally, Chinese Spring, a spring wheat variety that is widely used in genome studies and was the variety that was sequenced to 5X coverage in 2010 at the University of Liverpool by Brenchley *et al.* in the effort to develop a reference sequence for wheat (Brenchley *et al.,* 2012**)**.

Winter wheat seeds are known to have strong dormancy that prevents germination. They can be planted in Autumn and will survive the winter months to mature in the spring for harvest around summertime. In contrast spring wheat seeds have no or weak dormancy. They tend to be planted in the spring with harvest soon following in summer (Lei *et al.,* 2013). Higher yields are associated with the winter wheat varieties as a result although spring wheat does tend to offer a high quality for bread making (Columbia University Press, 2005).

1.6.1 Evolution of wheat

The allohexaploid (AABBDD) wheat genome is derived from three diploid progenitor genomes. It was produced from two separate hybridization events (figure 1.10). The AA genome is from *Triticum urartu*, the BB from an unknown species but likely to be of the Sitopsis section (includes *Aegilops speltoides*), and the DD from *Aegilops tauschii* (Brenchley *et al.,* 2012). In the first hybridization event AABB tetraploids appeared less than 0.5 million years ago (Dvorak *et al.*, 2006). It is thought that Emmer tetraploid wheat developed from the domestication of such natural tetraploid populations. The wheat that we have today formed around 8000 years ago by the hybridization of the unrelated diploid wild grass *Aegilops tauschii* (DD genome) with the tetraploid *Triticum turgidum* or Emmer wheat (AABB genome) (Dubcovsky and Dvorak, 2007). Each hybridization was followed by chromosome doubling in the new hybrid enabling normal bivalent formation at meiosis and thus production of fertile plants and it has been reported as likely that this hybridization occurred several times independently with the novel hexaploid (genome AABBDD) being selected by farmers for its superior properties (Shewry, 2009). The resultant hexaploid bread wheat carries 6 genomes each with 7 chromosomes and thus 42 chromosomes in total (Winfield, 2011). The coding regions of the 3 homeologous diploid wheat genomes share over 90% homology (Kawaura *et al.,* 2009).

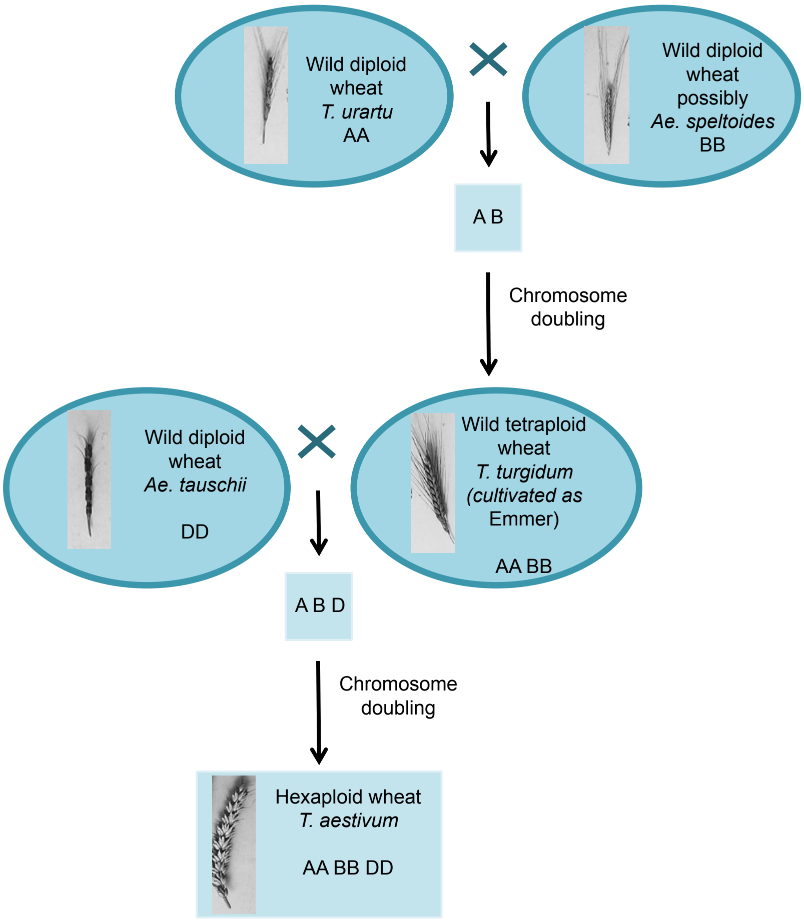


Figure 1.10. Evolution of hexaploid bread wheat. Detailing the 2 separate hybridization events that allowed the evolution of bread wheat *Triticum aestivum* (Winfield, 2011).

1.6.2 Gene enrichment in wheat

The wheat variety Chinese Spring was sequenced to 5X coverage primarily within the University of Liverpool in 2010 using Roche's 454 GS FLX. A draft wheat genome assembly was constructed from the 454 reads using the de novo DNA sequence assembly software package Newbler. Whole genome re-sequencing of wheat is prohibitively expensive due to its large size and repetitive nature. The volume of data that would be produced from such an analysis is a challenge to analyze and as such this method is unlikely to become the method of choice for determination of wheat mutants and for other wheat genome analyses. It is essential to allow analyses to be affordable and less prohibitively complex in an organism such as wheat with a large, repetitive, largely uncharacterized genome, where routine whole genome re-sequencing is an undesirable option and a completed reference is not yet available.

To overcome this limitation researchers have employed capture arrays to reduce the genome complexity. These allow enrichment of and thus the sequencing of targeted regions to high coverage, leading to confident SNP scoring. The utilization of enrichment in combination with mapping-by-sequencing as a mechanism to rapidly identify genes responsible for key agricultural traits one of the main challenges of this project. Such targeted sequencing in wheat reduces the cost associated with re-sequencing the entirety of the genome and the recent sequencing of the wheat genome using 454 pyrosequencing (Brenchley *et al.,* 2012) has opened up the possibility of utilizing targeted capture re-sequencing arrays for wheat analyses by generating a sequence around which an array can be designed.

Within this project capture arrays in solution are used to perform targeted re-sequencing however a selection of other methods exist that can effectively target subsets of a genome for sequencing and are detailed as follows; Firstly RNA sequencing or RNA-seq involves next-generation sequencing of RNA transcripts using reverse transcription to convert them into cDNA's that can be sequenced. This effectively targets transcribed regions for sequencing and provides information to reveal the transcriptional structure of genes and/or the level of expression for each gene (Wang *et al.,* 2009). Secondly, restriction site associated DNA sequencing involves using next-generation sequencing technology to only sequence at specific sites that are defined by restriction enzymes. Such methodology can identify and score thousands of genetic markers across the genome from one or a group of individuals i.e. enriching for potential marker regions (Davey and Blaxter, 2010). Cot filtration uses the principles of DNA renaturation kinetics to separate repetitive DNA sequence from gene rich regions to allow selective sequencing of non-repetitive genic regions. It utilizes the fact that a specific sequence will renature at a rate proportional to the number of times it occurs in the genome i.e. repetitive sequences more quickly than non-repetitive, and this technique has been successfully applied to wheat by Lamoureux *et al.,* although theoretically it may select against large gene families that become normalized. Finally methylation filtration in wheat produces genomic libraries enriched in hypomethylated, typically genic, sequence using a bacterial methylation-dependent restriction endonuclease. This method eliminates hypermethylated repetitive sequence, however, it can also eliminate hypermethylated gene sequence (Rabinowicz *et al.,* 2005).

The NimbleGen exome capture array in solution that was utilized within this project was developed for the wheat genome (~41 Mbp) using the Roche custom array design service. The exome is comprised of the coding exons in the genome i.e. the small sections of DNA that encode for proteins. Current knowledge of the genome reveals a large majority of DNA changes causing genetic diseases are within the exome, which is why it is often referred to as the most relevant portion of the genome (NimbleGen, Roche, 2011). This initial array design was based on cDNA as at the time the full genomic sequence of wheat was unavailable. The exome capture array acts as a proof of principle that such a method is applicable for gene enrichment and subsequent SNP calling in wheat. It also allowed development of array design methodology. The issue with using cDNA sequence is that only the subset of genes that were expressed in the sequenced sample at the time of processing will be included in the array.

The development of the exome capture array is detailed in Figure 1.11b. This initial array was based on the cDNA sequence that was generated for the hexaploid wheat strain Chinese Spring by the Roche 454 sequencer. This cDNA was assembled into contigs using the software package Newbler that is specialized for de novo DNA sequence assembly of 454 reads. The cDNA contigs that were generated by Newbler were then taken through a BLAST search against various databases to aid elimination of duplicate sequence, repetitive sequence or sequence from the chloroplast/mitochondrial genomes. Homeologous genes were collapsed into 1 sequence to allow generation of a single probe set that is capable of enriching all 3 wheat genomes. The remaining contigs were filtered for regions of low complexity and used as the basis for the array probe targets. Array probes were tiled across these target sequences (known as design-space sequence). Winfield *et al.* hypothesized that if the wheat exome is thought to represent about 170–340 Mb and the wheat exome capture array contains only unique sequences, as it was designed to do, and given that the total length of the features on the array is ~57.5 Mb; it is therefore capable of capturing 50% of the genes in a diploid exome as a minimum and that this figure could be potentially much higher.

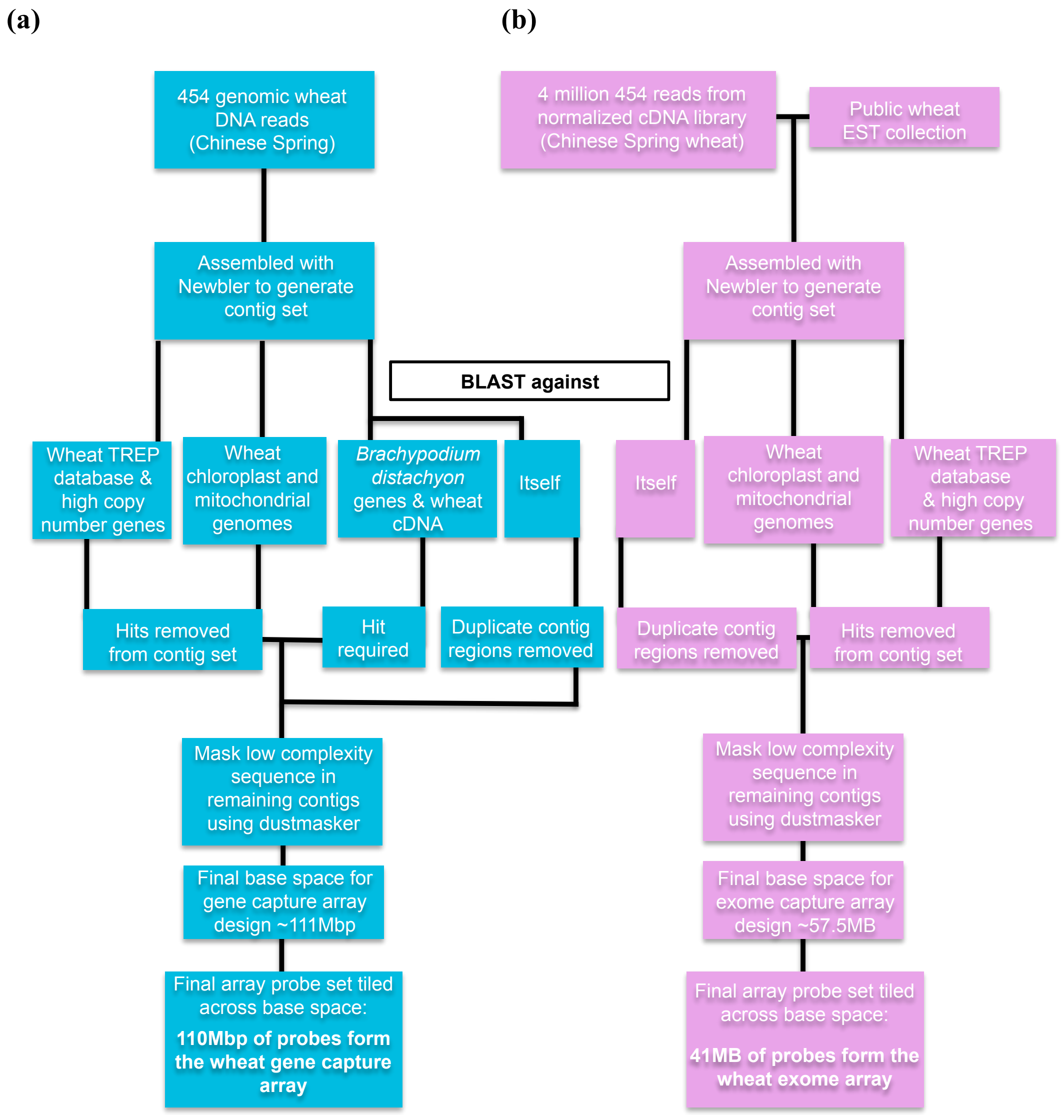

(a)
(b)
454 genomic wheat DNA reads (Chinese Spring)
Assembled with Newbler to generate contig set
BLAST against
Wheat TREP database & high copy number genes
Wheat chloroplast and mitochondrial genomes
Brachypodium distachyon genes & wheat cDNA
Itself
Hits removed from contig set
Hit required
Duplicate contig regions removed
Mask low complexity sequence in remaining contigs using dustmasker
Final base space for gene capture array design ~111Mbp
Final array probe set tiled across base space:
110Mbp of probes form the wheat gene capture array
4 million 454 reads from normalized cDNA library (Chinese Spring wheat)
Public wheat EST collection
Assembled with Newbler to generate contig set
Itself
Wheat chloroplast and mitochondrial genomes
Wheat TREP database & high copy number genes
Duplicate contig regions removed
Hits removed from contig set
Mask low complexity sequence in remaining contigs using dustmasker
Final base space for exome capture array design ~57.5MB
Final array probe set tiled across base space:
41MB of probes form the wheat exome array

Figure 1.11. Development of 2 wheat capture arrays in solution. (a) Development of a wheat gene capture array that is based on the genic regions of wheat (transcribed and non-transcribed) **(b)** Development of a wheat exome array that is based on wheat cDNA sequence.

As the wheat 454 genomic sequence assembly became available in 2010 a further array in solution was developed using similar techniques. This array was to contain the majority of the genic regions of wheat (transcribed and non-transcribed) and as such will be referred to as the wheat gene capture array.

The development of the wheat gene capture array is detailed in Figure 1.11a. The genomic DNA wheat 454 reads were assembled into contigs using Newbler. The contigs were selected that hit genes in the closely related species *Brachypodium distachyon* or that hit the wheat cDNA sequence when BLAST searches were conducted. Homeologous genes were again collapsed into 1 sequence and the remaining contigs were processed as for the wheat exome capture array to eliminate redundancy, repetitive sequence and chloroplast/mitochondrial sequence. The final contig set that remained is referred to as the design-space sequence and was used as the target sequence for wheat gene capture array. Array probes were tiled across these target sequences.

The wheat gene capture array is a clear improvement on the exome capture array; it will contain all of the genes that were sequenced whether they were expressed or not expressed and it allows inclusion of exon and intron sequence in the array. With growing importance being attributed to polymorphisms affecting splice sites (Baralle and Baralle, 2005) that are mostly located in introns, the analysis of SNPs in intronic sequence along with those in exonic sequence is desirable.

Here custom capture probe sets had to be designed due to there being no pre-existing marketed whole exome or genic capture array for wheat. NimbleGen's capture probe sets utilize DNA probes that are typically less than 100bp in length and tiled across the genome i.e. high density overlapping baits. This allows sensitive detection of small variants. NimbleGen will work closely with a researcher that is developing an array to allow custom development and validation of a capture probe set for an organism of interest and to ensure an effective end product. The main rival for custom capture probe design is Agilent (Sure select). These capture probe sets utilize 120bp RNA baits that can also overlap. Agilent, at the onset of this study were less involved in the development process of custom capture probe sets, with the design being mainly down to the customer. Over the course of this study Agilent have become more involved in the design process. For the development of the exome and genic wheat capture arrays NimbleGen was a clear favourite, at the time, due to the added design assistance and reported lower off-target enrichment/greater on-target enrichment when utilizing such methodology compared to an Agilent Sure select capture probe set. Clark *et al.* found that 7.2% more of the targeted bases were covered by NimbleGen enriched sequence at a depth of 10x or more and 3.5% less off-target sequence was generated compared to using Agilent enrichment (Clark *et al.*, 2011). In section 5 an Agilent Sure select custom capture probe set was designed and utilized for a methylation study in preference to a NimbleGen array since at this time NimbleGen did not support the use of their capture probe sets for the study of methylation patterns using bisulfite treatment.

Capture arrays are used to enrich target sequence by hybridization to bait probes. After hybridization, any DNA that has not hybridized is washed away and the desired captured DNA can be eluted for downstream processing e.g. sequencing. Arrays can be designed using bait probes that are on a solid support (Winfield *et al.,* 2012) or more recently in solution (Sulonen *et al.*, 2011). In the latter case, that is implemented in this project, a standard DNA fragment library is hybridized to biotinylated bait probes in solution and streptavidin beads are used to collect the complexes of probes and bound DNA fragments from which the enriched DNA fragment pool are eluted. Saintenac *et al.* reported the use of SureSelect, an in-solution targeted capture technology to examine a tetraploid wheat genome whilst Winfield *et al.* used the NimbleGen in-solution targeted capture genomic DNA probe set that is detailed here that was designed to capture a significant proportion of the wheat exome (Saintenac *et al.,* 2011; Winfield *et al.,* 2012).

1.7 Identifying mutations responsible for traits in complex organisms

1.7.1 Summary

Organisms vary in a multitude of ways including morphology, behavior, physiology, development, and susceptibility to disease. Some of these phenotypes are controlled by a single gene and are known as monogenic (Mendelian) traits whilst other phenotypes are controlled by multiple genes and as such are known as mulitgenic or genetically complex traits (Glazier *et al*., 2002). All of the mutants that will be analyzed here in the two key genomes of interest, wheat and *Arabidopsis,* are monogenic. Forward genetics, which is used in this project, identifies the underlying genotype that is responsible for a given phenotype.

Many of the mutants that are analyzed in the *Arabidopsis* genome are ethylmethanesulfonate (EMS) induced or deletion mutants. EMS is a well-known alkylating agent that is commonly used to induce point mutations in plants. Treatment with EMS results in a high mutation frequency without preference for specific genomic regions (Watanabe *et al.,* 2007; Kim *et al.,* 2006). If a selected plant mutant phenotype has been EMS induced in a diploid plant the cause of this is anticipated to be a point mutation or homozygous SNP (single nucleotide polymorphism). In contrast to this the mutants that are analyzed in the wheat genome are typically sources of natural variation that induce a given phenotype.

The pipelines for mapping and SNP calling of next generation sequencing data that will provide the foundation of this project have been outlined. Besides providing a knowledgebase about the plant being studied in general and allowing comparative analyses between strains etc., such analyses provide the initial information that is required to perform downstream analyses such as mutant identification. Here mutant identification will primarily be carried out using mapping-by-sequencing, a newer technique that vastly decreases time input in comparison to the more outdated methods such as map-based cloning. Mapping-by-sequencing analyses and other methodologies are outlined here. Such methods can be used with DNA samples that have been prepared in a variety of ways depending on the desired output of the analysis i.e. with a variety of plant crosses. As such, a background to plant breeding, with focus on the techniques employed within this project, will also be outlined. Sliding window mapping-by-sequencing analyses have largely benefitted from a complete reference sequence and employed whole genome re-sequencing of diploid organisms (Nordstrom *et al.,* 2013). The feasibility of the application of such studies to an enriched polyploid wheat dataset will be tested in this project.

1.7.2 Methods for identifying traits

Until recently, identifying a mutated gene required the tedious process of map-based cloning that can take, from skilled post-doctorial research, from 6-12 months. The original idea of map-based cloning was to identify a molecular marker linked to the gene of interest as markers consistently associating with the mutant phenotype with low recombination frequency indicate close proximity to the mutant allele i.e. in linkage disequilibrium (Maniatis *et al.,* 2004). Chromosome 'walking' would then be performed to the gene using overlapping clones (e.g. YAC/BACs) i.e. using a probe for a marker near a gene to select a genomic clone near the gene and moving toward the gene by repeatedly selecting for overlapping clones until you have a clone that contains the gene. This method of chromosome 'walking' proved difficult to apply to larger, more complex uncharacterized plant genomes and was only necessary under the assumption that the markers that were identified were not physically close to the gene of interest. With the advent of physical molecular linkage maps (also comparative genetic linkage maps-allowing determination of the relative order of genes among related species that evolved from a common ancestral genome) markers that were tightly linked to genes of interest could be identified. Markers could then be used to screen YAC/BAC libraries and isolate the clone containing the gene of interest directly. This is particularly useful for the *Arabidopsis* genome for which both genetic and physical maps had been generated along with a complete reference sequence by 2000 (The Arabidopsis Genome Initiative, 2000) and this became the main strategy by which map-based cloning could be applied to isolate genes underlying complex traits in plant species for which such information was available (Tanksley *et al.,* 1995).

Map-based cloning in *Arabidopsis* involves crossing the mutant plant with a divergent accession e.g. Ler-0 to create a mapping population. Different crosses can be carried out these include outcrossing, backcrossing or introgression. Outcrossing can introduce unrelated genetic material into a breeding line by crossing two genetically unrelated individuals. This tends to always be the first step in a linkage analysis (Mooney and McGraw, 2007). Outcrossing to produce an F1 population can be followed by backcrossing of the F1 population (where an F1 hybrid strain is crossed to one of its parental strains) to produce an F2 population. Introgression involves the gene flow from one species to another by repeated backcrossing of a hybrid with one of its parents (Frisch and Melchinger, 2005). Outcrossing to produce an F1 population can alternatively be followed by one or more intercrosses i.e. F1 brother and sister crosses to produce a highly homozygous F2 etc. During map-based cloning in the F2 generation the mutant phenotype is scored and molecular markers are then used to rough map the gene i.e. markers consistently associating with the mutant phenotype. Plants with intra-chromosomal recombination events can be used to narrow down the genetic interval (Lukowitz *et al*., 2000*)*. This process becomes difficult if natural variation exists in the phenotype being mapped between the two parental lines or if the mutant phenotype is subtle and assaying for it is labour intensive (Ashelford *et al.*, 2010). Ashelford *et al.* encountered such a problem when determining a novel circadian clock mutation in *Arabidopsis*. In this case simply identifying the mutant phenotype in a set of plants could be difficult and time consuming.

There was an obvious requirement for method development with regard to identifying mutations responsible for traits at this point since the process of genetic mapping, linking a phenotype to its causal mutation, is widely acknowledged to be long-winded and time-consuming. Direct re-sequencing has already been successfully used as an alternative to map-based cloning to identify point mutations in the 15.4 Mb genome of the yeast *Pichia stipitis* (Smith. *et al.,* 2008) and in *Caenorhabditis elegans* (Sarin *et al.,* 2008). Whole genome re-sequencing approaches like that of Sarin *et al.* are of limited use if, like in *Arabidopsis*, the EMS mutation load is high. Therefore, a method of reducing the number of point mutations must be considered in these cases. However, it would still rely on the ability to accurately score mutants in an F2 mapping cross and has all the limitations with regards to map-based cloning in this respect (Ashelford *et al*, 2010). In an 'in house' study mentioned previously by Ashelford *et al.* the 120 Mb genome of a novel *Arabidopsis* clock mutant early bird (*ebi-1*) and the corresponding wild type, Wassilewskija (*Ws-2*) were re-sequenced using the Applied Biosystems SOLiD sequencing by ligation technology. Sequencing a backcrossed line reduced the number of point mutations, investigating gene expression data for mutated genes further narrowed the SNPs down and finally the new SNP data was used to exclude a known clock gene and identify a SNP in the gene AtNFXL-2 as the likely cause of the *ebi-1* phenotype.

Mapping-by-sequencing is the process of taking this direct re-sequencing approach and combining it with genetic mapping. The direct re-sequencing that has been previously outlined is one example of such methodology that can be used to 'roughly' map the causal mutation. Other methodologies involve sequencing a bulk segregant or pooled mutant population (typically an F2 population that has been previously outcrossed and then selfed or backcrossed) to uncover the mutation, potentially a single nucleotide polymorphism (SNP) that is responsible for the phenotype. This approach benefits from a reference parent genome that the mutant dataset can be mapped to (Nordstrom *et al.,* 2013). SNPs can then be called from the mapping output and used to study allele frequencies and to define regions of conserved homozygosity with the mutant dataset.

Bulk segregant analysis can be used to map monogenic traits and involves a pooled DNA sample of typically F2 individuals from a cross who share a single phenotypic trait of interest differing from the normal population and as such they will differ genetically only at the trait locus. All of the pooled plants with the desired trait will have high homozygosity in this region that is conserved with the mutant parent, with a high frequency of heterozygosity across the rest of the genome, while in the pool without the specific trait such conservation with the mutant parent will not be seen (Quarrie *et al.,* 1999).

It is common for these techniques to become integrated, for example, in assays to find a phenotype inducing mutation or loci that are demonstrated here; Parent 1 is crossed with parent 2 that contains a desirable phenotype (outcrossing to produce an F1 population). The progeny of the cross is crossed back to parent 1. The F2 progeny of this cross is selected for the desirable phenotype and then crossed back to parent 1 and such backcrossing is repeated. The aim of this being to create a line as identical as possible to parent 1 while still having the donor gene of interest from parent 2 enabling easier identification of the gene by narrowing the donor interval. This method can be combined with bulk segregant analysis for monogenic traits to allow further ease of donor gene identification through looking at allele frequency distribution. High heterozygote frequency across the genome due to pooling of samples allows the homozygous region harbouring the trait of interest to be easily identified (Michelmore *et al.,* 1991).

Earley & Jones used phenotype-based selection and introgression to backcross a food preference loci in *Drosophila simulans* into *D. sechellia* (has opposite food preference at loci). Populations of *D. simulans* and *D. sechellia* were hybridized and then selected for the food preference phenotype across multiple generations of backcrosses. The trait of interest (*D. simulans* food preference in *D. sechellia*) was selected in each generation and the offspring mated with *D. sechellia*. Backcrossing continued for 15 generations and then the final generation was inbred for 2-3 generations to ensure introgressed loci were mostly homozygous. DNA from 30 females was then pooled and sequenced using Illumina technology. By looking for enrichment of *D. simulans* SNPs in a *D. sechellia* background the breakpoints of introgressions were identified and thus the regions harbouring the genes influencing the trait (Earley and Jones, 2011).

The programs SHORE and SHOREmap enable mapping-by-sequencing in one streamlined pipeline that allows simultaneous mapping (SHORE) and mutation identification (SHOREmap) to output an approximate mapping interval that would contain the phenotype inducing mutation plus a prediction of the phenotype inducing mutation itself (see 1.3.2 Analyses downstream of mapping). This program is intended for use on diploid organisms with a reference sequence (Schneeberger *et al.,* 2009). SHOREmap implements an analysis of local differences in parental allele frequencies across a complete reference sequence that have been introduced via mutant phenotypic selection (Galvao *et al.*, 2012). With use of a bulk-segregated F2 population that is derived from an initial outcross and subsequent selfing of F1 offspring Schneeberger *et al.* identified a mutation in *Arabidopsis* that was causing slow growth and light green leaves owing to lesions in an unknown gene using SHOREmap. They sequenced a single genomic DNA sample (prepared from mutant F2 plants) using the Illumina® (Genome Analyzer^{TM}), produced an 'interval' plot of the relative allele frequencies of the mapping parents to reveal a candidate region of interest on chromosome 4. They then used mutations relative to the reference within this region as input for SHOREmap 'annotate' (ranks base pair substitutions by distance from allele distribution peak and predicts base change effects). SHOREmap detected a mutation in the AT4G35090 gene as the causal mutation (Schneeberger *et al.,* 2009).

Such mapping-by-sequencing analyses have largely required complete chromosome sequences and have involved the use of whole genome re-sequencing. In species such as wheat, no finished genome reference sequence is available. Additionally due to its vast size and highly repetitive content, complexity reduction methods such as targeted enrichment sequencing have been proposed to reduce the need for whole genome re-sequencing analyses that are prohibitively expensive. In this analysis it is proposed that routinely using the gene capture array to target wheat genic regions prior to sequencing will greatly reduce the cost associated with sequencing the wheat genome eliminating much of the repetitive sequence from the analysis while still, essentially, allowing mapping-by-sequencing analyses to be performed. As such it is this application of mapping-by-sequencing to, firstly, a polyploid and, secondly, an enriched dataset that is of interest here.

Trick *et al.* employed the use of Near Isogenic Lines (NILs) that are created by outcrossing an organism with a phenotype of interest with the wild type parent to produce an F1. The F1 population is intercrossed to produce an F2 population and F2 offspring with the phenotype of interest are repeatedly backcrossed to the wild type parent to generate a NIL line that will be almost identical to that of the wild type parent except for the genomic segment harbouring the phenotype-inducing gene. In this project they sequenced the mRNA (a technique known as RNAseq) of tetraploid wheat lines that differed for the grain protein content gene (*GPC-B1*) in order to identify the region containing the gene of interest. They identified inter-varietal SNPs between the parental lines and examined the relative frequencies of these SNPs in two bulked samples of near isogenic lines (NILs) differing for the GPC phenotype. Marker assays were designed for any enriched SNPs and they were mapped to using each set of bulked DNA leading to identification of a ~0.4cM interval including ~70% of the SNPs and the gene of interest in wheat with use of synteny of marker sequences with the closely related grass *Brachypodium* (Trick et al. 2012).

Galvao *et al.* demonstrated the enrichment of a subset of marker linked genomic DNA sequences and that mapping of this data back to the marker sequences could successfully identify an interval of interest in *Arabidopsis*. They also found that by sorting *Arabidopsis* cDNA sequences into *Brassica rapa* based pseudo-chromosomes, using synteny between the two, a confidence interval could be identified in *B. rapa* that could be translated back to a position in *Arabidopsis*. They hypothesized that enrichment and mapping-by-sequencing analyses, based on the SHOREmap methodology, were compatible however likely to need additional fine mapping due to likely exclusion of the target region from the enriched dataset (Galvao *et al.,* 2012). These approaches have yet to be combined and fully tested for a mapping-by-sequencing analysis i.e. full gene enrichment and organization into pseudo-chromosomes, based on synteny with a related organism, in a polyploid species such as wheat, that is lacking in extensive annotation and a finished genome reference sequence. For their enriched dataset Galvao *et al.* analyzed allele frequency at each individual marker position to look for homozygous regions, as did Trick *et al.* and sliding window analyses were reserved for whole genome sequencing projects with a defined reference genome (Galvao *et al.,* 2012; Trick et al. 2012). Within this project the wheat pseudo genome is utilized to perform a sliding window analysis of allele frequencies in the mapping population along each pseudo-chromosome to identify a region in wheat directly. The extensive coverage of the majority of wheat genetic sequence in the gene capture array will allow enrichment whilst maintaining the likelihood of inclusion of the target region within the dataset.

1.8 Wheat methylation studies

Methylation of the cytosine residues in eukaryote DNA is thought to act as a mechanism of gene expression control. In plants, it occurs typically at CpG (Finnegan *et al.,* 1998) residues but can also occur at CpNpG sites, where N is any nucleotide, and any CpHpH site, where H represents adenine, cytosine or thymine (Lukens and Zhan, 2007). It was noted in *Arabidopsis* that most methylation that would occur in the gene body was mainly at CpG sites, whilst methylation elsewhere and in repetitive regions could be at CpG, CHH and CHG sites (Widman *et al*., 2009).

Cytosine methylation within gene promoter regions is thought to inhibit binding of regulatory proteins and repress transcription; it can also silence the transposable elements (TEs) that would otherwise disrupt DNA sequence by transposition. TE transposition can result in altered gene expression, novel regulatory networks, gene deletions, duplications, increases in genome size, illegitimate recombination and chromosome breaks/rearrangements (Cantu *et al.,* 2010). As such, reduced DNA methylation is known to disrupt normal plant development. Methylation within introns and downstream exons has been highly correlated and if such gene body methylation is found it has been associated with highly expressed genes in some studies (Zhang *et al.,* 2006) while other studies have found little or no association in this context. Brenet *et al.* discovered that methylation downstream of the transcriptional start site i.e. in the first exon region was strongly linked to gene silencing, even more so than methylation of the upstream promoter region. The effects of gene body methylation can be seen to remain controversial and largely without clarification. It is clear that the location of methylation within or around a gene is important, however, the reasoning for this is, as of yet, poorly understood (Brenet *et al.,* 2011).

In a study by Rabinowicz *et al.* a small subset of whole genome sequencing data for hexaploid wheat was analyzed and subsequently methylation filtration was utilized in an attempt to isolate hypomethylated genic material from hexaploid wheat for sequencing. Isolated sequenced material was later used in a BLAST search to identify wheat genes. From the whole genome sequencing data wheat was found to have a large number of gene-like sequences relative to other plants (1597 sequencing reads, ~500bp in length, 1.44% genes) while in the enriched data gene enrichment was comparatively low (1548 sequencing reads, ~500bp in length, 6.78% genes). They predicted that the apparent excess of genes combined with poor enrichment could be due to high levels of methylated pseudogenes (recently amplified and then silenced), reducing the number of active genes to a level closer to that which was expected (Rabinowicz *et al,* 2005).

Here the study of methylation patterns in wheat was to be used to test a number of hypotheses; firstly, if differential methylation exists between the A, B and D genomes. Secondly, using two growth temperatures for the Chinese Spring to test if temperature is capable of altering the methylation state and to see if this is both genome specific and genome independent. Finally, to investigate if it is this underlying methylation that can control both genome specific and temperature dependent changes in gene expression.

There are three main methods used in the laboratory for the study of methylation patterns;
-Bisulfite treatment deaminates un-methylated cytosine residues converting them to uracil. The conversion of these un-methylated but not methylated residues to uracil allows, after PCR and sequencing, effective discrimination of the methylation status at every cytosine residue making this method the gold standard in methylation studies (Darst *et al.,* 2010).
-Differential enzymatic cleavage uses methylation-sensitive restriction enzymes to fragment genomic DNA that can then be analyzed. Enzymatic cleavage is limited by the number of enzyme recognition sites (New England Biolabs, 2009).
-Affinity based methods use antibodies or proteins that bind to methylated DNA resulting in the enrichment of the methylated DNA in the experimental sample to allow downstream analysis (New England Biolabs, 2009).

The clear significance of the impact of methylation on the genome makes it an obvious area for research. In order to study the general effects and patterns of methylation in the hexaploid wheat, without encountering the problems previously detailed due to the large size of the wheat genome, development of a methylation target enrichment array would be the best way forward i.e. the ability to enrich for regions of interest in the genome and to study methylation patterns therein. Sodium bisulfite treatment is an increasingly popular method for epigenetic profiling and combined with the use of Agilent's SureSelect Methyl-Seq Target Enrichment System allows the study of methylation patterns in target regions. The Agilent enrichment system utilizes 120bp biotinylated RNA baits in solution to capture user-defined regions based on primary DNA sequence. In this system a standard DNA fragment library is hybridized to biotinylated bait probes in solution and streptavidin beads are used to collect the complexes of probes and bound DNA fragments from which the enriched DNA fragment pool are eluted. The enriched DNA fragments are then bisulfite converted; PCR amplified to convert uracil residues in the sample to thymidine, indexed if necessary and finally sequenced using next generation sequencing technology (Illumina recommended) (Agilent Technologies Inc., 2012). Bioinformatical analysis can then be carried out to differentiate methylated cytosines from un-methylated cytosines and to determine their implications. Such methodology opens up the possibility of cost effective epigenetic profiling in large genomes.

The RNA baits for the SureSelect Methyl-Seq Target Enrichment are based on primary DNA sequence. As such, when designing the wheat methylation array, design-space contigs for the wheat gene capture array were adapted for this purpose. This ensured that probe sequences were unique, non-repetitive, gene-rich and evenly distributed across the wheat genome. In this project it is demonstrated that an enrichment array can be used to give an overview of methylation patterns across the genic regions in the wheat genome and designed to target a 6Mb subset of the genic regions of wheat using the 5x Roche 454 genomic DNA wheat sequence generated by Brenchley, R. *et al.* (subset distributed across the contigs that were selected previously for the gene capture array design-space) (Brenchley *et al*. 2012).

Modification of gene expression by methylation can be tissue-specific or developmental stage dependent (Wang *et al.,* 2011a). It has been reported that methylation levels between members of the same species can differ, resulting in disease (Langevin and Kelsey, 2013) and can also differ in response to environmental factors or stresses e.g. temperature (Hashida *et al.,* 2006) or salt stress in plants (Wang *et al.,* 2011a). Further to this allele specific methylation has also been observed in animals and plants. Notably Wei *et al.* found allele specific methylation in humans that resulted in allele–specific expression (ASE) of death-associated protein kinase 1 (DAPK1) and predisposition to chronic lymphocytic leukemia (CLL) (Wei *et al.,* 2013).

The potential for differential methylation of homeologous genes in a polyploidy species is an important question in this study. Differential methylation was observed in maize correlating with differential expression of maternal and paternal alleles in the genes *r* and *dzr1* (Kermicle, 1978; Chaudhuri and Messing, 1994). In tetraploid cotton silencing or unequal homeologs expression was observed with epigenetic induction implicated; the proportion of genes with only partial homoeoalleles expressed was predicted to be as high as 25% (Adams *et al.,* 2003). For hexaploid wheat the percentage of genes with partial homoeoalleles expressed i.e. genome-wise differential gene expression, is thought to be 29%, typically one of the three homeoalleles present is silenced (Bottley *et al.,* 2006; Wang *et al.,* 2011b).

The array was to be used to test a number of hypotheses in wheat; firstly, Chinese Spring hexaploid wheat DNA could be enriched using the array to see if differential methylation exists between the A, B and D genomes. A list of naturally occurring homeologous SNP positions within the array bait sequences would allow identification of differential methylation between the A, B and D genomes in this analysis. Such SNPs would make it possible to associate sequencing reads with a homeologous SNP allele and ultimately a particular wheat genome. Secondly, using two growth temperatures for the Chinese Spring (12°C to represent a lower more ambient temperature for wheat growth in the UK and 27 °C to represent a contrasting high temperature for wheat growth) such DNA could be enriched using the array to test if temperature is capable of altering the methylation state and to see if this is both genome specific and genome independent. Finally, with use of RNAseq datasets for Chinese Spring at the same two growth temperatures (12°C and 27 °C) gene expression patterns could be identified and correlated with differential methylation under the hypothesis that it is this underlying methylation that can control both genome specific and temperature dependent changes in gene expression. To generate this gene expression data cDNA was generated, sequenced and analyzed by Mark Quinton-Tulloch by mapping the sequence data to the methylation array design-space. The program BitSeq was utilized to allow identification of gene expression levels.

BitSeq is an additional bioinformatical software tool with two main stages: transcript expression estimation and differential expression assessment. For the transcript expression estimation; sequencing reads are taken as input and aligned to the transcriptome using Bowtie to then allow calculation of the probability of a read originating from the transcript to be calculated and finally transcript expression level estimation. For differential expression assessment expression estimates are generated from replicates of 2 or more conditions; it infers the condition mean transcript expression and ranks transcripts based on the likelihood of differential expression (Glaus *et al.,* 2012).

1.9 Aims

Here the SOLiD^{TM} system and Illumina HiSeq and Genome Analyzer IIx are used primarily as re-sequencing tools to allow analysis of closely related species in relation to a reference genome with a view to SNP detection. The primary challenge is the efficient mapping of these short sequences to a reference genome and ultimately the further analysis of mapped reads to enable SNP detection. This information can be applied to analyses such as mutant identification by mapping-by-sequencing using where possible SHOREmap or else the principles that it employs. The complexity of the wheat genome makes it a great challenge to analyze in such a way. Initially a mapping, SNP calling and mutant identification analysis will be carried out on *Arabidopsis* to enable technique and skill development before beginning analyses on wheat. Further to this, and to manage the large, complex genome of wheat the gene and exome capture arrays will be validated and implemented in comparative and mutant identification analyses. The methylation enrichment array, which is developed here, will also be implemented to perform one of the first genome-wide studies of methylation patterns in hexaploid bread wheat.

The main aims of this project are therefore to develop high throughput pipelines for mapping, SNP calling and mutant identification in *Arabidopsis* and then adapting these optimized approaches for use with diploid and then polyploid wheat combined with use of target enrichment. These mapping, SNP calling and enrichment techniques will then be applied to enable a study of methylation patterns in a subset of wheat. These aims can be split up as follows:

-To perform a mapping and SNP identification analysis on the *Arabidopsis* clock mutant early bird (*ebi-1*) as an introduction to the techniques in the hope of gaining results largely similar to those gained by Ashelford, K *et al.*

-To perform mutant identification analyses on various *Arabidopsis* mutants to produce an interval of interest to be investigated further in the hope of determining the phenotype inducing SNP with use of SHOREmap

-To perform analyses and implement the use of the gene and exome capture NimbleGen arrays to allow validation plus a comparative inter-varietal hexaploid wheat SNP study

-To perform mutant identification to identify a mapping position and possibly a causative SNP in a simulated diploid, tetraploid and hexaploid wheat mutant (simulated complete chromosomes) to allow method development using SHOREmap or else the SHORE mapping-by-sequencing principles

-To identify a mapping interval and possibly a causative SNP in a diploid wheat mutant employing a combination of exome enrichment and SHOREmap or else the SHORE mapping principles with use of pseudo-chromosomes based on ordered enriched fragments by synteny with the closely related *Brachypodium*

-To ultimately produce a streamlined pipeline to enable mutant identification using sliding window mapping-by-sequencing analyses in enriched hexaploid mutants with the use of pseudo-chromosomes

-To design and validate a wheat methylation array based on the NimbleGen gene capture array to enable a comprehensive study of methylation patterns in a subset of wheat

Chapter 2. Mutant identification in the model plant *Arabidopsis thaliana*

Here accelerated mutant identification techniques are developed using mapping-by-sequencing analyses that combine whole genome sequencing with genetic mapping. Such methods have largely required a complete reference sequence and are typically implemented on a mapping population with a common mutant phenotype of interest. Here mutant identification was demonstrated on the model diploid plant *Arabidopsis thaliana* as a proof of principle of the methodology. It was also demonstrated on a simulated hexaploid mutant that was developed using the *Arabidopsis* reference genome.

2.1 Introduction

This chapter serves as an introduction to bioinformatics and to the techniques that will later be applied to the enriched genic regions of hexaploid bread wheat. It is sensible here to first apply mapping, SNP calling and mutation identification techniques to a more simple, well-annotated and smaller diploid genome, such as *Arabidopsis,* to develop a working pipeline before transferring their application to a more complex genome. This allows the understanding of the principles, limitations and time constraints of the various softwares to anticipate and solve problems that may be encountered when implementing them on a larger more complex genome. Here artificial sequencing datasets were also created using the *Arabidopsis* reference genome (Columbia strain, Col-0) allowing a simulation of mutant identification for firstly a diploid, then a tetraploid and finally a hexaploid mutant. Thus demonstrating the feasibility of the mapping-by-sequencing approach for polyploid genomes and allowing pipeline development before producing data sets.

Section 2.2 served as an initial familiarization study for the popular mainstream-mapping tool BWA and the SNP caller GATK. A mapping and SNP calling pipeline was developed (see figure 1.6) using these tools and SNP calls were generated for the same *Arabidopsis* datasets that were previously analyzed in a study by Ashelford *et al.* Study of a previously analyzed dataset provided a benchmark for comparison and validation of the pipeline under development here.

Section 2.3-2.4 demonstrated the use of the program's SHORE and SHOREmap for mapping and mutant identification in various *Arabidopsis* mutant plant datasets, all of which were grown, phenotyped, bulk segregated and typically backcrossed to the wild type parental line by Jonathan Napier's research group in Rothamsted. The samples were sequenced by the CGR at the University of Liverpool. In this, its original form, SHOREmap analyses heterozygote frequency, which is high due to the pooling of mutant plant DNA, and assumes that the homozygous character of the location harbouring the causal mutation will reduce the density of heterozygote markers to enable identification of the region (Schneeberger *et al.,* 2009). A firm grasp of such mapping-by-sequencing methodology that is used within SHOREmap will be needed if it is to be ultimately adapted for use in the hexaploid wheat. SHOREmap itself is currently intended only for use in a diploid genome.

In section 2.4.2 these techniques were combined to demonstrate the initial trial of a bespoke novel mutant identification pipeline. This pipeline was developed here using a combination of BWA mapping and VarScan SNP calling methodologies plus implementation of the ideas demonstrated by SHOREmap for allele frequency analysis and identification of an interval containing the phenotype inducing SNP. Due to the complexity of this task the method was first developed using the diploid *Arabidopsis* Ws-2 mutant (section 2.4.1), in effect re-creating the result from this section that was gained using SHOREmap, using these pipelines and methodologies to enable comparison and subsequent validation. Development of this novel mutant identification pipeline will ensure a clear and thorough understanding of the principles used for the analysis and the mechanism of action at each stage of the pipeline that will allow easier manipulation for its future use on a polyploid genome.

The novel mutant identification pipeline that is introduced in section 2.4.2 was further developed and utilized in sections 2.5-2.7. Here, it is demonstrated that it is in fact possible to use such a mapping-by-sequencing based method to identify a region harbouring a causal mutation in a diploid, tetraploid and finally a hexaploid with use of artificial sequencing datasets. These datasets were all created using the *Arabidopsis* Col-0 reference genome as a starting point. This analysis allows the focus for this initial methodology development to be on mutant detection in a polyploid and to eliminate, for now, the added complexity of fragmented, incomplete chromosome sections that will later be encountered in the study of the enriched and poorly defined wheat genome. In the hexaploid mutants that are studied here the phenotype inducing SNP is likely to only be homozygous in 1 of the 3 genomes, and therefore present in only ~1/3 of sequencing reads, thus potentially more difficult to pick out than a homozygote SNP in a diploid organism where it will be seen in ~100% of the sequencing reads.

The artificial dataset that is utilized and developed in sections 2.5-2.7 was also designed to represent data similar to that which we would expect from a mutant scored F2 bulk segregant, backcrossed line with a single homozygous phenotype inducing SNP. This design ensured that the novel mutant identification pipeline that is based on SHOREmap's mapping-by-sequencing principles could be more easily employed. In the hexaploid mutant (section 2.7) the genomes were also designed to be approximately comparable to the 3 wheat genomes i.e. three diploid genomes with similar numbers of SNPs conserved and different between them and no major structural differences as co-linearity appears to be retained between them (Gu *et al.,* 2004).

2.2 *Ebi-1* SNP identification

In the study by Ashelford *et al.* the 120 Mb genome of a novel EMS induced *Arabidopsis* clock mutant *early bird* (*ebi-1*) and the corresponding wild type, Wassilewskija (Ws-2) were re-sequenced using the Applied Biosystems SOLiD sequencing by ligation technology. Screening an EMS-mutagenized population for mutants with CAB2 oscillating with a short period allowed identification of the *ebi-1* mutant. This mutant was then backcrossed four times with the original parent line Ws-2 in the hope of removing EMS-induced SNPs that were not associated with the phenotype. The data was mapped and SNP calling was carried out using the SOLiD System Analysis Pipeline Tool (Corona Lite) and the *Arabidopsis* Col-0 TAIR9 reference genome. SNPs that were novel to *ebi-1* were selected that could be potential candidates for causing it's phenotype. Investigating gene expression data for mutated genes further narrowed down the list of SNPs that were identified and finally a SNP in the gene *AtNFXL-2* was determined as the likely cause of the *ebi-1* phenotype (Ashelford *et al*, 2010).

With the current development of mapping tools such as BWA and Bowtie enabling efficient sequence data mapping independent of sequencing technology, the corona lite tool, that was developed specifically for mapping colour-space reads and subsequent SNP calling, is now largely redundant. In this study, re-mapping and SNP calling in the sequencing datasets that were generated by Ashelford *et al.,* enabled a comparison of the SNP calls that were generated using a bespoke mapping and SNP calling pipeline to ensure that the pipeline picked up the phenotype inducing SNP plus the majority of the other confirmed SNPs that were identified in this study. This served as a validation for the SNP calling pipeline that has been developed here. The data mapping and SNP calling for the SOLiD sequenced *Arabidopsis* clock mutant *ebi-1* and the corresponding wild type Ws-2 datasets generated by Ashelford *et al.* was therefore repeated using BWA (v 0.5.8) and GATK (version 1.0.4418).

The short SOLiD sequencing reads that were generated for Ws-2 were initially mapped to Col-0, to the TAIR9 reference genome, using the alignment tool BWA. The same was done for *ebi-1* dataset. The steps involved in the BWA alignment and subsequent GATK SNP calling are shown in figure 1.6. Indexing of the reference sequence, involved use of the 'IS' algorithm that is preferred if the genome is smaller than 2GB. The path for BWA fragment short read mapping was followed (prior to the 2013 update to BWA-backtrack) and parameters to allow processing of colour-space SOLiD reads were used. The program csfasta2fastq was used to convert the SOLiD sequencing csfasta/qual files into a fastq file, which is required as input for BWA's alignment step. For sequencing read alignment the mismatch number was altered and the alignment re-run allowing 2, 3, 4 and 6 mismatches per sequencing read. All example BWA/GATK commands can be found in the command outline appendix section 1, 2 and 3 showing default parameters used unless otherwise stated. When 4 mismatches were used on average approximately 20x coverage was gained and over 85% of the genome was mapped to in both datasets.

Using the GATK SNP detection was carried out between Ws-2 and Col-0 and *ebi-1* and Col-0. This allowed subsequent identification of those homozygote SNPs that differed between *ebi-1* and Ws-2 relative to Col-0 i.e. SNPs novel to *ebi-1* could be potential candidates for causing it's phenotype. Four such lists were created corresponding to the 4 different mapping analyses that were carried out for each sample (allowing 2, 3, 4 and 6 mismatches per sequencing read). SNPs that could be found in all 4 lists made up the final list.

The following main filters were applied to the SNPs during variant filtration that were found to enable distinction between true SNPs and false positives in this dataset: Discard SNPs within the 10bp flanking region around a potential indel; discard SNPs covered by 3 or fewer reads (per sample); discard SNPs with a QUAL score below 50 (phred scaled probability that call is correct); in any 10bp window if there are 3 or more SNPs discard them all; discard SNP if depth of coverage is greater than 100 and finally discard heterozygote calls. These filters were determined to be most appropriate for this particular dataset and where possible were tailored to be as close as possible to those used in the study by Ashelford *et al.*

The *ebi-1* mutant was compared with Ws-2 rather than simply Col-0 as it shares greater similarity with Ws-2. With this methodology SNPs could be eliminated as unlikely to cause the mutant phenotype if both Ws-2 and *ebi-1* had them when mapped to Col-0. The results gained supported those identified by Ashelford *et al.* and are summarized in figure 2.1.

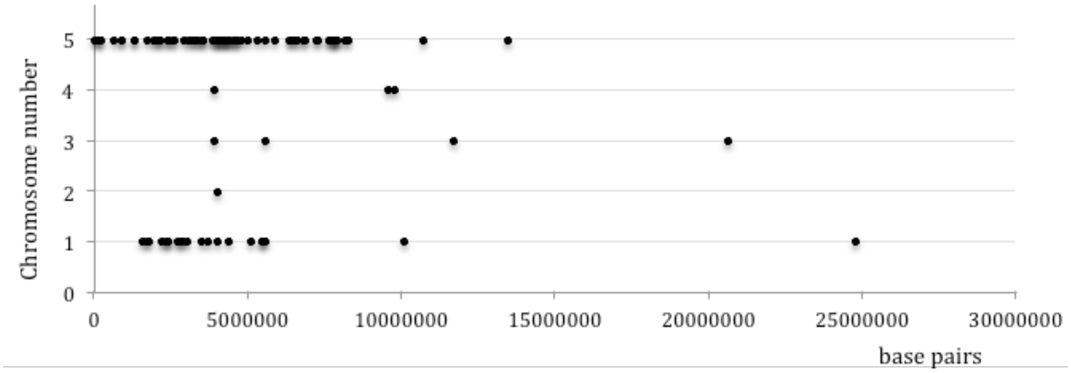


Figure 2.1. Positions of *ebi-1* SNPs along the *Arabidopsis* chromosomes 1-5. Homozygous SNPs that were unique to the *Arabidopsis ebi-1* mutant were identified using GATK and are detailed here

The original study identified 109 SNPs that were unique to *ebi-1*, plus the *ebi-1* phenotype inducing SNP that was found in the gene *AtNFXL-2* (At5g05660). The *AtNFXL-2* protein shares homology with the mammalian zinc finger transcription factor. In this study 95 SNPs were identified that were unique to *ebi-1* (for full list see Appendix 1, table 1) plus the *ebi-1* phenotype inducing SNP. Two trends were also found in the original study that were reinforced with this study; a high proportion of SNPs that were found in the north arm of chromosome 5 where the phenotype inducing SNP was found. This is a result of SNPs being carried through with the *ebi-1* mutation during backcrossing. There was also, to a lesser extent, a group of SNPs on chromosome 1 due to backcrossing *ebi-1* with the original parent.

2.3 Mutation identification in the *Arabidopsis* strain Ws-2

In this study an EMS Ws-2 fatty acid metabolism mutant, which was provided by Jonathan Napier's research group in Rothamsted, was sequenced by the CGR using the Illumina Genome Analyzer IIx (standard fragment library preparation). A pool of over 100 F_2 backcrossed lines were sequenced that all displayed the mutant phenotype. Analysis of the allelic frequency of EMS SNPs that co-segregate with the mutant SNP could potentially locate a mapping interval to allow identification of the mutation causing SNP. Bulk segregation of the mutant makes it an ideal candidate to be analyzed using the SHORE and SHOREmap software (version 0.5.0 and version 1.1). The steps that were followed using SHORE and SHOREmap to allow the identification of possible causative SNPs responsible for the mutant phenotype of the organism are summarized in figure 2.2. The actual commands that were used are detailed in the command outline appendix section 5. All settings that have been used are default unless otherwise stated.

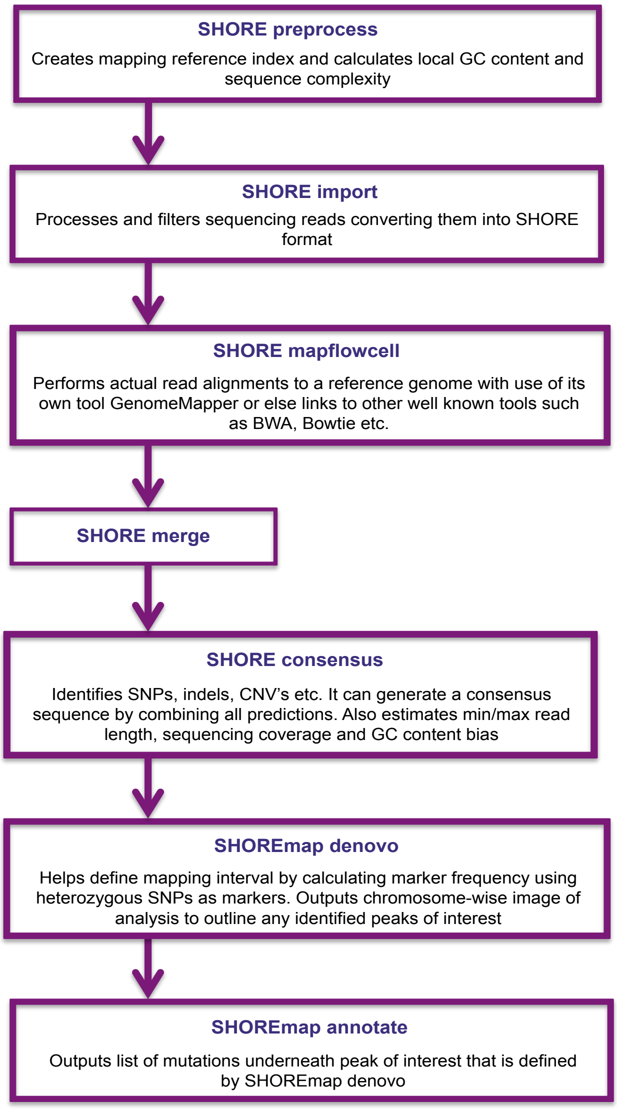

SHORE preprocess
Creates mapping reference index and calculates local GC content and sequence complexity
SHORE import
Processes and filters sequencing reads converting them into SHORE format
SHORE mapflowcell
Performs actual read alignments to a reference genome with use of its own tool GenomeMapper or else links to other well known tools such as BWA, Bowtie etc.
SHORE merge
SHORE consensus
Identifies SNPs, indels, CNV's etc. It can generate a consensus sequence by combining all predictions. Also estimates min/max read length, sequencing coverage and GC content bias
SHOREmap denovo
Helps define mapping interval by calculating marker frequency using heterozygous SNPs as markers. Outputs chromosome-wise image of analysis to outline any identified peaks of interest
SHOREmap annotate
Outputs list of mutations underneath peak of interest that is defined by SHOREmap denovo

Figure 2.2. Pipeline for mutant identification using SHOREmapping. 'SHORE preprocess' takes the reference sequence as input and produces an Index folder as output. Then 'SHORE import' can take as input csfasta/qual files or fastq files and outputs SHORE file format reads to the Flowcell Folder. 'SHORE mapflowcell' performs read alignments using the Index and Flowcell folder producing a map.list file that is written to the Flowcell folder. 'SHORE merge' can merge together alignments for parallel analysis downstream as one map.list file. 'SHORE consensus' takes the map.list alignment file and the reference sequence from the Index folder and outputs its analysis (SNPs, indels, CNV's etc.) to an Analysis folder. 'SHOREmap denovo' takes input from the Analysis folder (heterozygous SNPs) and the Index folder reference genome and generates a plot for each chromosome of the relative allele frequencies per user-defined window (200,000bp default) in order to define the phenotype inducing SNP's mapping interval. Finally 'SHOREmap annotate' extracts a list of homozygous SNPs that lie within the interval identified by 'denovo'.

In this instance ‘SHORE preprocess’ was implemented with use of Col-0, the TAIR10 reference genome, as the reference input. The colour-space (–c) option was not implemented, as colour-space indexing was not required due to having Illumina not SOLiD reads. ‘SHORE import’ was then used to convert the sequencing reads into SHORE read file format. ‘SHORE mapflowcell’ was used to perform read alignments with the –v option as ‘Fastq’ due to Illumina output in fastq files as opposed to ‘Solid’ used when SOLiD generated sequence data is inputted. At this step, again no –c is required due to Illumina sequencing reads and an average depth of coverage of 18.5 was calculated. ‘SHORE merge’, ‘SHORE consensus’ and ‘SHOREmap denovo’ were then each run in turn. In this case the program ‘SHOREmap denovo’ was re-run using 100,000, 200,000 and 300,000bp windows to allow downstream determination of the most suitable parameters for this dataset and 200,000bp was confirmed as such (gave the smoothest baseline and most convincing peaks of interest). The ‘denovo’ output plot is created in pdf format and for this dataset is shown in figure 2.3.

It is clear from figure 2.3 that one peak has been clearly identified at the beginning of chromosome 3. The peak of interest was noted in the region 1,000,000-3,000,000 bp on chromosome 3 and was within the preliminary mapping interval that has been identified for this mutation. This region was used as input for ‘SHOREmap annotate’ with the processed ‘denovo’ output along with the reference sequence (Index folder) and a list of homozygote SNPs that were identified by ‘SHORE consensus’ and outputted to a file called homozygous_snp.txt in the Analysis folder. A text file list of homozygous mutations, prioritized by their distance to the highest peak in the user-defined interval, was generated as output. The top SNP in the list generated by ‘annotate’ is likely to be the causative mutation (Schneeberger *et al.*, 2009). The results gained are shown in table 2.1 and outline the top SNPs identified and their positions. Subsequent analysis of these SNPs through the The *Arabidopsis* Information Resource (TAIR, 2011) allowed further information to be gained regarding associated genes and annotations.

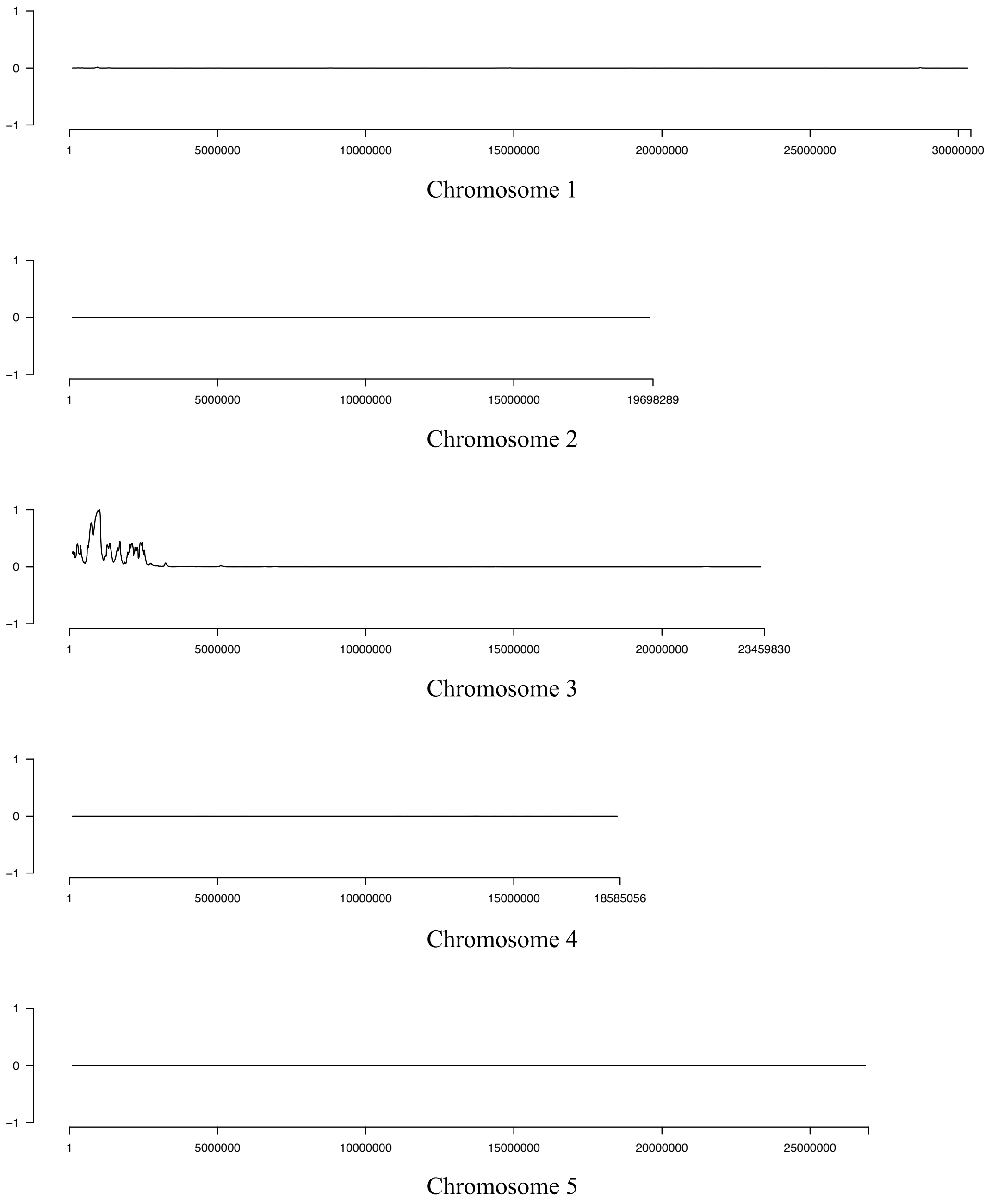


Figure 2.3. SHOREmap denovo output pdf file for an *Arabidopsis* Ws-2 mutant. Mapping analyses carried out using the *Arabidopsis* Col-0 TAIR10 genome as a reference sequence. A window size of 200,000bp was used.

SNP Position	Base Change	Associated Gene	Function	Location	% Reads Supporting SNP	Amino Acid Change
1052392	C→T	AT3G04050	Pyruvate kinase family protein	~900bp downstream	100	n/a
1082660	C→ T	AT3G04120	Encodes cytosolic GADPH	Exonic	97	Proline → Leucine
1163605	C→T	AT3G04380	Encodes nucleolar histone methyl-transferase	Intronic	100	n/a
856061	C→T	AT3G03560	Unknown	Intronic	100	n/a
1287110	C→ T	AT3G04721	Unknown	Promoter	100	n/a

Table 2.1. Top candidate homozygous SNPs for the phenotype inducing SNP of an *Arabidopsis* Ws-2 mutant. SNPs taken from output of 'SHOREmap annotate' and additional annotation provided with use of the *Arabidopsis* TAIR website.

Professor Johnathan Napier and his group at Rothamsted supplied the original bulk segregated mutant DNA that was sequenced at the University of Liverpool. Prof. Napier's group are now taking the SNPs that are detailed in table 2.1, along with a short list of SNPs in the immediate vicinity, forward for further investigation in the laboratory. The table 2.1 list was extended to ensure the phenotype inducing SNP was not missed since the peak in chromosome 3, although very convincing, was spread over a large interval. The extended list encompassed 32 SNPs ranked in order of confidence, that were spread across the entire peak interval rather than simply the peak tip, and is detailed in full in Appendix 1, table 2.

2.4 Further mutation identification in the *Arabidopsis* strain *Ws-2*

2.4.1 SHORE mapping pipeline

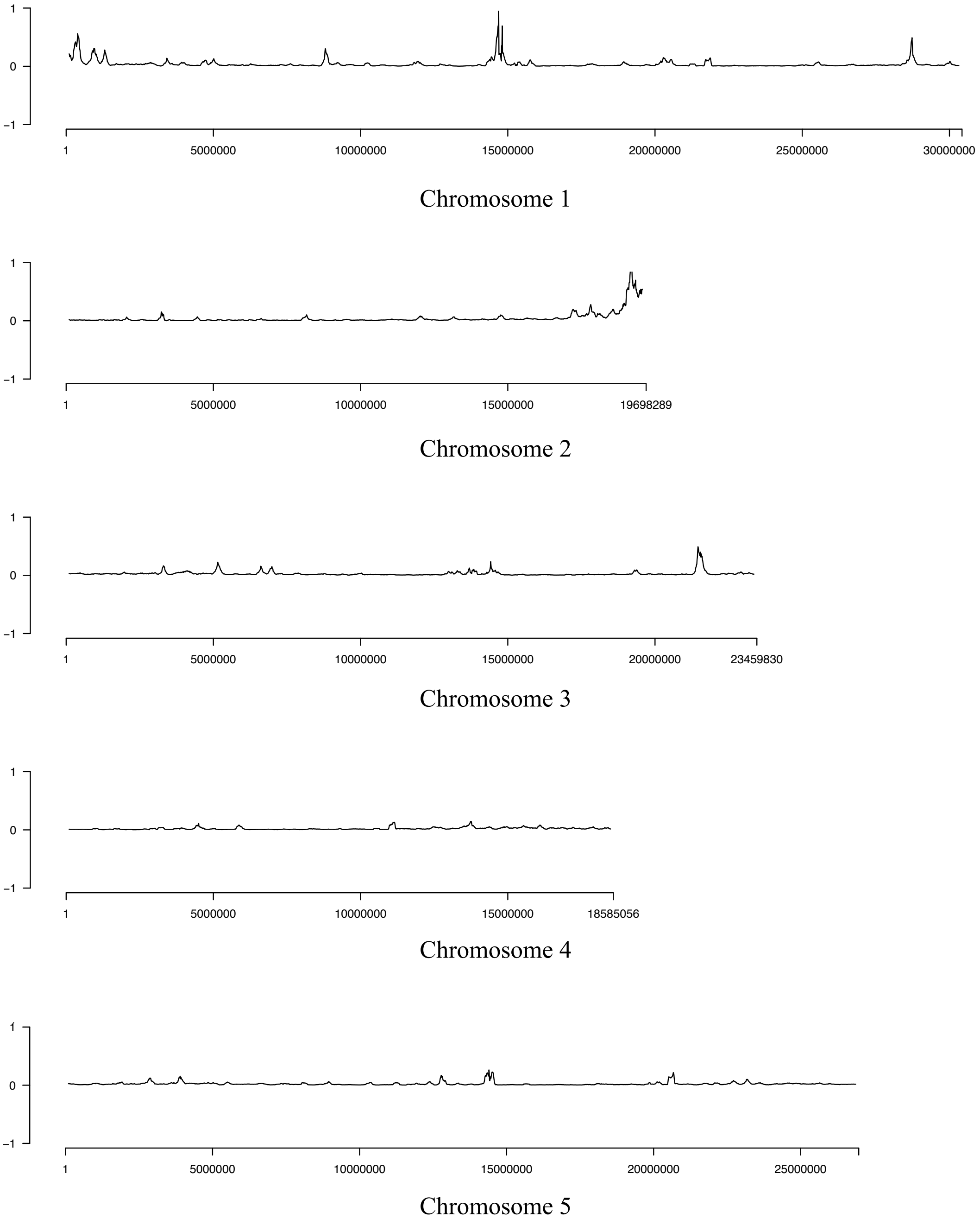
In a parallel study to that detailed in section 2.3 an EMS *Arabidopsis* Ws-2 fatty acid metabolism mutant *(*pool of F_2 backcrossed lines all displaying the mutant phenotype) that was provided by Jonathan Napier's research group in Rothamsted, was sequenced at the CGR using the SOLiD sequencing by ligation technology (fragment library used). The aim of this investigation was to identify a potential SNP that was novel to the mutant that could be inducing the mutant phenotype.

The SOLiD sequencing reads were initially taken through the pre-assembly error correction SAET-SOLiD Accuracy Enhancement Tool. This software first builds a list of all k-mers present in the reads and those with a frequency higher than a default threshold are considered trusted. Each read is corrected, where possible, so that it contains only trusted k-mers. This tool corrects missing and miss-calls in the reads thus increasing their mappability by up to 3 times. It helps to reduce false SNP calls due to colour-space to allow us to distinguish between true SNPs and sequencing errors (Applied Biosystems, Life Technologies, SAET, 2011). In an assay that was carried out on a sub-set of *Arabidopsis* SOLiD sequencing reads the use of SAET before mapping resulted in a 20% increase in the number of reads that were subsequently mapped. The tool requires as input the csfasta and qual file output standard to SOLiD sequencing. The approximate size of the genome that has been sequenced is also required.

The steps that were followed using SHORE/SHOREmap to allow the identification of possible causative SNPs responsible for the mutant phenotype of the organism follow the same method used in section 2.3 with use of the Col-0 TAIR10 reference genome and parameters to allow use of SOLiD colour-space reads. In the penultimate step of the pipeline 'SHOREmap denovo' was re-run using 100,000, 200,000 and 300,000bp windows to allow downstream determination of the most suitable parameters for this dataset. 200,000bp was again confirmed as the correct choice. The 'baseline' in the plots deviated from the smooth baseline that was seen in the exemplary dataset shown in figure 2.3. This was potentially due to the fact that in order to create that mutant bulk segregant dataset 100 plants were pooled, here only ~50 plants were pooled.

An initial run through of this method gained inconclusive results in the 'denovo' output and as such the mapping success of the dataset was analyzed. Around 93% of the reference genome was mapped to with a mean depth of ~17 and a median depth of 15. The mean and median depths of mapping seem to be comparable to the successful analysis in section 2.3 and therefore adequate. On closer inspection the SNPs that had been identified and were being used to generate the 'denovo' output had, on average, a depth of 7.05. This depth was much lower than the overall average. In a repeat analysis only SNPs that had a depth of coverage greater than or equal to 10 were considered. Subsequent filtering of SNPs allowed

SHOREmap ‘denovo’ to be re-run with greater success and the graphical output that was generated is shown in figure 2.4.



Chromosome 1

Chromosome 2

Chromosome 3

Chromosome 4

Chromosome 5

Figure 2.4. SHOREmap denovo output pdf file for an *Arabidopsis* Ws-2 mutant. Mapping analyses carried out using the Col-0 TAIR10 *Arabidopsis* genome as a reference sequence. A window size of 200,000bp was used for this analysis.

Two peaks were clearly identified from figure 2.4. These peaks were noted in the region 14,000,000-15,000,000bp on chromosome 1 and in the region 18,000,000-19,698,289bp on chromosome 2. These two regions were used as input for 'SHOREmap annotate' with the processed 'denovo' output. The results gained are shown in table 2.2 and outline the top homozygous SNPs that were identified and their positions. Subsequent analysis of these SNPs through the The *Arabidopsis* Information Resource (TAIR, 2011) allowed further information to be gained regarding associated genes that is also detailed in the table.

It should be noted that the two top candidate SNPs that were found by 'SHOREmap annotate' in chromosome 1 (table 2.2) are transposable element genes and as such will be disregarded as these SNPs are likely to be located within the centromeric region of the chromosome. SNPs found in chromosome 2 were not found to have the same issue and as such can be taken forward for further investigation in the laboratory by Prof. Johnathan Napier and his group who supplied the original bulk segregated mutant DNA that was sequenced here at the University of Liverpool. The SNP locations that were identified in chromosome 2 were within the preliminary mapping interval that Prof. Napier's group identified for this mutation.

SNPs with a depth below 10 were removed from this analysis as low coverage, potentially false positive SNPs, appeared to be hindering effective peak discrimination within the genome-wide analysis. However, these SNPs were reinstated onto the final SNP list given to Prof. Napier if they fell within the proximity of the identified peak. This was considered to be sensible in case a real SNP was wrongly classified as a false positive due to low coverage. The list totalled 36 SNPs and is detailed in full in Appendix 1, table 3.

Chrom	SNP Position	Base Change	Associated Gene	Function	Location	Amino Acid Change
1	14608287	T→A	AT1G39350	Transposable element gene	n/a	n/a
1	14608141	A→ G	AT1G39350	Transposable element gene	n/a	n/a
2	19186718	G→A	AT2G46700	CDPK-related kinase 3	Promoter	n/a
2	19256651	G→A	PPA3	Inorganic pyrophosphatase activity	1000bp down-stream of gene	n/a
2	19328802	G→A	AT2G47040	Enhance growth of pollen tube in style and transmitting tract tissues	Coding regon	Alanine→ Valine
2	19336514	G→A	AT2G47070	DNA binding proteins/putative transcription factors	Promoter	n/a
2	19357996	G→A	AT2G47160	Boron transporter	Coding region	Serine → Phenylalanine
2	19388595	G→A	AT2G47230	Plant specific DUF724 protein family	Coding region	Aspartic acid → Asparagine
2	19443789	G→A	AT2G47390	Serine-type peptidase activity	Coding region	Leucine → Phenylalanine
2	19540806	G→A	AT2G47650	Similar to UDP-glucuronic acid decarboxylase	Coding region	Glutamine → Stop codon
2	19548657	G→A	AT2G47680	Zinc finger helicase family protein	Coding region	Proline → Serine
2	19566463	G→A	AT2G47760	Asparagine-linked glycosylation	Coding region	Leucine → Phenylalanine
2	19577036	G→A	AT2G47800	Plasma membrane localized ATPase transporter	Coding region	Glycine → Aspartic acid
2	19672565	G→A	AT2G48100	Exonuclease family protein	Intron	n/a
2	19678764	G→A	AT2G48110	Unknown protein	Coding region	Glycine → Glutamic acid
2	19691150	G→A	AT2G48160	Tudor/PWWP/MBT domain-containing protein	Coding region	Threonine → Methionine

Table 2.2. Top candidate homozygous SNPs for the phenotype inducing SNP of an *Arabidopsis* Ws-2 mutant. Homozygous SNPs in over 80% of the sequencing reads and taken from output of 'SHOREmap annotate' and additional annotation provided with use of the *Arabidopsis* TAIR website.

2.4.2 Development of a novel pipeline for mutant identification with use of BWA and SAMtools

SHOREmap is currently intended only for use in a diploid genome. There is a need to develop a mutant identification pipeline that can be applied to polyploid species. Here the initial trial of a bespoke novel mutant identification pipeline is implemented using the diploid *Arabidopsis* Ws-2 mutant (section 2.4.1), in effect re-creating the result from this section that was gained using SHOREmap, to enable pipeline validation prior to implementation on polyploid species.

This bespoke mutant identification pipeline was initially trialed on the *Arabidopsis* mutant dataset that was introduced in section 2.4.1 and involved mapping the data, SNP calling and finally analysis of allele frequency to define an interval containing the phenotype inducing SNP. The tools selected for use in the pipeline include; mapping tool BWA (v 0.5.8), SAMtools (SAMtools-0.1.16) and VarScan (VarScan.v2.2.3.jar). Here VarScan was implemented in preference to GATK for the following main reasons; firstly, with more simple statistics VarScan proves to be more effective in pooled samples easily identifying low frequency alternate alleles (Koboldt *et al.,* 2009); the tools are fairly similar otherwise and a repeat of the investigation in section 2.2 using VarScan yielded a SNP list that was over 70% identical to that gained with GATK and more importantly included similar SNP trends and the mutation of interest; VarScan can output multiple alternate alleles in one position in a simple user-friendly output that allows easier adaptation downstream to a polyploid genome; its input SAMtools pileup file also allows easy user manipulation and calculation of alternate allele frequency; finally, at this time the GATK had no such polyploid adaptation.

The *Arabidopsis* Ws-2 mutant bulk segregant sequencing data was mapped to the Col-0 TAIR10 reference genome using the BWA short read-mapping pipeline, SAMtools was implemented for filtering, duplicate reads were removed and finally the VarScan SNP calling pipeline was followed (see figure 1.6.). All commands can be found in the command outline appendix section 1 and 4. During indexing of the reference parameters to allow a colour-space index to be built and the 'IS' algorithm to be used were selected. During alignment of reads to the reference 4 mismatches per read alignment were allowed and the query was reversed but not complemented (required for a colour-space alignment). No other parameters were altered to gain a comparable mapping output (SAM/BAM file) to that of section 2.4.1. Mapping coverage was analyzed and details are shown in table 2.3.

Chromosome	Size (bp)	Bp mapped	% Mapped	Mean coverage	Median coverage	SD of coverage
1	30,427,671	28,066,281	92.24	15.95	15.00	20.34
2	19,698,289	18,630,523	94.58	19.26	15.00	75.87
3	23,459,830	22,153,984	94.43	16.54	15.00	57.87
4	18,585,056	17,420,143	93.73	15.73	15.00	33.15
5	26,975,502	25,263,754	93.65	16.10	15.00	32.73

Table 2.3. Descriptive mapping statistics for an *Arabidopsis* Ws-2 mutant. Statistics given per chromosome for the mapping of an *Arabidopsis* mutant to the TAIR10 reference genome using BWA.

SNP calling was performed from the filtered BAM file using SAMtools. Initially SAMtools mpileup was used to output a summary of the bases seen across all sequencing reads mapping to each genomic position. This command took as input the reference fasta file and the BAM file. The output pileup format file was then run through VarScan to filter out SNPs and to calculate alternate allele frequencies at each location. The parameter 'min-coverage' was used and defined as 10 i.e. the minimum depth of coverage for a SNP to be called was set to 10. This produces a text file including columns detailing the following: SNP position, reference base, consensus base, number of reads supporting the reference allele, number of reads supporting the alternate allele, % of reads with the alternate allele, average quality of mapped reads supporting the reference allele, average quality of mapped reads supporting the alternate allele and finally the variant allele itself.

A homozygote SNP that could potentially be inducing the phenotype in this mutant was defined as a SNP with more than 80% of the sequencing reads containing the alternate allele. These SNPs were filtered from the VarScan output into a homozygote's file. A heterozygote SNP was defined as having greater than or equal to 20% and less than or equal to 80% of sequencing reads containing the alternate allele. These SNPs were also filtered from the VarScan output into a heterozygote's file.

Both of these files were then run through a preliminary Perl script (Allele-frequency-interval-determination.pl) that was developed to take VarScan formatted files as an input and calculate how many SNPs in the input file fall within each 200,000 base pair window along each chromosome. The output from this Perl script takes the form of a simple text file with the interval number in column 1 and the number of SNPs in the input file that fall within it in column 2. From this data the homozygote to heterozygote ratio per window could be calculated and used to produce the frequency plot shown in figure 2.5. A higher ratio would suggest a homozygote region with fewer heterozygotes present (as we would expect in a region that harboured the causal mutation). Centromeric regions were masked to avoid false positive peaks.

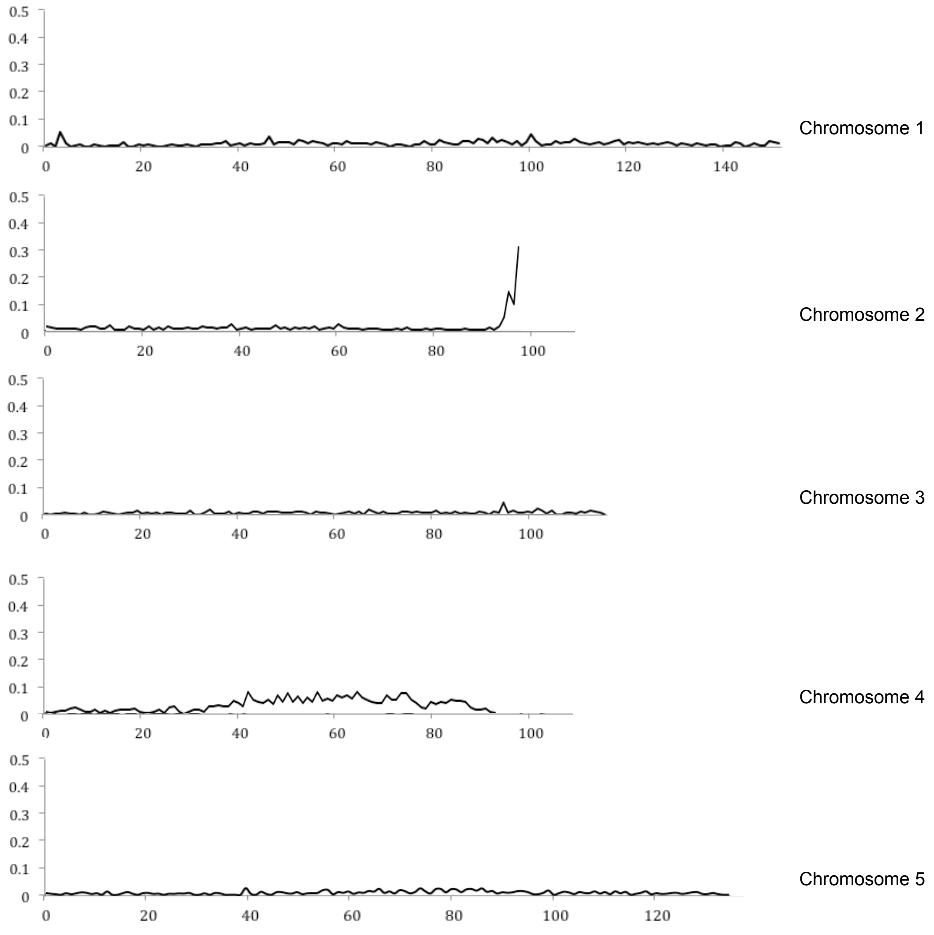


Figure 2.5. Allele frequency analysis of an *Arabidopsis* Ws-2 mutant. Plots describe the homozygote to heterozygote ratio (x axis) per 200,000bp interval along each chromosome (y axis).

In figure 2.5 a clear solitary peak can be seen at the end of chromosome 2 within the interval 95-97. This interval corresponds approximately to chromosome location 19,200,000-19,698,289 bp and covers the same region that was discovered by SHOREmap. The absolute peak (highest point) is observed in the interval 97 i.e. location 19,600,000-19,698,289bp and in theory the SNPs of interest could be narrowed down further to only those in this narrow region. Experimental validation is required before this conclusion could be confirmed. The SNPs that were identified in this study that fall within the peak intervals 95-97 on chromosome 2 are detailed in table 2.4 and a comparison of the chromosome 2 SNPs in table 2.4 and table 2.3 shows that both lists are identical apart from 2 SNPs; one at position 19,347,162 bp in chromosome 2 appears in the SAMtools data in table 2.4 and not the SHOREmap data in table 2.3 due to it having insufficient depth of coverage when analyzed by SHOREmap (depth of 8); the other SNP at position 19,540,806 in chromosome 2 appears in the SHOREmap data and not the SAMtools data seemingly due to a lower mapping quality in the SAMtools data leading to the SNP not passing SNP calling filters marginally within SAMtools.

SNP Position	Base Change	Associated Gene	Function	Location	Amino Acid Change
19186718	G→ A	AT2G46700	CDPK-related kinase 3	Promoter	n/a
19256651	G→ A	PPA3	Pyrophosphatase activity	1000bp downstream of gene	n/a
19328802	G→A	AT2G47040	Enhance growth of pollen tube in style and transmitting tract tissues	Coding regon	Alanine→ Valine
19336514	G→A	AT2G47070	DNA binding proteins/putative transcription factors	Promoter	n/a
19347162	G → A	AT2G47115	Unknown protein	Promoter	n/a
19357996	G→A	AT2G47160	Boron transporter	Coding region	Serine → Phenylalanine
19388595	G→A	AT2G47230	Plant specific DUF724 protein family	Coding region	Aspartic acid → Asparagine
19443789	G→A	AT2G47390	Serine-type peptidase activity	Coding region	Leucine → Phenylalanine
19548657	G→A	AT2G47680	Zinc finger helicase family protein	Coding region	Proline → Serine
19566463	G→A	AT2G47760	Asparagine-linked glycosylation	Coding region	Leucine → Phenylalanine
19577036	G→A	AT2G47800	Plasma membrane localized ATPase transporter	Coding region	Glycine → Aspartic acid
19672565	G→A	AT2G48100	Exonuclease family protein	Intron	n/a
19678764	G→A	AT2G48110	Unknown protein	Coding region	Glycine → Glutamic acid
19691150	G→A	AT2G48160	Tudor/PWWP/MBT domain-containing protein	Coding region	Threonine → Methionine

Table 2.4. Top candidate homozygous SNPs for the phenotype inducing SNP of an *Arabidopsis* Ws-2 mutant. Candidate homozygous SNPs taken from output of BWA mapping, SAMtools SNP calling and allele frequency analysis and additional annotation provided with use of the *Arabidopsis* TAIR website. All positions that are detailed can be found on chromosome 2.

2.5 Mutant identification in a diploid using an artificial dataset

Here the bespoke mutant identification pipeline that was introduced in section 2.4.2 was trialed on an artificial sequencing dataset that was created for the diploid plant *Arabidopsis.* This analysis was carried out as a precursor to the development of the artificial sequencing dataset into a tetraploid and finally a hexaploid based dataset on which the pipeline would later be trialed.

2.5.1 Development of an artificial diploid sequence dataset

The development of sequencing data for the artificial diploid mutant was based on the evolution of the A genome of wheat. The program evolver was used to create an evolved *Arabidopsis* genome (known as *Arabidopsis* A genome within this study) that was as different from the *Arabidopsis* reference strain Col-0 than the wheat A genome is from its donor *T. turgidum*. To output an approximation of an evolved sequence that has accumulated SNPs at random over a given time period, evolver took the *Arabidopsis* Col-0 reference sequence as input, along with a 'branch length' of 0.0093. This branch length, or expected amount of sequence change per site, was the value for genome A in wheat calculated by Gu *et al.* and creates a random substitution approximately every 108 bases. No major structural changes were added into this approximation of an evolved genome for simplicity within the analysis.

For this evolved genome to be run through a SNP identification pipeline, sequencing data for the *Arabidopsis* A genome was required. The SAMtools program wgsim, the short read simulator, creates artificial paired end Illumina sequencing data for an input fasta sequence with a base error rate of 0.02, a standard distance between the two ends of 500, a standard length of read of 70 and an option to determine the number of read pairs created that in this case was set at 100,000,000.

The sequencing data required a characteristic homozygous hotspot (containing the phenotype inducing SNP) fading into the low homozygote/high heterozygote frequency found elsewhere in the genome and characteristic of a bulk segregant sample of mutant F2 backcrossed lines. To create this effect the fasta file described in figure 2.6 was created and used as input into wgsim. Duplicating the genome sequence within this input file would result in approximately 50% of the sequencing reads being created corresponding to each genome sequence and tripling the genome would result in 33% of the reads corresponding to each genome sequence etc. Therefore if two genome sequences were included as wgsim input and the second differed from the first, as *Arabidopsis* Col-0 differs from the evolved *Arabidopsis* A genome i.e. every ~108bps, then at every differing position a heterozygote with an approximately 50:50 distribution of reads will be created.

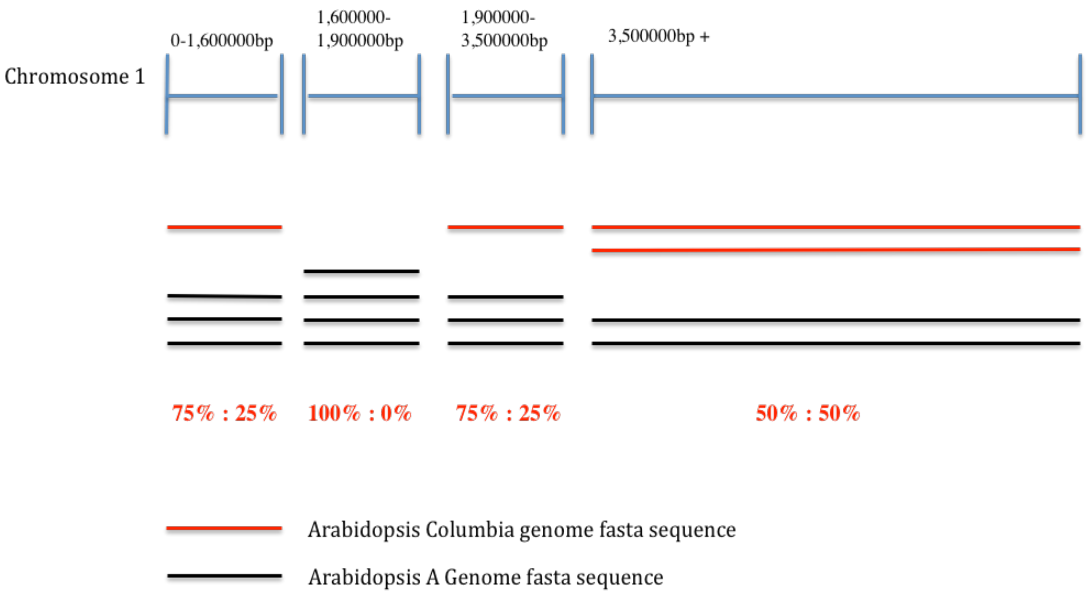


Figure 2.6. Outline of sequence input into SAMtools wgsim. Outline of the fasta sequence input (red and black lines) into SAMtools wgsim in order to create a diploid mutant with a homozygous hotspot. Chromosome 1 is shown, all other chromosomes are of the same design as for the 3,500,000bp+ region of chromosome 1.

The mutation and the homozygous hotspot were placed on chromosome 1 in the region 1,600,000-1,900,000bp. The reference genome sequence that the data would eventually be mapped to was the TAIR10 *Arabidopsis* Col-0 genome as such the mutation was added to the *Arabidopsis* A Genome sequence. Figure 2.6 demonstrates how for each section of the *Arabidopsis* genome four DNA sequences were inputted into wgsim. For chromosomes 2-5 and chromosome 1 region 3,500,000bp+ this was broken down into two sequences for the *Arabidopsis* Col-0 genome and two for the evolved *Arabidopsis* A genome i.e. approximately 50% reads generated for each in these regions. This ensured that on average every 108bp where there was a difference between the two genomes a heterozygote could be seen in the sequencing reads (higher heterozygote frequency). The only homozygotes found in these regions were due to errors in sequencing reads generated by wgsim to create a dataset with more variability that is typical of a real dataset, this is also why heterozygotes were not always in exactly 50% of reads.

In figure 2.6 the regions of chromosome 1 approaching the mutation; 0-1,600,000bp and 1,900,000-3,500,000 show a demonstration of the phasing into the homozygote region that we would expect to see in real data (although we would expect more interim stages between 50:50 and 100:0 than just 75:25 this would have proved too difficult to simulate). These sections both required three input sequences for the *Arabidopsis* A genome and only one for the *Arabidopsis* Col-0 genome to allow more potential homozygote alternate allele calls against the *Arabidopsis* Col-0 reference as the proportion of reads with an allele from the *Arabidopsis* A Genome, at positions with a difference, reached ~75%.

Finally in the homozygote hotspot region of chromosome 1 (1,600,000-1,900,000bp) only four *Arabidopsis* A genome sequences were inputted into wgsim ensuring a high homozygote alternate allele frequency to the *Arabidopsis* Col-0 reference sequence in this area harbouring the mutation as the proportion of reads with an allele from the *Arabidopsis* A Genome, at positions with a difference, reached ~100%. The mutation that was selected for analysis was in chromosome 1 position 1,700,408 and was a G to A substitution. Although four sequences were used at each point in the genome a maximum of two alleles were possible at any point due to use of only the *Arabidopsis* A genome and the ancestor *Arabidopsis* Col-0 genome sequences therefore this mutant was diploid.

2.5.2 Mutant identification in the artificial diploid dataset

After the Illumina data was generated from the input fasta file it was firstly analyzed, as per section 2.3, using the SHORE/SHOREmap pipeline. *Arabidopsis* Col-0 (TAIR10) was implemented as the reference sequence. SHOREmap 'denovo' was run on the output files with a window size of 100,000bp due to the small 300,000bp window of the artificial homozygote region. The minor allele frequency text file, used as input for 'denovo' acts as a list of heterozygotes. The output from 'denovo' can be seen below in figure 2.7a and an obvious peak appears in the target region of chromosome 1 at ~1,600,000-1,900,000bp. SHOREmap 'annotate' was run over this peak and a list of 641 homozygous SNPs were identified (see Appendix 1, table 4) of which the mutant SNP was 170th. It was identified correctly as a G→A SNP and found in 100% of sequencing reads with a depth of coverage of 81.

The correct interval was identified containing the SNP of interest, but the mutant SNP was not the top ranked SNP, and the longer length of the list of SNPs that were identified was not ideal. This was anticipated due to the previously detailed difficulties of phasing into the homozygote region when creating the dataset and this resulted in a longer length homozygote region. The correct identification of the region containing the mutant SNP provides the proof of concept, however, additional phasing into the homozygote region, as would be seen in a real sample, (the more backcrossing or pooled samples the smaller the interval) could have allowed further homing in on the SNP of interest.

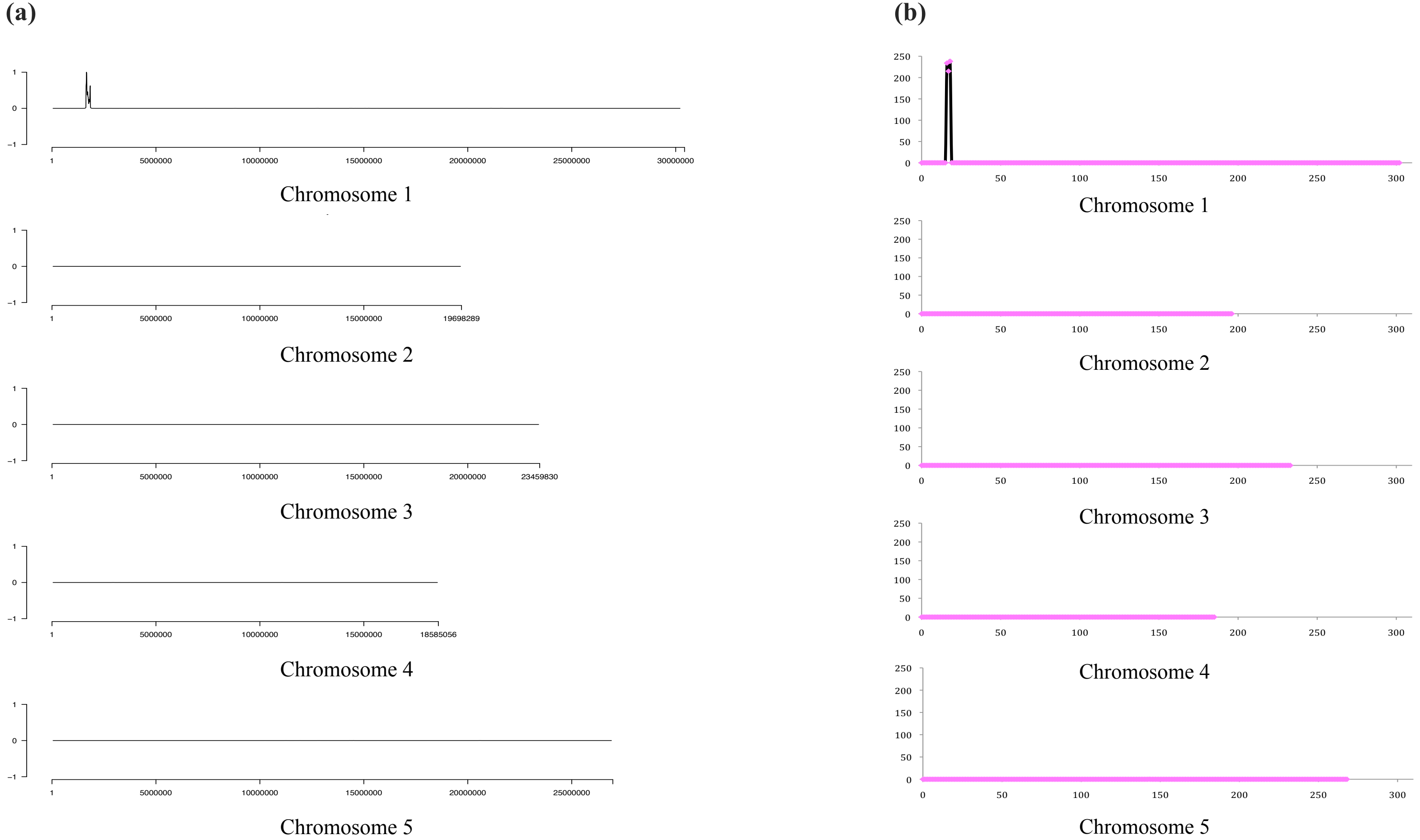


Figure 2.7. Allele frequency analysis of a simulated *Arabidopsis* diploid mutant. (a) ‘SHOREmap denovo’ output pdf file for an artificial *Arabidopsis* mutant mapped to the TAIR10 *Arabidopsis* reference genome (window size 100,000bp used). **(b)** Output from bespoke allele frequency analysis of artificial *Arabidopsis* mutant. Plots describe the homozygote to heterozygote ratio (y axis) per 100,000bp window along each chromosome (x axis)

This investigation was repeated using the bespoke mapping, SNP calling pipeline and allele frequency analysis that was developed here as per section 2.4.2 without the need for additional parameters to account for colour-space reads and the TAIR10 *Arabidopsis* Col-0 sequence was used as the reference. SNPs with more than 80% of the sequencing reads containing the alternate allele were filtered from the output into a homozygote's file. On this occasion a heterozygote SNP was defined as having greater than or equal to 30% and less than or equal to 70% of sequencing reads with the alternate allele and these SNPs were filtered from the VarScan output into a heterozygote's file. Both files were run through the Perl script (Allele-frequency-interval-determination.pl) to calculate the numbers of heterozygotes and homozygotes per 100,000bp interval along each chromosome and the homozygote to heterozygote ratio was calculated. These ratios were plotted and are shown in figure 2.7b. In both analyses in figure 2.7a and 2.7b average depth of coverage was calculated to be ~100.

A peak appears in figure 2.7b in the target region around chromosome 1 windows 16, 17 and 18 corresponding to 1,600,000-1,900,000bp. 3436 SNPs were found within this peak in the VarScan output, these could be narrowed to 430 relevant homozygous SNPs (see Appendix 1, table 5). Within this list the mutant SNP position was identified correctly as a G→A SNP in 100% of sequencing reads with a depth of coverage of 92. Either the bespoke pipeline that was developed here or SHORE can be used effectively for determination of the interval containing the phenotype inducing SNP position within a diploid mutant.

2.6 Mutant identification in a tetraploid using an artificial dataset

2.6.1 Development of an artificial tetraploid sequence dataset

This novel mutant identification pipeline was then trialed on an artificial tetraploid sequencing dataset; an additional diploid mutant was developed to add to the existing diploid mutant created in section 2.5.1 to create this tetraploid mutant. This second diploid was based on the evolution of the B genome of wheat. The program evolver was again used to create this evolved *Arabidopsis* genome (*Arabidopsis* B genome) that was as different from the *Arabidopsis* reference strain Col-0 than the wheat B genome is from its most likely donor *T. turgidum*. To output this evolved sequence evolver took the *Arabidopsis* Col-0 reference sequence as input along with a 'branch length' of 0.0056. This branch length was the value for the genome B in wheat as calculated by Gu *et al.* and creates a substitution approximately every 179 bases.

In the original diploid mutant that was created in section 2.5.1 *Arabidopsis* Col-0 was used to create heterozygosity in the data (sequence data from one input sequence will create homozygotes only). In this case to create the B diploid genome it was more suitable to create a B' genome to compliment the B genome and create heterozygosity within it. The B' genome was a further minimally evolved version of genome B and as such the difference between genomes A and B was calculated and divided by 10 and this figure (0.00037) was used as the branch length along with the *Arabidopsis* B genome for input into evolver.

A phenotype inducing mutation in a tetraploid is likely to be homozygous and in a homozygous hotspot within one genome. As such, the A genome Illumina data that was generated in section 2.5.1 could be re-used, and the B genome data added to it in equal amounts. With no mutation in the B genome, data could be generated for it as follows; the *Arabidopsis* B genome and the *Arabidopsis* B' genome sequences were used once each as input for SAMtools wgsim with the same parameters to create another 100,000,000 read pairs. This Illumina data was merged with *Arabidopsis* Genome A sequencing data, 100,000,000 read pairs for each excluded bias. The resultant dataset contains approximately equal numbers of reads for each of the 2 diploid genomes that make up the tetraploid.

The same features as detailed in figure 2.6 were expected for this tetraploid however the addition of the B genome altered the proportions of reads effected. Due to the presence of two genomes, heterozygotes in one of the two genomes along the chromosomes 2-5 and chromosome 1 3,500,000bp+ were typically found in approximately 25% of sequencing reads (unless two identical heterozygotes were present in both genomes or more than 2 alternate alleles were present; both were found to be rare occurrences which would not effect results). In the 'phased' regions of chromosome 1 approaching the mutation; 0-1,600,000bp and 1,900,000-3,500,000 alternate alleles would be expected in around 37-38% of reads. Finally in the homozygote hotspot region of chromosome 1 (1,600,000-1,900,000bp) a high homozygote alternate allele frequency to the *Arabidopsis* Col-0 reference in this area generated by the *Arabidopsis* genome A mutation would result in an alternate allele in ~50% of sequencing reads.

2.6.2 Mutant identification in the artificial tetraploid dataset

After the Illumina data was generated for the tetraploid it was analyzed identically to the data in section 2.5.2 using the SHORE/SHOREmap pipeline and later the bespoke mutant identification pipeline, again using *Arabidopsis* Col-0 (TAIR10) as the reference sequence. SHOREmap 'denovo' was run on the output files with a window size of 100,000bp. The minor allele frequency text file, used as input for 'denovo', that is intended to act as a representation of heterozygote frequency along the genome, was filtered using awk to remove any minor alleles with a frequency in the sequencing reads lower than 15% and greater than 35% since we expect heterozygotes in this dataset to be found in ~25% of sequencing reads. The outputs from 'denovo' using the filtered minor allele frequency file can be seen in figure 2.8a.

The VarScan output that was generated by the bespoke mutant identification pipeline that this dataset was also taken through was filtered for SNPs with more than 45% and less than 55% of the sequencing reads containing the alternate allele. These SNPs were added into a homozygote's file. A heterozygote SNP was defined as having greater than or equal to 15% and less than or equal to 35% of sequencing reads containing the alternate allele and these SNPs were filtered from the VarScan output into a heterozygote's file. Both files were then run through the Perl script (Allele-frequency-interval-determination.pl) and the numbers of heterozygotes and homozygotes per 100,000bp interval along each chromosome were calculated along with the homozygote to heterozygote ratio. These ratios were plotted to produce the output in figure 2.8b. In both analyses in figure 2.8a and 2.8b average depth of coverage was calculated to be ~200.

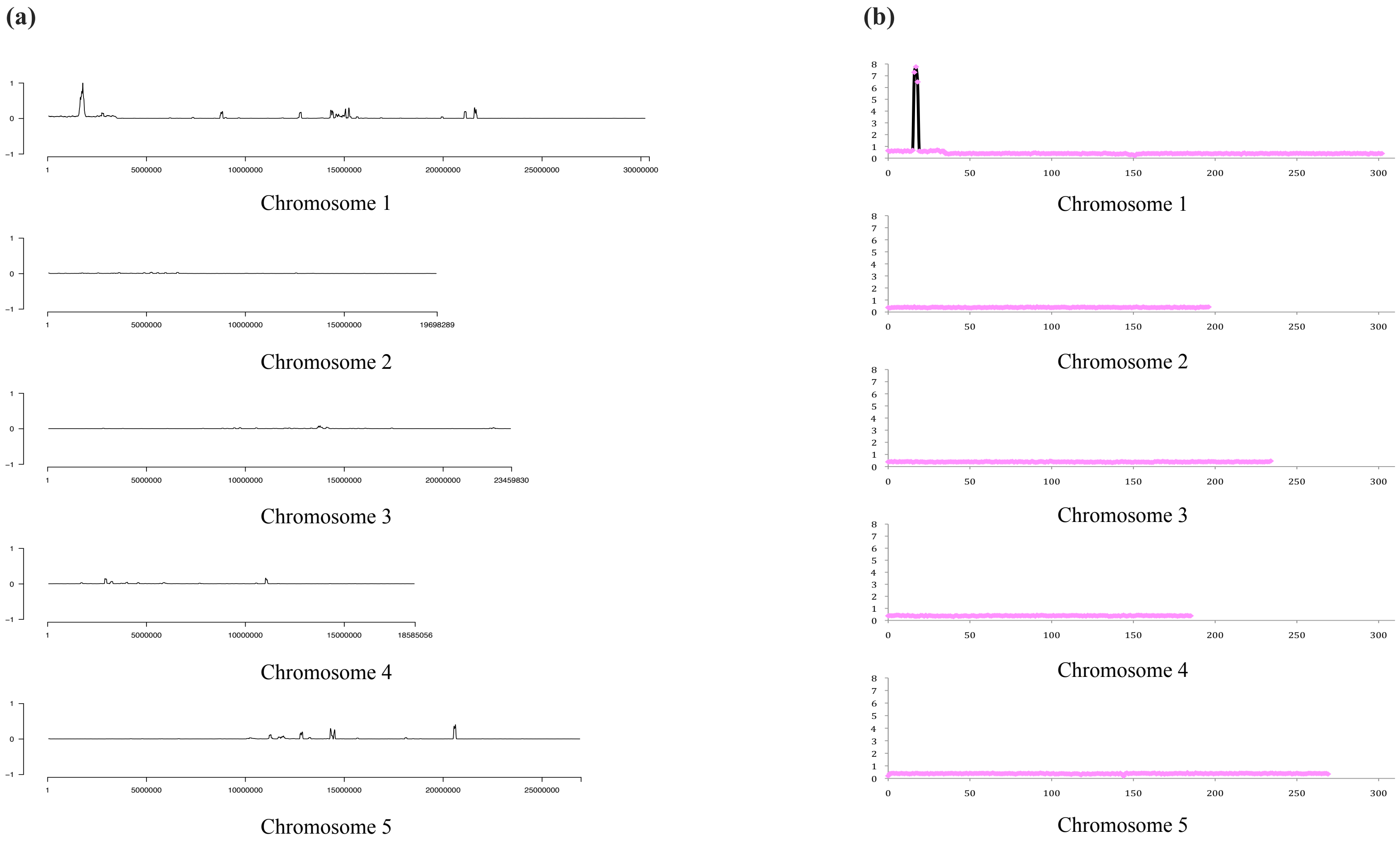


Figure 2.8. Allele frequency analysis of a simulated *Arabidopsis* tetraploid mutant. (a) ‘SHOREmap denovo’ output pdf file for an artificial *Arabidopsis* mutant mapped to the TAIR10 *Arabidopsis* reference genome (window size of 100,000bp used). **(b)** Output from bespoke allele frequency analysis of artificial *Arabidopsis* mutant. Plots describe the homozygote to heterozygote ratio (y axis) per 100,000bp window along each chromosome (x axis).

In figure 2.8a, a peak appears in the target region around chromosome 1 at ~1,550,000-1,940,000bp. SHOREmap 'annotate' was run over this peak and a list of 3006 homozygous SNPs in ~50% of sequencing reads were identified of which the mutant SNP was 737th (see Appendix 1, table 6). It was identified correctly as a G→A SNP and found in ~48% of sequencing reads with a depth of coverage of 100. Figure 2.8b supports these findings since a peak appears in the target region in chromosome 1 windows 16, 17 and 18 corresponding to ~1,600,000-1,900,000bp and on closer inspection 1,600,000-1,850,000bp. 2528 homozygous SNPs in ~50% of the reads were found within this peak in the VarScan output (see Appendix 1, table 7). Within this list the mutant SNP was identified correctly as a G→A SNP and found in 49.57% of sequencing reads with a depth of coverage of the alternate allele of 114. Either the bespoke pipeline or SHOREmap can be used effectively for determination of the correct interval containing the phenotype inducing SNP position within a tetraploid genome.

2.7 Mutant identification in a hexaploid using an artificial dataset

2.7.1 Development of an artificial hexaploid sequence dataset

This novel mutant identification pipeline was finally trialed on an artificial hexaploid sequencing dataset; an additional diploid mutant was developed to add to the existing tetraploid mutant created in section 2.6.1 to create this hexaploid mutant. Creation of this diploid was based on the evolution of the D genome of wheat. Evolver was implemented with the *Arabidopsis* Col-0 reference sequence as input and a 'branch length' of 0.0137 to create an *Arabidopsis* D genome that was as different from Col-0 as the wheat D genome is from its ancestor Ae. *tauschii*. This branch length was the value for the genome D in wheat as calculated by Gu, Y. Q. *et al.* and creates a substitution approximately every 73 bases. It was not ideal that the A and B genome's branch lengths were taken from their evolution from *T. turgidum* while the D genomes branch length was taken from its evolution from *Ae. tauschii,* however, these plants act as hexaploid wheat's genome donors to allow a guideline for diversity between the 3 genomes since little is known about their long-term divergence.

Here, as for the *Arabidopsis* B genome, a minimally evolved *Arabidopsis* D' genome was developed to compliment the *Arabidopsis* D genome and to create heterozygosity within it. To create this the difference between wheat genomes A and D was calculated and divided by 10, this figure (0.00044) was used as the branch length along with the *Arabidopsis* D genome as input for evolver.

A phenotype inducing SNP in a hexaploid is likely to be homozygous and in a homozygous hotspot within one genome, as such the Illumina data that was generated in section 2.6.1 for genomes A and B could be re-used and the D and genome data simply added to it. With no mutation in the *Arabidopsis* D genome, data could be generated for it as follows; the *Arabidopsis* D genome and the *Arabidopsis* D' genome sequences were used once each as input for SAMtools wgsim with the same parameters to create another 100,000,000 read pairs. This Illumina data was merged with the section 2.6.1 *Arabidopsis* Genome A and B sequencing data, 200,000,000 read pairs in total resulting in 100,000,00 read pairs for each genome ensuring no bias between genomes.

The same features as detailed in section 2.6.1 were expected for this hexaploid however the addition of the *Arabidopsis* D genome altered the proportions of reads effected. Due to the presence of 3 genomes heterozygotes in one of the 3 genomes along the chromosomes 2-5 and in chromosome 1 3,500,000bp+ were, in general, in approximately 16.6% of sequencing reads (unless identical heterozygotes were present in 2 or even 3 genomes or more than 2 alternate alleles were present; both were found to be rare occurrences which would not effect results). In the 'phased' regions of chromosome 1 approaching the mutation; 0-1,600,000bp and 1,900,000-3,500,000 alternate alleles were expected in ~25% of reads. In the homozygote region of chromosome 1 (1,600,000-1,900,000bp) a high homozygote alternate allele frequency to the *Arabidopsis* Col-0 reference in this area generated by the *Arabidopsis* genome A results in an alternate allele in ~33% of sequencing reads.

2.7.2 Mutant identification in the artificial hexaploid dataset

After the Illumina data was generated for the hexaploid it was analyzed identically to the data in both of the previous sections using the SHORE/SHOREmap pipeline and later the bespoke mutant identification pipeline, using *Arabidopsis* Col-0 (TAIR10) as the reference sequence. SHOREmap 'denovo' was run on the output files with a window size of 100,000bp. The minor allele frequency text file was filtered using awk to remove any minor alleles with a frequency in the sequencing reads lower than 10% and greater than 20% since we expect heterozygotes in this dataset to be found in ~16.6% of sequencing reads. The output from 'denovo' gave inconclusive results using the filtered minor allele frequency file. The minor allele frequency text file was then re-filtered using awk to remove any minor alleles with a frequency in the sequencing reads lower than 16% and greater than 17% and the output from 'denovo' using the re-filtered file is shown in figure 2.9a.

The VarScan output of the mutant identification pipeline that this dataset was also taken through was filtered for SNPs with more than 28% and less than 38% of the sequencing reads containing the alternate allele. These SNPs were added into a homozygote's file. A heterozygote SNP was defined as having greater than or equal to 10% and less than or equal to 20% of sequencing reads containing the alternate allele and these SNPs were also filtered from the VarScan output into a heterozygote's file. Both files were then run through the Perl script (Allele-frequency-interval-determination.pl) and the numbers of heterozygotes and homozygotes per 100,000bp interval along each chromosome were calculated along with the homozygote to heterozygote ratio. These ratios were plotted to produce the frequency plot shown in figure 2.9b. In both analyses in figure 2.9a and 2.9b average depth of coverage was calculated to be ~300.

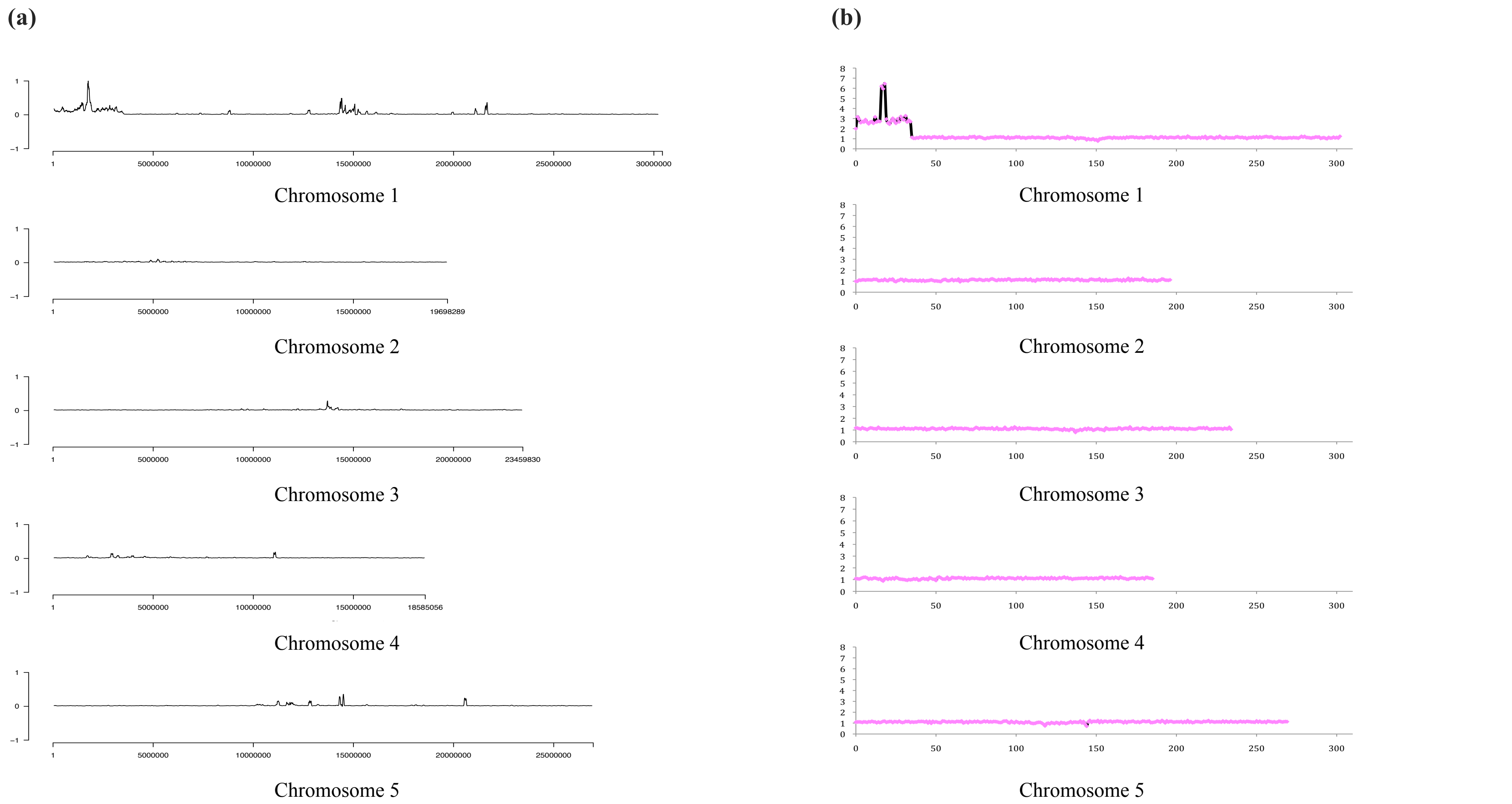


Figure 2.9. Allele frequency analysis of a simulated *Arabidopsis* hexaploid mutant. (a) 'SHOREmap denovo' output pdf file for an artificial Arabidopsis mutant mapped to the TAIR10 *Arabidopsis* reference genome (window size of 100,000bp used). **(b)** Output from bespoke allele frequency analysis of artificial *Arabidopsis* mutant. Plots describe the homozygote to heterozygote ratio (y axis) per 100,000bp window along each chromosome (x axis).

In figure 2.9a a peak appears in the target region around chromosome 1 and on closer inspection this maps to ~1,360,000-1,920,000bp. SHOREmap 'annotate' was run over this peak and a list of 9228 homozygous SNPs in ~33% of sequencing reads were identified of which the mutant SNP was 2368th (see Appendix 1, table 8). It was identified correctly as a G→A SNP and found in ~37% of sequencing reads with a depth of coverage of 105. Figure 2.9b supports these findings since a peak appears in the target region in chromosome 1 windows 16, 17 and 18 corresponding to 1,600,000-1,900,000bp. 6197 homozygous SNPs were found within this peak in the VarScan output (see Appendix 1, table 9). Within this list the mutant SNP was identified correctly as a G→A SNP and found in 36.67% of sequencing reads with a depth of coverage of the alternate allele of 117. Either the bespoke pipeline or SHORE/SHOREmap can be used effectively for determination of the correct interval containing a phenotype inducing SNP position within a hexaploid mutant.

2.8 Conclusions

Section 2.2 acts as a mapping and SNP calling pipeline validation. It demonstrates that the pipeline that was developed using BWA and GATK was successfully able to identify a large proportion of the *ebi-1* specific SNPs that were identified by Ashelford *et al*. (~90%). The SNP that is responsible for causing the *ebi-1* phenotype was successfully identified in the dataset along with the two trends (high proportion of SNPs on the north arm of chromosome 5 and a group of SNPs on chromosome 1) that were found in the original studies.

Here all of the analyses that were carried out using SHORE/SHOREmap on *Arabidopsis* Ws-2 mutants (sections 2.3-2.4) produced clear peaks of interest and a short list of potential phenotype inducing SNPs to be taken forward for further analysis. Schneeberger *et al.* demonstrated that SHOREmap ranks the SNPs that it finds in the defined interval so that the top SNP in the list is likely to be the causative mutation. Here lists of SNPs were taken forward for further investigation, rather than a single top candidate, due to higher numbers of SNPs being close to the identified peak whilst also close to one another within the genome. This is possibly due to false positive SNPs creeping into data and 'cluttering' results (due to low coverage or too few pooled plants) or simply due to the SNP dense nature of the identified region. Despite these issues this is a very useful analysis as the recommended SNP list to be investigated experimentally as potential phenotype inducers has been reduced to a relatively small and manageable number that can be tentatively ranked in order of confidence. The more narrow a region and smaller a list of SNPs for experimental validation is the less time consuming and expensive the analysis. The SNP list can be further narrowed by investigating the functions of the genes associated with the SNPs and their relevance.

In section 2.4 figure 2.4 had secondary smaller peaks in centromeric regions. The repetitive nature of such regions is as such that mapping coverage tends to be excessively high and SNPs found within these regions tend to be ignored. The peaks identified in these regions were therefore disregarded (Round, E. K. *et al.,* 1997). Otherwise, the exemplary smooth 'baseline', with only one peak, that was demonstrated in figure 2.3 (section 2.3) and in the Schneeberger *et al*. analysis, was not seen in figure 2.4, although clear peaks could still be easily identified within the data. In the example dataset tested by SHOREmap's developers Schneeberger *et* al. ~500 plants were pooled and in the dataset detailed in figure 2.3 ~100 plants were pooled (Schneeberger *et al.,* 2009). In the dataset that is shown in figure 2.4 only ~50 plants were pooled. Pooling fewer plants yields fewer markers for definition between a homozygote region (harbouring the phenotype inducing mutation) and a heterozygote region. This could increase the possibility of erroneous calls/false peaks or cause a more uneven baseline. James *et al.* demonstrated that if a sample had been backcrossed to the wild type parent a balance between sequencing coverage and the number of pooled plants would hold the key to successful interval identification. It was noted that at good coverage (upwards of 25x) and a larger pool size (~70 or more) effective interval determination could be achieved, but from this point the effect of further increasing coverage, larger pools or more backcrosses was diminished and similar sized intervals were identified (James *et al.,* 2013). In section 2.3/2.4 coverage was not ideal in either dataset (less than 20x) however the large pool size of ~100 allowed effective interval identification while a pool size of ~50 compromised the analysis. In this dataset phenotype subtlety, i.e. difficulty in selecting plants of the correct phenotype for pooling, was reported which could have been a factor.

Using a standard mapping and SNP calling pipeline (BWA/SAMtools/VarScan) combined with a bespoke allele frequency analysis (section 2.4.2) the investigation successfully produced largely the same results as SHOREmap produced for the same dataset (section 2.4.1) (~87% SNP conservation). It also identified the interval containing the mutant phenotype inducing homozygote in a diploid organism. The challenge was the adaptation of this analysis to a polyploid wheat strain. In section 2.5-2.7 it has been demonstrated that it is in fact possible to use this method to identify a region harbouring a causal mutation in an artificial diploid, tetraploid and even a hexaploid genome despite the fact that the phenotype inducing SNP was only homozygous in 1 of the 3 genomes i.e. present in ~1/3 of sequencing reads and thus potentially more difficult to detect.

This polyploid mutant identification analysis was successful not only using a bespoke pipeline but also through manipulation of the SHOREmap output/input files to almost 'trick' it into identifying homozygotes and heterozygotes in only one of a number of polyploid genomes. This adaptation of SHOREmap shows that the methodology is applicable to polyploid species although this was unnecessarily complex due to this being an un-intended use for SHOREmap. The bespoke pipeline that was developed here was quicker to implement and it was more straightforward to use. Adaptation of SHOREmap for a polyploid could only be achieved as follows (detailed for a hexaploid); SHOREmap's lower limit of detection of a homozygote and heterozygote could be re-defined i.e. to just below 1/3 and 1/6 respectively, however, the upper limit of detection of a heterozygote could not be altered and remained at 80% resulting in an overlap of homozygote and heterozygote calls. No definition between heterozygotes and homozygotes hindered mapping interval identification. Therefore multiple intermediate SNP files that are utilized by SHOREmap had to be manually located and edited to only include what was deemed in the hexaploid context to be heterozygote/homozygote calls. Since SHOREmap defines SNP files separately, and in different formats, as heterozygote, homozygote and minor allele these files were re-defined and re-formatted as required. It was these filtered files that could be used as input for 'SHOREmap denovo' and 'SHOREmap annotate'.

This adaptation of SHOREmap was not ideal; it was long winded and required a relatively high degree of programming knowledge. This outlined the need for an alternative mutant identification method that can be easily tailored to use on a polyploid genome. When considering all of the artificial wheat mutants, SHOREmap identified longer lists of potential mutation inducing SNPs compared to the bespoke pipeline, increasing the downstream effort to locate the SNP of interest e.g. in the hexaploid mutant SHOREmap defined a peak that was 260Kbp longer and contained 3031 more SNPs. This was observed using comparable parameters and heterozygote definitions where possible. For all these reasons, the bespoke mutant identification pipeline will be the analysis of choice to be used, and/or adapted for use, for any further frequency based mutant identification investigations within this project.

The artificial datasets themselves, although not a perfect simulation of wheat, allowed the trial and development of this bespoke mutant identification pipeline on a polyploid species *in silico* prior to the costly generation of sequencing datasets for polyploid wheat species. A similar approach to simulate an artificial sequencing dataset has been used by Brenchley *et al.* to simulate maize whole genome shotgun sequencing data (Brenchley *et al*. 2012).

NB: The Perl script that is used in this analysis; Allele-frequency-interval-determination.pl, is here in its preliminary developmental stage. Further development that was ongoing throughout this project, utilizing knowledge that was gained from multiple datasets, has resulted in further pipeline development and the depreciation of this particular script.

Chapter 3. Validation of a wheat gene capture array

Here a genomic target enrichment approach was validated and used to capture the gene rich regions of several hexaploid bread wheat varieties, reducing the sequencing cost while still allowing SNP calling and varietal comparison across the majority of wheat's genic sequence. A detailed analysis and validation of the gene capture array is provided to outline its target sequence and improved enrichment capability over the original exome capture array. A pseudo-chromosome based reference sequence was developed from the gene capture array target sequence with a long-range order of genes based on synteny of wheat with *Brachypodium distachyon.*

3.1 Introduction

This chapter will assess the utility of mapping-by-sequencing as a methodology to rapidly identify genes responsible for key agricultural traits in complex, largely uncharacterized genomes. Mapping-by-sequencing typically involves the generation of high coverage shotgun sequence of a scored F2 mapping population. In wheat, due to its vast 17 Gb size and high repeat content (Choulet *et al.*, 2010), it is expensive to generate high coverage datasets and challenging to analyze such data. To reduce this genome complexity methods such as transcriptome sequencing (Trick et al. 2012) or targeted enrichment sequencing (Winfield et al 2012), have been proposed to reduce the need for whole genome re-sequencing.

In section 1.6.2 two NimbleGen in solution capture probe sets were introduced that enable the application of targeted DNA enrichment to wheat. The development of the exome capture array is detailed figure 1.11b. This initial array was based on the cDNA sequence that was generated for the hexaploid wheat strain Chinese Spring using the Roche 454 sequencer. This cDNA formed the design-space contigs across which capture probes were tiled. At that time the full genomic sequence of wheat was not yet available and as such the exome capture array was used for gene enrichment and subsequent SNP calling in hexaploid wheat. The issue with using cDNA sequence is that only those genes that were expressed in the sequenced sample at the time of processing are included in the array.

As wheat's full genomic sequence and the resulting assembly became available in 2010 a further in solution capture probe set was developed using similar techniques (see figure 1.11a). Genic regions of this wheat genomic sequence formed the design-space for the gene capture array that the capture probes were tiled across and contained the majority of the genic regions of wheat. This was a clear improvement on the exome capture array, containing all of the genes that were sequenced, whether they were expressed or not expressed, plus transcribed and non-transcribed sequence. It is therefore this gene capture array that has been preferentially utilized for mutant identification later in this project.

Here a detailed analysis and validation of the gene capture array, plus a comparison of its features, is provided to outline it's improved enrichment capability over the original exome capture array. Using the array probe set in solution to enrich wheat plant(s) of interest and subsequent sequencing and mapping of the data back to the array design-space facilitated such an analysis. Each array's design-space was used to map non-enriched data to and then to map data that had been enriched to allow a comparison between the two. This analysis highlighted the benefits of enrichment and demonstrated the array's efficacy in general whilst discriminating which array performed better (section 3.4).

For the exome and gene capture arrays in solution an analysis was carried out on each design-space, to determine the approximate number of genes that were represented, that would ultimately be present in the enriched sequence that was generated. Section 3.5 includes a detailed comparative analysis of the predicted target sequence for each array in wheat. In section 3.2 the BLAST (Basic local alignment search tool) analysis comparing wheat cDNA sequence to the wheat gene capture array design-space allowed approximate determination of the proportion of coding and non-coding sequence that is contained within the enrichment array.

The gene capture array design-space contigs were ordered and concatenated into 7 pseudo-chromosomes that are representative of the genic regions of wheat. This method is detailed in section 3.3 and is based on each contig's synteny to the *Brachypodium* genome, a close grass relative of wheat. Sorting the array probes employed comparative genomic organization with a genome for which a complete and well-annotated reference genomic sequence is available. Rice (*Oryza sativa* L.), a model species, has been previously used in comparative analyses of wheat for molecular mapping and gene isolation (Liu and Anderson, 2003). Synteny and gene homology between rice and the other cereal genomes e.g. wheat is extensive (Goff *et al.,* 2002) but numerous studies show that co-linearity between the two species can frequently break down due to translocations, deletions and gene duplications (Bennetzen and Ma, 2003). The *Brachypodium* genome has become a popular alternative to rice due to its completed and annotated genomic sequence and data suggesting better co-linearity exists between it and wheat than between rice and wheat (Cao1 *et al.,* 2012). For this reason markers were designed between *Brachypodium* and wheat to allow association of contigs with a region of wheat to enable ordering and sorting into pseudo-chromosomes.

Pseudo-chromosomes assist in the application of chromosome dependent mutant identification methods and visualization tools. Such methods, e.g. those derived from the SHOREmap approach in *Arabidopsis,* are based on a sliding window analysis along a chromosome. Also for visualization of mapping statistics e.g. analysis of variable sequencing coverage, such analyses of sliding windows along each chromosome can be very useful.

As detailed in section 3.6 the gene capture array was used to enrich a number of common hexaploid wheat varieties including; Truman, Rialto, Utmost and Chinese Spring. NimbleGen carried out this enrichment and subsequent sequencing externally. Enrichment of the hexaploid wheat varieties plus the standard reference variety Chinese Spring enabled identification of varietal SNPs between the wheat crops allowing downstream comparisons between the varieties and visualization of any regions/hot spots for SNP conservation or difference. Here the hope is to develop methods to enable easy identification of important genes/mutants/characteristics within or between wheat plants to facilitate crop improvement with use of target enrichment to facilitate such analyses.

3.2 Intron and exon modeling

The gene capture array design-space was used as input for a BLAST search against wheat cDNA sequences (pre-determined e-value cutoff of 0.001). Regions of alignment identified predicted exons within the input design-space sequences. The wheat cDNA sequences used were 4 million 454 reads from a normalized Chinese Spring library that was sequenced by Brenchley, R. *et al.* (Brenchley *et al.,* 2012). 71% of the design-space contigs had, at least in part, a hit to the wheat cDNA. Of these 71%, the average percentage of each design-space contig that was found to be exonic sequence was ~49%. The average percentage overall of gene capture design-space that was found to be exonic sequence was ~35%.

As a basis for comparison the same investigation as above was carried out for the exome capture array. We would expect a much higher proportion of this array to be correctly identified as exonic sequence due to its cDNA based design, although perhaps not 100%, as although the cDNA sequence produced by Brenchley *et al.* was used to design this array, additional sequence from the public wheat EST collection was also used. The exome array design-space was used as input for the BLAST alignment against the wheat cDNA sequences. This time 70% of contigs had an alignment hit to the wheat cDNA sequence. Of these the average percentage of each design-space contig that was found to be exonic sequence was 91% i.e. the percentage of the array design-space that was found to be exonic sequence was higher at ~64% (detailed in table 3.2).

3.3 Ordering array probes

Data is available detailing 800 wheat markers (see Appendix 2, table 1), their positions along the wheat chromosomes and their respective alignment positions within the *Brachypodium* genome. If we know which *Brachypodium* genes that our array probes align to then they can be ordered into 7 pseudo wheat chromosomes according to the known associated marker positions.

Although 3 separate genomes exist in wheat, gene co-linearity appears to be retained between them (Gu *et al.,* 2004) and it is estimated that homeologous gene copies differ by only 1 in 100bp with varietal SNPs occurring at the rate of approximately 1 per 233bp (Barker and Edwards, 2009). Array probes are designed to be able to hybridize to a target in the presence of a limited number of mismatches between the probe and the target allowing subsequent target enrichment. Given this we can conclude that, in most cases, if a single sequence is used to represent the 3 genomes of wheat, one probe under 100bp in length is likely to enrich all 3 homeologs.

Downstream mapping of the enriched sequence data can also be successfully implemented in the presence of mismatches to the reference genome (typically 1-4 per 100bp sequencing read) to allow SNP detection. With such a low frequency of homeologous SNPs one ordered probe set is therefore also likely to be sufficient as a reference sequence for downstream mapping and SNP calling of enriched sequencing data, even in the presence of inter-varietal SNPs. One collapsed wheat reference representing all 3 genomes of wheat may in fact be desirable for downstream mapping analyses; considering homeologous SNP frequency, a large proportion of ~100bp mapping regions of the A, B and D genomes would be identical or have a small number of differences, making allocation of some of the ~100bp mapped sequencing reads to a single genome impossible and introducing a high proportion of non-uniquely mapped reads. Such reads are usually removed from SNP calling analyses resulting, in this case, in potentially a large amount of unused data.

Here BLAST (version 2.2.17) was used to search for similarities between the gene capture array design-space and *Brachypodium*. BLASTn was implemented with an e-value cutoff of 0.001. The most likely gene hit for each array design-space contig was filtered out ordered by lowest e-value, highest score and then longest length hit. 68% of the contigs had an alignment to *Brachypodium*. The mid-point of the aligned contig in *Brachypodium* was taken as the ‘contig position’ to enable calculation of the nearest wheat marker as a measure of distance along the relevant *Brachypodium* chromosome. The probes could then be ordered into pseudo-chromosomes using wheat marker positions (see figure 3.1 for illustration).

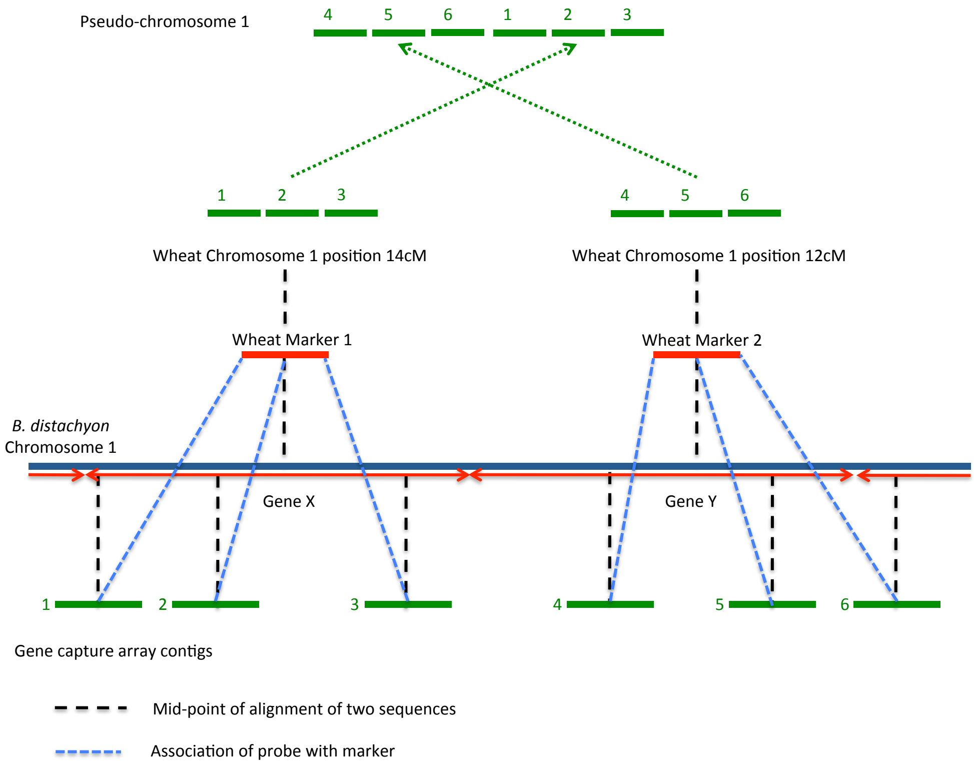

Pseudo-chromosome 1
4
5
6
1
2
3
1
2
3
4
5
6
Wheat Chromosome 1 position 14cM
Wheat Chromosome 1 position 12cM
Wheat Marker 1
Wheat Marker 2
B. distachyon
Chromosome 1
Gene X
Gene Y
1
2
3
4
5
6
Gene capture array contigs
Mid-point of alignment of two sequences
Association of probe with marker

Figure 3.1. Construction of pseudo-chromosomes from the gene capture array design-space contigs using *Brachypodium*-wheat markers. Here the ordering of 6 design-space contigs (green) into a section of pseudo-chromosome 1 is illustrated. The central point of a contig’s alignment with *Brachypodium* is used to calculate it’s nearest *Brachypodium*-wheat marker as a distance in *Brachypodium*. The order of these contigs in *Brachypodium* determines local ordering of contigs around the marker. These ordered groups of contigs that are associated with an individual marker are then further assembled and concatenated into pseudo-chromosome sequences based on marker positions in wheat.

This analysis was replicated, as per figure 3.1; on this occasion the newly sequenced dataset that was available for barley (*Hordeum vulgare L.*) was used in combination with 1822 barley-wheat markers (The International Barley Genome Sequencing Consortium, 2012b). For this purpose only barley sequences that were assigned to a specific chromosome position in the genome would be useful and as such, this dataset contained contigs that spanned 3.9

Gb and had been assigned genetic positions (The International Barley Genome Sequencing Consortium, 2012). The aim of this analysis was to increase the available input for the pseudo-chromosomes since barley is a closer relative to wheat than *Brachypodium* is. When a similar BLAST search to the one above was carried out, using the gene capture array design-space and the barley contigs, an increased number of wheat design-space contigs could be aligned to the barley sequence (81%) and pseudo-chromosomes could then be constructed from these using the barley-wheat markers. This increased the number of design-space probes that could be included in the pseudo-chromosomes, improving the analysis.

Over 90% of the design-space contigs were aligned to a barley region where another wheat contig, aligning to a different barley sequence, had the same reported location. This is because only 12% of the barley contigs that were used in the BLAST alignment were associated with a unique position on a barley chromosome. Multiple, partially overlapping, barley contigs have been anchored to the barley genome using the same marker and have been assigned this same position, although its exact location along the contig is not detailed. Therefore their relative positions cannot be accurately discriminated. Figure 3.2 details the problems that were encountered as a result of this. Ultimately although a group of design-space contigs could be correctly assigned a chromosome position using anchored barley contigs and wheat-barley markers, at this chromosome position, ordering of the array contigs using barley could only be an approximation at a local level.

The barley-based pseudo-chromosomes were compared to the *Brachypodium* based pseudo-chromosomes to determine which were more suitable for further use. The bulk B pooled wheat mutant dataset that was studied in section 4.2 was analyzed using both pseudo-chromosome assemblies. When an initial allele frequency analysis was run along each chromosome, plotting the raw frequency of bulk B homozygotes that were conserved with its parental line per 100,000bp window, the plot that was gained using *Brachypodium* was found to generate significantly less noise and a much more convincing peak of interest was observed at the end of chromosome 3 (see Appendix 2, figure 1). Therefore the *Brachypodium* based order of array probe sequences for the pseudo-chromosomes was used preferentially throughout this project. The benefit of a complete assembled and annotated reference genome for pseudo-chromosome assembly is clear.

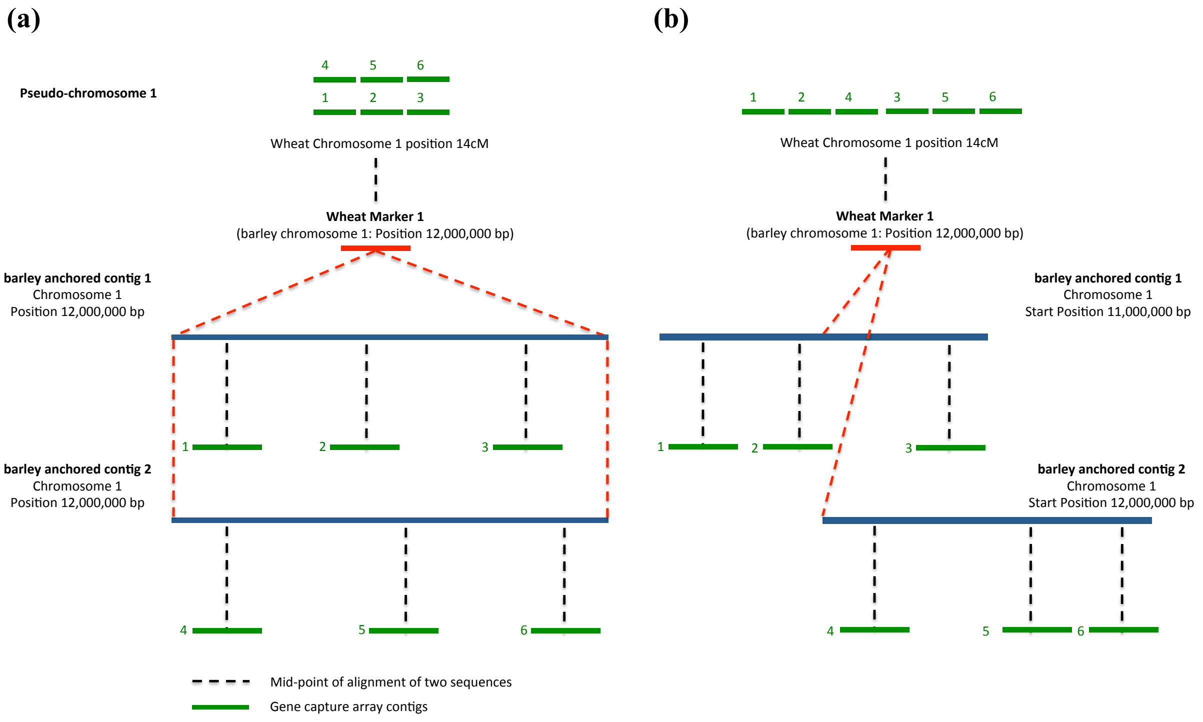

(a)
Pseudo-chromosome 1
4
5
6
1
2
3
Wheat Chromosome 1 position 14cM
Wheat Marker 1
(barley chromosome 1: Position 12,000,000 bp)
barley anchored contig 1
Chromosome 1
Position 12,000,000 bp
1
2
3
barley anchored contig 2
Chromosome 1
Position 12,000,000 bp
4
5
6
(b)
1
2
4
3
5
6
Wheat Chromosome 1 position 14cM
Wheat Marker 1
(barley chromosome 1: Position 12,000,000 bp)
barley anchored contig 1
Chromosome 1
Start Position 11,000,000 bp
1
2
3
barley anchored contig 2
Chromosome 1
Start Position 12,000,000 bp
4
5
6
Mid-point of alignment of two sequences
Gene capture array contigs

Figure 3.2. Using barley for the construction of pseudo-chromosomes from the gene capture array design-space contigs. Adapted from figure 3.1; ordering 6 design-space contigs (**green**) into a section of pseudo-chromosome 1 utilizing barley for comparative analyses. **(a)** Multiple barley contigs (**barley anchored contigs 1 and 2**) have been assigned the same barley position but its exact location along the contig is not detailed therefore the relative positions of the barley contigs cannot be discriminated. As a result non-redundant wheat gene capture array design-space contigs appear to align to overlapping regions in barley. These ordered groups of contigs (numbered **1-3** and **4-6**) could be assigned a chromosome position using wheat-barley markers although local ordering at this location is an approximation. **(b)** Defining relative contig positions in barley enables local ordering of wheat sequence i.e. here the start positions of the barley-anchored contigs 1 and 2 have been defined.

The release of the wheat chromosome assemblies by the International Wheat Genome Sequencing Consortium (IWGSC) allowed further interrogation of the current pseudo-chromosome model. These chromosome sequences were in fact a collection of contigs that had been sorted by wheat chromosome (no large scale chromosomal anchoring or order was available). Whilst these sequences would not aid validation of the ordering of the pseudo-chromosomes they would give an indication of how much of each pseudo-chromosome's sequence had been placed into the correct chromosomal bin. A BLAST search (BLASTn with an e-value of 0.001) was carried out using the gene capture array design-space and the wheat chromosome sorted IWGSC contigs. Over 80% of the design-space contigs that were concatenated into the pseudo-chromosomes were found to have been associated with the correct chromosome. The error rate for the chromosomal sorting of the IWGSC contigs is unknown but cases where scaffolds appear to have been associated with incorrect chromosomal arms have been found as the purity of the flow-sorted DNA used to generate the sequence is not 100% (IWGSC, 2014).

3.4 Comparative enrichment study (exome array versus gene capture array)

3.4.1 Mapping analysis

The exome array bait probes were tiled across 198,056 design-space contigs that were used as reference sequences in all relevant mapping analyses and ranged from a minimum length of 51 to a maximum length of 2467. Similarly in all analyses that involved the gene capture array, these bait probes were tiled across 169,345 design-space contigs that were used as reference sequences in mapping analyses and ranged from a minimum length of 100 to a maximum length of 13168.

The hexaploid wheat variety Rialto was enriched using the gene capture array and sequenced externally using Illumina sequencing technology (Genome Analyzer IIx). The resultant paired end sequencing dataset was mapped to the gene capture array design-space as a fragment library, due to the shorter length of the reference sequence contigs, using BWA (v 0.6.2) short read mapping (prior to the 2013 update to BWA-backtrack). Indexing of the reference sequence involved use of the 'IS' algorithm and during alignment of reads to the reference 4 mismatches were allowed per sequencing read. All unmapped, non-uniquely mapped and duplicate reads were later removed using SAMtools. The steps involved in this analysis are shown in figure 1.6 and example commands are outlined in the command line appendix sections 1 and 4. Rialto was also enriched using the exome capture array in solution and sequenced using SOLiD sequencing technology. This dataset was mapped to the exome capture array design-space using the same steps, filtering and parameters although additional parameters to allow processing of colour-space SOLiD reads were used.

Non-enriched DNA (whole genome sequencing) from the wheat variety Rialto, that was sequenced externally using SOLiD sequencing technology, was also mapped separately to both the gene capture array design-space and the exome capture array design-space using the same steps and parameters as detailed for the enriched Rialto above ensuring that parameters to allow processing of colour-space SOLiD reads were used. The program csfasta2fastq was used to convert the SOLiD sequencing csfasta/qual files into a fastq file, which is used as input for BWA.

SAMtools mpileup (v 0.1.16) was implemented on all mapped datasets and finally SNP calls were filtered out using VarScan (VarScan.v2.2.3.jar) with the following parameters: discard SNPs covered by 20 or fewer reads, discard sequencing reads with a quality less than 20 and if the alternate allele has less than 2 supporting reads passing the quality filter discard it. For this SNP analysis the tool awk was implemented to remove indels from the VarScan output.

This analysis enabled a comprehensive comparison of enrichment quality between the original exome capture array and the gene capture array. The results gained are shown below in table 3.1. The non-enriched data maps with a deeper coverage on average to the gene capture array compared to the exome array, it also maps to more of the gene capture array i.e. ~42% compared to ~31%. Although this gene capture array is double the size of the exome capture array it has fewer design-space contigs since average contig length is ~654bp while for the exome capture array it is ~205bp. It is therefore not surprizing that a higher depth and more extensive coverage has been achieved using the gene capture array with use of significantly longer reference contigs. This could also account for the ability to confidently call a higher number of SNPs using the gene capture array.

Wheat Variety	Mean depth	% Of array probes mapped	Mean % coverage array probe	StdDev coverage depth	% Of reads mapped	Total reads	SNP No.
cDNA array							
Rialto (enriched)	36.5	86.4	86.1	18.8	29.4	106,435,597	146527
Rialto (non-enriched)	11.5	58.8	53.4	6.5	0.38	1,725,138,247	57222
Genomic DNA array							
Rialto (enriched)	268.3	98.6	96.1	183.0	49.9	642,311,196	517022
Rialto (non-enriched)	17.1	88	48.1	14.7	1.25	1,725,138,247	124021

Table 3.1. Exome capture array versus gene capture array. Mapping statistics for enriched and non-enriched Rialto to the exome capture array design-space and also to the gene capture array design-space.

The enriched Rialto data, as anticipated, had a much deeper coverage than the non-enriched Rialto on average across both of the arrays. Approximately ~6x more sequence data was generated for the gene capture array than for the exome capture array and, as it is double the size of the exome capture array, we would expect ~3x more coverage overall, yet the enriched Rialto maps with over 7x deeper coverage to the gene capture array. It is likely that this is because the gene capture array generates less off target sequence data; almost 50% of its sequence reads can be mapped compared to ~30% of the exome capture array reads. Data also maps to ~95% of the gene capture array design-space but to only ~65% of the exome capture array design-space suggesting that a greater proportion of the gene capture array bait probes are enriching effectively in comparison to the exome capture array. Of the unmapped sequencing data typically ~63% of sequencing reads include repetitive sequence.

For the exome capture array overall 21,077 contigs out of 198,056 were not mapped to (11% of reference contigs). However, for the gene capture array overall only 579 contigs out of 169,345 were not mapped to, this amounts to less than 1% of the reference contigs and shows a great improvement. 2,399 contigs (~1%) in the exome array had a high depth of coverage (over 3 SD from the mean) whilst only 2,415 contigs had a high depth of coverage in the gene capture array amounting to less than 1% of the reference contigs. The gene capture array performed enrichment more efficiently.

3.4.2 SNP analysis

A comparison of the homeologous SNPs that could be identified in enriched and non-enriched Rialto data for each array was carried out to ensure that enrichment did not affect SNP calling i.e. that they enrich all three wheat genome copies effectively. The low average non-enriched data coverage was potentially an issue. It was possible that SNP alleles found in the enriched data would not be picked up at all, or at a depth high enough to confidently call SNPs, in non-enriched data due to comparatively low coverage. Conversely it was possible that low frequency SNP alleles in the non-enriched data could be proved to be false positive SNP calls in the higher coverage of the enriched data. The following technique was adopted for this particular situation; SNPs were only considered for comparison between the non-enriched and enriched datasets if they were in regions that were mapped to in both datasets with a depth greater than or equal to 20 and if the alternate allele from one dataset is found in the raw reads of the other (or in the case of an ambiguous alternate allele if both alleles represented are seen) then a SNP was defined as conserved. As detailed in section 3.6 VarScan outputs multiple alternate alleles for one position if they pass quality filters therefore even if multiple homeologous SNP alleles for one position are seen then they could all be validated.

For the exome array 86% of SNPs that were found in the enriched data could be identified in the non-enriched data. The remaining 14% of SNPs from the enriched dataset that were not identified in the non-enriched tend to be low frequency alternate alleles making them difficult to define in the low coverage of the non-enriched data. 96% of SNPs found in the non-enriched data could be identified in the enriched dataset. The remaining 4% of non-enriched data SNPs that could not be identified in the enriched data also showed evidence of SNPs with low frequency alternate alleles or low quality and as such may have been proved to be false positives in the high coverage gained by the enriched data.

For the gene capture array 97% of SNPs found in the enriched data were found in non-enriched data with only 3% that could not be seen at all. 96% of SNPs found in the non-enriched data were found in the enriched data with only 4% that could not be seen at all. The same reasoning as seen in the exome capture array could be attributed to the 3 and 4% of SNPs that could not be seen. Importantly the proportion of unseen SNP alleles was consistently low in both of the enriched datasets although an improvement was noted with use of the gene capture array.

3.5 Comprehensive analysis of the array targets

Both the exome capture array and gene capture array were run through a BLAST search of the *Brachypodium* gene database (International Brachypodium Initiative, 2010) (e-value cutoff of 0.001). When the exome capture design-space was analyzed 96,014 contigs had hits (48%) whilst 70.4% of the *Brachypodium* genes were covered. 100,006 contigs in the gene capture design-space had hits to *Brachypodium* (59%) and 82.7% of the *Brachypodium* genes were covered. As anticipated, the same pattern was observed when the *Brachypodium* gene database was replaced with the *Brachypodium* exon database (International Brachypodium Initiative, 2010) whereby 45% of the exome capture design-space contigs had hits and 84% of the *Brachypodium* genes had exons that were hit. 55% of the gene capture design-space contigs hit *Brachypodium* exons and 93% of the *Brachypodium* genes had an exon that was hit. As a comparison, wheat cDNA sequences from the full Chinese Spring library that was sequenced by Brenchley, R. *et al* and used to develop the exome array were subjected to the same analysis; 71% of the cDNA sequence was hit and 82% of the *Brachypodium* genes had 1 or more exon hit (highly comparable to the 82.7% of genes hit by the gene capture array) (Brenchley *et al*. 2012).

Of the remaining array contigs that did not hit the *Brachypodium* gene database, an additional 37,685 contigs from the gene capture array hit the wheat cDNA to make a total of 81% of the array contigs hitting either *Brachypodium* genes or wheat cDNA. This is an improvement on the 56,180 additional contigs from exome capture array hitting the wheat cDNA to make a total of only 77% of the array contigs hitting either *Brachypodium* genes or wheat cDNA i.e. likely to be hitting wheat genes.

In a separate investigation how much in total of each array design-space hits the wheat cDNA sequence (1.8Gb) alone was analyzed i.e. potential gene space and in turn how much of the wheat cDNA is hit was established. For the gene capture array 119,503 contigs hit the cDNA (71%) and ~99% of the cDNA was hit. For the exome array 138,368 contigs hit the cDNA (70%) and ~93% of the cDNA was hit. It was not surprising that a larger proportion of the exome array design-space was predicted to be transcribed i.e. hit wheat cDNA (~64% compared to ~35% in gene capture array) as this array was modeled solely on cDNA sequence. The gene capture array is still an improvement hitting slightly more wheat cDNA whilst still including non-coding sequence information. These results are summarized in table 3.2.

	Number of design-space contigs hitting wheat cDNA	**Average % of each design-space contig hitting wheat cDNA**	**% Wheat cDNA that is hit by design-space contigs**	**% design-space contigs that hit Brachy genes**	**% Brachy* genes hit by design-space contigs**	**% design-space contigs hitting wheat cDNA or Brachy genes**	**% design-space contigs hitting Brachy exons**	**% Brachy genes with an exon hit by an design-space contig**	**Non-mapped design-space contigs**	**Contigs with depth of coverage >3 SD from mean**
cDNA array	138368 (70%)	91	92.8	48	70.4	77	45	84	21077 (11%)	2399 (~1%)
Genomic DNA array	119503 (71%)	49	98.6	59	82.7	81	55	93	579 (<1%)	2415 (~1%)

Table 3.2. Exome capture array targets versus gene capture array targets. Summary of array design-space contig targets for the exome capture and the gene capture array. Number of *Brachypodium* genes: 32255, number of wheat cDNA contigs: 97481 and number of *Brachypodium* exons: 167291. *Brachy; *Brachypodium distachyon.*

3.6 Mapping and SNP identification in four wheat varieties

Following the previous enrichment and sequencing of the hexaploid wheat variety Rialto, three additional hexaploid bread wheat varieties; Chinese Spring, Truman and Utmost were enriched using the gene capture array and paired end sequenced using Illumina sequencing technology (GAIIx). Enrichment and sequencing was carried out alongside the three varieties in a repeat analysis for Rialto and all four datasets were created externally by NimbleGen. Here the datasets were mapped to the gene capture array design-space and SNP calling was performed using the same pipeline/parameters detailed for the enriched Rialto dataset that was mapped to the gene capture array (section 3.4.1). Mapping efficacy is summarized below in table 3.3. The analysis of the four varieties would allow varietal SNP identification within the enriched gene set.

Wheat variety	Further details	Sequencing technology	Mean Depth of coverage	Coverage of reference (%)	Median Depth of coverage	Q95* Depth of coverage	SNP No.
Rialto (enriched)	winter wheat	Illumina	206	98.6	194	571	733,576
Rialto (non enriched)	winter wheat	SOLiD	17.1	88	7	42	124,021
Truman	winter wheat	Illumina	197	98.7	186	548	722,801
Utmost	spring wheat	Illumina	254	98.8	235	720	713,291
CS	spring wheat	Illumina	165	99.5	156	464	614,885

Table 3.3. Mapping statistics for four varieties of wheat. Coverage data for 4 varieties of wheat mapped to the gene capture array design-space. *Q95 of coverage indicates the depth of coverage for which 95% of data points are lower than or equal to.

When identifying SNPs across hexaploid datasets we expect, in the main, to see heterozygotes or homozygotes in 1 of the 3 genomes only, thus one reference and one alternate allele overall. At the time of this investigation SNP callers were mainly intended for use on datasets derived from a diploid organism and these tools fail to detect regions deviating from the one reference and one alternate allele genotype i.e. where more than one different alternate allele is seen. This typically results in less than a 5% loss in SNPs and will have little or no impact on many analyses. However, since this analysis is a comprehensive comparison of 4 hexaploid wheat varieties it would be desirable to record all alternate alleles for each SNP position. With this in mind for this analysis any alternate allele in >10% of the sequencing reads will be recorded as a potential SNP allele.

VarScan is fine tuned for pooled datasets and is able to pick up the low frequency alternate alleles that we are also likely to see in wheat. It also outputs one line for every alternate allele that passes filters for a SNP call per position thus multiple lines per SNP position can potentially be seen i.e. multiple alternate alleles. This is useful for defining regions with multiple alternate alleles, although since VarScan is expecting a diploid with only 1 alternate allele, for each individual alternate allele it calculates the percentage of reads with that alternate allele as a proportion of the number of those reads plus the reference reads. As a result the percentage of reads containing the alternate allele can be overestimated if there is more than one alternate allele since this additional alternate allele's sequencing reads are not included in the total coverage at that position. Figure 3.3 shows a snapshot of a typical VarScan output where there is more than one alternate allele and the accurate percentage of each alternate allele in the sequencing reads at that position is compared to the VarScan analysis output. At this position 50 reads contain the reference allele A, 50 reads contain a G and 50 reads contain a T so the correct proportion of reads for each alternate allele G and T is ~1/3 and the proportion of reads with the reference allele is also ~1/3, as shown, VarScan defines this incorrectly.

						VarScan	Accurate
Probe	Position	Reference	Alternate	Reference	Alternate	% Alternate	% Alternate
		Allele	Allele	Reads	Reads	Allele	Allele
1	100	A	G	50	50	50	33.3
1	100	A	T	50	50	50	33.3

Figure 3.3. Extract from a VarScan SNP call output at a position with more than one alternate allele. Here the first 7 columns are extracted directly from a VarScan SNP call output and the last column details an accurate calculation of the percentage of sequencing reads containing the alternate allele to allow comparison with the VarScan analysis output.

As a solution to this problem for each VarScan SNP position the original SAMtools base pileup column was associated with that position (from the SAMtools mpileup output file that is used as input for VarScan). This column details every read mapping to the position and its allele, reference or alternate, to allow identification of any number of alternate alleles so that one all-inclusive SNP call per position could accurately be made. This method utilizes the fact that each VarScan SNP location will be a position where mapping quality and number of reads mapping to that position for the identified SNP allele and reference allele are greater than the set threshold, in this case 20. Since we are processing only SNP positions that VarScan has identified, this is indicative of a region mapped well by high quality reads. The risk of including additional reads that are below recommended quality filters when using the pileup column directly to identify any additional alternate alleles in >10% of the sequencing reads was quantified. In 99% of cases this was determined to not be an issue and we were not including poor quality reads i.e. calculated depths of coverage for alleles using the pileup output directly were within 5% of those that were accurately made, from positions with one alternate allele, using VarScan's strict quality filters. This is not surprising as strict mapping parameters and filters should largely exclude poorly mapped reads prior to the SAMtools mpileup file generation.

To implement this methodology a bespoke Perl script (3rd_base_script.pl) was developed and used to extract, for every SNP position that was identified by VarScan; the position, reference base, an ambiguous base representing all alleles present, depth of coverage, % of sequencing reads with A, % of sequencing reads with C, % of sequencing reads with G, % of sequencing reads with T and the base pileup column. The reference base can be associated with the corresponding % of sequencing reads for that allele and any alternate alleles in more than 10% of the sequencing reads are easily identified. Using IUPAC nucleotide ambiguity codes, the appropriate ambiguous base is assigned to the position representing all of the alleles that are present. The output, which will be referred to as the "polyploid SNP list", was produced for each of the 4 varieties of wheat that were enriched with the genomic DNA array using their individual VarScan SNP calls.

The 4 polyploid SNP lists were compared using a bespoke Perl script (ID_varietal_SNPs.pl) that took the 4 files as input and extracted the ambiguous base representing all alleles present at each position. The output of this script known as the "varietal SNP list" consists of a line for every SNP position that was identified over the 4 input files detailing; the position, Truman ambiguous base, Rialto ambiguous base, Chinese Spring ambiguous base and Utmost ambiguous base. The ambiguous base was recorded as a "-" if a particular variety did not have a SNP at that position i.e. it has only the reference allele at this position. If one variety or more was unmapped at that position then the SNP position was discarded. There were 886,888 SNPs in the varietal SNP list (see file Four_varieties_SNPs_beta_array_final.txt).

The reference, although based on the variety Chinese Spring included no ambiguous bases despite the possibility of their presence in the genome. As such, it was quite possible for a SNP to be called in Chinese Spring sequencing data against the Chinese Spring based mapping reference. This however, was likely to happen less often than for the other varieties and explains the lower number of SNPs found in the Chinese Spring data. As a result the Chinese Spring variety sequencing data SNP calls were treated as the reference bases from this point in the analysis. If an ambiguous base was found in the Chinese Spring sequencing data this was to be the reference base or if a "-" was found in the Chinese Spring sequencing data then the original reference base call was supported and could be used. This allowed expansion of the reference to include all three genomes alleles to increase accuracy of the analysis.

This varietal SNP list output was then used to parse out varietal SNPs and SNPs conserved between varieties. Any position, at which one variety had a different base to the other 2 varieties and the new reference Chinese Spring allele, was classified as a varietal SNP. The other 2 varieties plus Chinese Spring could appear all as a "-", supporting the original reference base, or additionally the other 2 varieties could appear as the same ambiguous base as that of the Chinese Spring sequencing data. This methodology was used to identify Truman, Rialto and Utmost specific SNPs i.e. varietal SNPs.

A similar method was applied to identify SNPs that were conserved across the 3 varieties of wheat (Rialto, Truman and Utmost) but differed from the Chinese Spring sequencing data. In addition, SNPs that were conserved across any of the 2 varieties of wheat yet differing from Chinese Spring data were recorded. These SNPs can be used for downstream analysis of areas of SNP conservation or hot spots for differences between varieties. A summary of the SNPs that were identified is shown in figure 3.4.

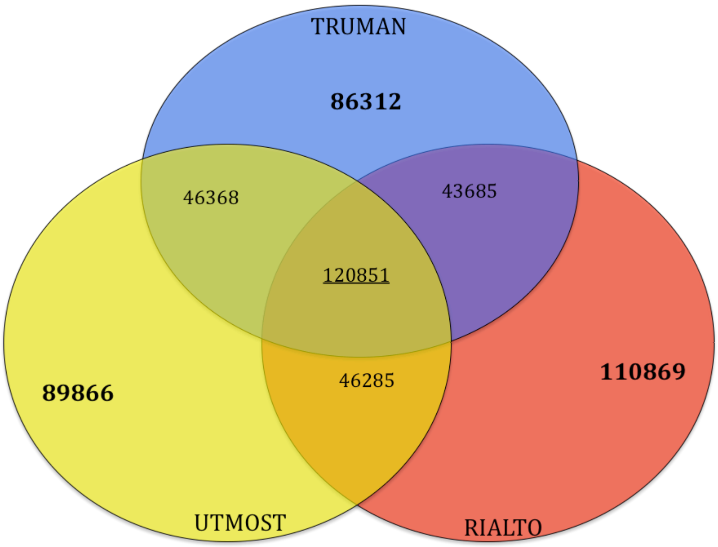


Figure 3.4. SNPs that have been identified in three varieties of wheat. Varietal SNPs are shown in bold type i.e. SNPs in one variety that differ from the other two varieties and the reference Chinese Spring. SNPs that are conserved in all 3 varieties but are different to Chinese Spring are underlined and SNPs that are conserved in each of the 2 varieties in turn that differ from Chinese Spring and the 3rd variety are shown in normal type.

Similar numbers of varietal SNPs were identified for each of the three varieties using Chinese Spring as a reference. Numbers were more strongly conserved across Utmost/Truman with neither variety deviating by more than ~2% from their average number of 88089 varietal SNPs. Rialto, however showed a frequency of varietal SNPs which was ~26% higher than the Truman/Utmost average number. The number of SNPs that are conserved in all three varieties but are different to Chinese Spring i.e. non-reference SNPs in all varieties was relatively high at 120,851. When looking at SNPs that are conserved in two varieties (that differ from Chinese Spring and the third variety) the numbers for Utmost/Rialto, Rialto/Truman and Utmost/Truman were highly similar with none deviating from the average of 45,446 by greater than ~4% i.e. no two varieties appeared to have more or less similarity to each other in comparison to the third variety and Chinese Spring.

Figure 3.5 shows the varietal SNP positions on the gene capture array design-space contigs that were translated into their respective positions on the pseudo-chromosome sequences. This image, created using the software Circos, allows, using heat maps, visualization of the distribution of each type of SNP per 80,000bp window along each pseudo-chromosome.

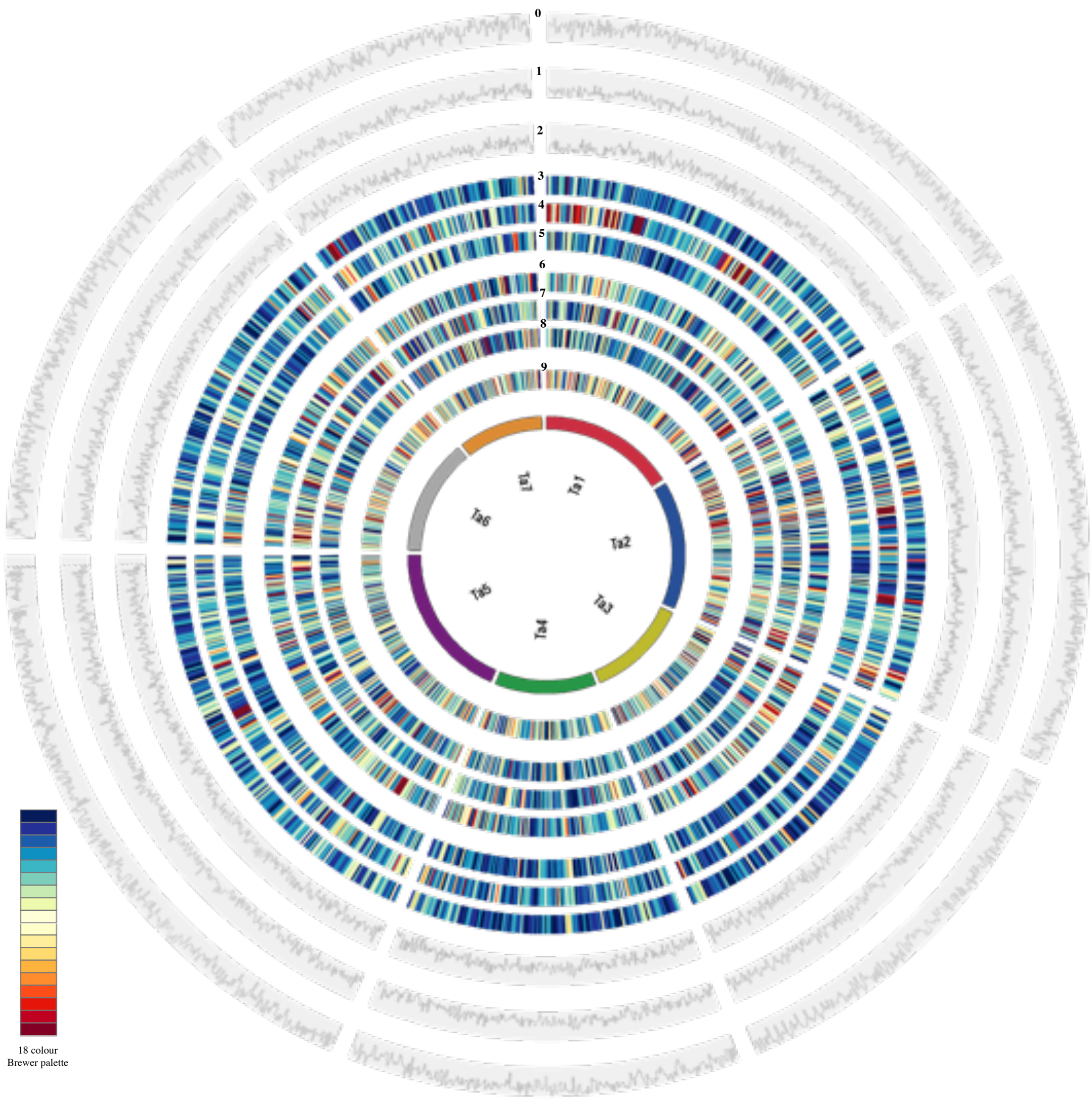


Figure 3.5. Circos plot outlining the three wheat varieties Rialto, Truman and Utmost. Average depth of coverage of; **(0)** Utmost, (**1)** Truman and **(2)** Rialto. Varietal SNP count for; **(3)** Truman, **(4)** Rialto and **(5)** Utmost. Count of SNPs conserved in; **(6)** Utmost & Truman, **(7)** Utmost & Rialto, **(8)** Rialto & Truman and **(9)** Utmost/Truman/Rialto. SNPs are called against the Chinese Spring reference dataset. SNP counts and average depths are calculated per 80,000bp window. The heat maps represent SNP numbers with an 18-colour Brewer palette (see figure). Minimum and maximum values are set as q5 (lowest) and q95 (highest) values and other percentiles of SNP counts are divided equally between the 16 remaining interim palette colours i.e. percentiles; 1-5, 6-11, 12-16, 17-22, 23-27, 28-33, 34-39, 40-44, 45-50, 51-55, 56-61, 62-67, 68-72, 73-78, 79-83, 84-89, 90-95 and 96-100 are assigned to each of the 18 interim Brewer palette colours from blue to red for comparatively low to high SNP numbers respectively. Any windows with numbers falling above or below these minimum and maximum values will be coloured deep red or deep blue respectively.

In figure 3.5 the q5/q95 lowest/highest values per 80,000bp over all Rialto, Utmost and Truman varietal SNPs (33/169) were used for these 3 tracks 3, 4 and 5. This allowed comparison between the varieties Rialto, Utmost and Truman over all 7 chromosomes. The q5/q95 lowest/highest values respectively per 80,000bp across the datasets; Rialto/Utmost conserved SNPs, Truman/Utmost conserved SNPs and Rialto/Truman conserved SNPs (13/68) were used for these 3 tracks 6, 7 and 8 to also allow comparison between the frequency of Rialto/Utmost conserved SNPs, Truman/Utmost conserved SNPs and Rialto/Truman conserved SNPs over all 7 chromosomes. Finally the q5/q95 lowest/highest values of SNPs in all 3 varieties (50/137) per 80,000bp were used for this plot in track 9. Tracks 0-2 are line plots of the average depth of coverage per 80,000bp window. Minimum and maximum values are the q5/q95 for each dataset independently.

The Circos plot in figure 3.5 is suited to visualization of larger trends in SNP distribution across all chromosomes i.e. genome-wise trends. The majority of the plots show relatively consistent trends; smaller regions of SNP density or scarcity are numerous, scattered consistently across the genome and difficult to identify in the large volume of data that is represented in the plot. We can however conclude from figure 3.5's larger trends that when looking at Rialto, Utmost and Truman varietal SNP plots (tracks 3-5) chromosomes 3 and 4 appear to be more SNP sparse. The Rialto varietal SNP plot (track 4) appears to have a SNP dense region, compared to the other chromosomes and other varieties, at the beginning of chromosome 1.

In figure 3.6 the same dataset is represented. However the q5/q95 lowest/highest values per 80,000bp window across the Rialto, Utmost and Truman varietal SNP lists were used on a chromosome-by-chromosome basis for tracks 3, 4 and 5. The limits were as follows;

chromosome 1 (32/285)
chromosome 2 (32/206)
chromosome 3 (32/285)
chromosome 4 (29/129)
chromosome 5 (33/141)
chromosome 6 (33/124)
chromosome 7 (32/143)

The q5/q95 lowest/highest values respectively between Rialto/Utmost conserved SNPs, Truman/Utmost conserved SNPs and Rialto/Truman conserved SNPs were again used on a chromosome-by-chromosome basis per 80,000bp window for tracks 6, 7 and 8. The limits were as follows;

chromosome 1 (13/60)
chromosome 2 (14/65)
chromosome 3 (13/79)
chromosome 5 (13/68)
chromosome 6 (15/89)
chromosome 7 (12/76)

chromosome 4 (12/57)

Finally the q5/q95 lowest/highest numbers of SNPs in all 3 varieties for each chromosome individually per 80,000bp window were used for track 9. The limits were as follows;

chromosome 1 (50/141)
chromosome 2 (49/143)
chromosome 3 (50/148)
chromosome 4 (49/114)
chromosome 5 (50/119)
chromosome 6 (56/137)
chromosome 7 (54/131)

This chromosome-wise analysis allowed clarity of the frequency of Rialto, Utmost and Truman varietal SNP plots within each of the 7 chromosomes individually. It also allowed comparison between the frequency of Rialto/Utmost conserved SNPs, Truman/Utmost conserved SNPs and Rialto/Truman conserved SNPs within each of the 7 chromosomes. Any within-chromosome trends that had been previously masked could be identified e.g. if in figure 3.5 one chromosome had a higher number of SNPs in general then it would look SNP dense (red) and make other chromosomes comparatively look SNP sparse (blue), if this chromosome was viewed alone regions of its own high and low SNP density may then become clearer.

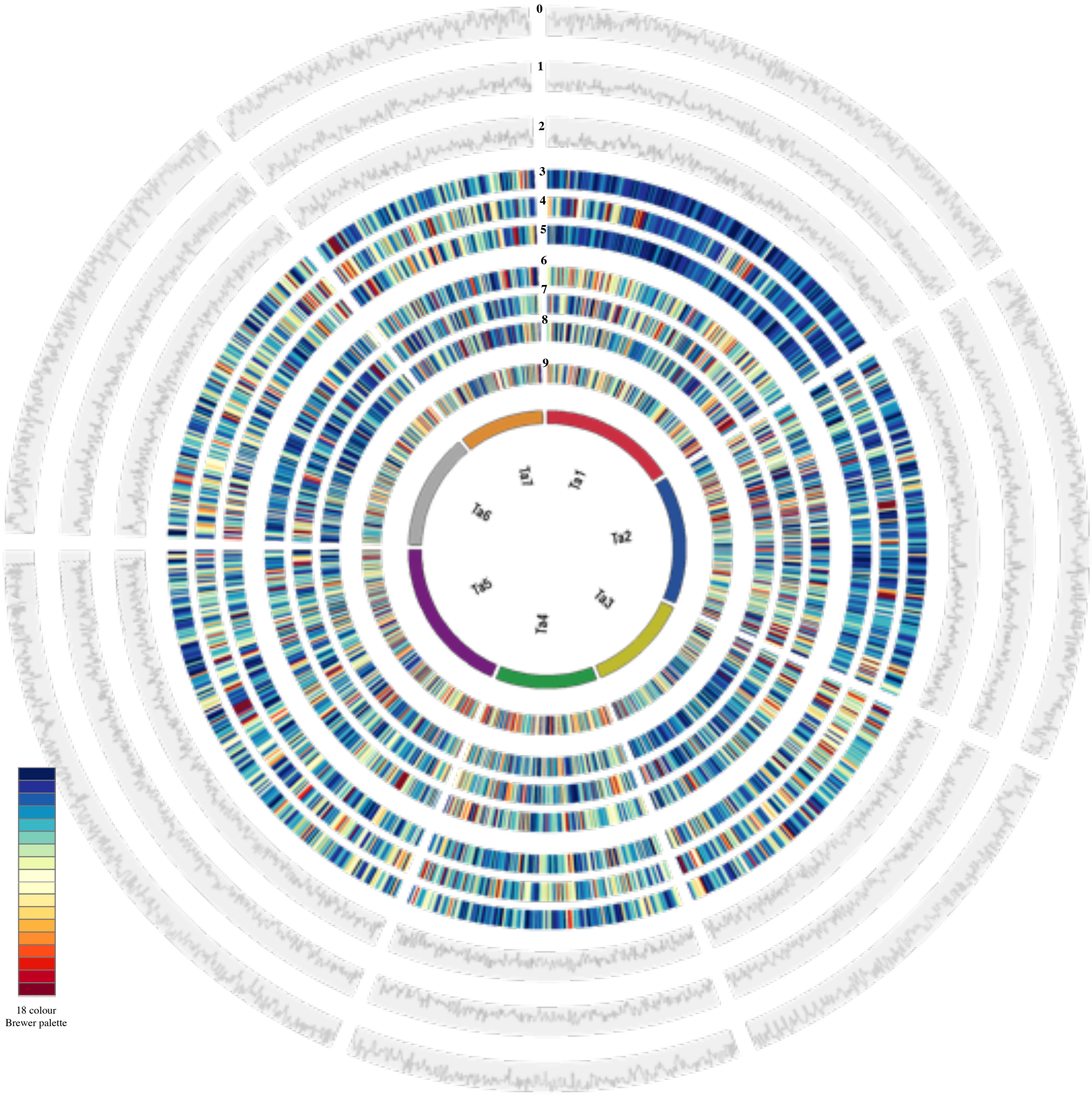


Figure 3.6. Circos plot outlining the three wheat varieties Rialto, Truman and Utmost. Average depth of coverage for; (**0**) Utmost, (**1**) Truman and **(2)** Rialto. Varietal SNP count for; **(3)** Truman, **(4)** Rialto and **(5)** Utmost. Count of SNPs conserved in; **(6)** Utmost/Truman, **(7)** Utmost/Rialto, **(8)** Rialto/Truman and **(9)** Utmost/Truman/Rialto. All SNPs are called against the Chinese Spring reference dataset. SNP counts and average depths are calculated per 80,000bp window along each pseudo-chromosome sequence. The heat maps represent SNP numbers with an 18-colour Brewer palette (see figure and full description in figure 3.5). Minimum and maximum values are set as q5/q95 lowest/highest values on a chromosome-by-chromosome basis as detailed previously.

Figure 3.6 allows visualization of larger trends in SNP distribution on a chromosome-by-chromosome basis i.e. trends within a within chromosome. Smaller regions of SNP density or scarcity appear to be numerous, scattered consistently within each chromosome over the majority of the plots and thus difficult to identify in the large volume of data that is represented in the plot. We can however conclude from figure 3.6's larger trends that when looking at Rialto, Utmost and Truman varietal SNP plots (tracks 3-5) chromosome 2 appears to be on the whole relatively SNP sparse with several small SNP dense regions rather than the consistent scatter of both dense and sparse regions that is seen in the other chromosomes. For these same varietal SNP plots (tracks 3-5) chromosome 1 appears to be very SNP sparse on the whole with little or no SNP dense regions in Truman and Utmost plots (tracks 3 and 5) but the same comparatively SNP dense region appears in the Rialto varietal SNP plot (track 4) at the beginning to middle of chromosome 1.

3.7 Conclusions

Synteny of the wheat gene capture array design-space contigs with *Brachypodium* allowed *Brachypodium*-wheat markers to be used to order and concatenate 68% of the contigs into 7 pseudo wheat chromosomes. These pseudo-chromosomes will assist in the application of chromosome dependent mutant identification methods/visualization tools to enriched wheat. Using the chromosome sorted IWGSC wheat sequence data over 80% of the array probes that were concatenated into the pseudo-chromosomes could be confirmed to have been associated with the correct chromosome (IWGSC, 2014).

In a comprehensive comparison of the quality of enrichment between the exome array and the gene capture array the following was observed over the non-enriched Rialto and enriched Rialto datasets after mapping analyses: the percentage coverage of the gene capture array reference and its average depth of coverage was always higher than that of the exome capture array; the enriched Rialto, as anticipated, has a much deeper coverage than the non-enriched Rialto on average across both of the arrays; for the exome capture array 11% of reference probe set were consistently unmapped with ~1% having a high depth of coverage (over 3 SD from the mean); this improved, and fell to <1% that were consistently unmapped and again <1% that had a high depth of coverage for the gene capture array. The gene capture array generates less off target sequence data and a greater proportion of the gene capture array bait probes appear to be enriching effectively in comparison to the exome capture array.

The results of a comparison of the homeologous SNPs identified in the enriched and non-enriched Rialto data for each array highlighted that both arrays enriched all three wheat genome's alleles efficiently. The key issue when working with a hexaploid is differential hybridization of homeologous SNPs and it appears at this stage that this will not be an issue with either array, as they are both, in the main, enriching data containing all of the anticipated homeologous SNP alleles. The gene capture array in particular does this with great accuracy as SNPs were largely conserved between enriched and non-enriched datasets in a minimum of 96% of cases in the comparative analyses (a minimum of ~86% of cases for the exome capture array). The small number of remaining SNPs that were unaccounted for tended to be low frequency alternate alleles, in areas of low coverage or low quality.

In a study of the array targets 77% of the exome array design-space contigs aligned to either *Brachypodium* genes or wheat cDNA as intended. This figure increased to 81% for the alignment of the gene capture array design-space contigs. The average percentage overall of the gene capture array probe design-space that was found to be exonic sequence was ~35%. As anticipated this figure increased greatly to ~64% for the exome capture array design-space. The gene capture array aligned to ~99% of the wheat cDNA sequence thus it effectively enriches the vast majority of wheat exonic sequence whilst still including non-coding sequence information, an improvement on the ~93% of the wheat cDNA that was hit by the exome array.

In almost all aspects of this study the gene capture array is confirmed to be an improvement upon the original exome capture array. It allows a higher quality of enrichment; higher reference coverage by mapped data, greater accuracy of SNP calls, the ability to pull out homeologous SNPs and a larger proportion of the intended target sequence hit. It will therefore be the enrichment array of choice for future analyses.

Four wheat varieties in total were successfully enriched using the genomic DNA array. A mapping and bespoke SNP calling pipeline was developed to map the data to the gene capture array design-space contigs and to identify a varietal SNP list in the context of a hexaploid i.e. multiple alternate alleles identified and represented by an appropriate ambiguous base call. Truman, Rialto and Utmost specific SNPs were identified against the reference dataset Chinese Spring. These varietal SNP numbers are loosely conserved across all three datasets and more strongly conserved across Utmost/Truman with Rialto having the most overall. A large number of SNPs are also present in all three varieties when compared to Chinese Spring. When looking at SNPs that are conserved in each of the two varieties in turn that differ from Chinese Spring and the third variety, no two varieties appear to have more or less similarity to each other in comparison to the third variety and Chinese Spring.

These conclusions were confirmed using the Circos plots (figure 3.5 and 3.6) where the varietal SNP positions were translated into pseudo-chromosome positions and their frequency summarized in plots per 80,000bp window. No notable (large scale conserved) differences were observed when comparing the tracks for SNPs that are conserved in each of the two varieties in turn that differ from Chinese Spring, as predicted. The majority of the plots showed relatively consistent trends in both plots; smaller regions of SNP density or scarcity are numerous, scattered consistently across the genome. A large number of SNPs were found to be present in all three varieties when compared to Chinese Spring and the consistent lighter to red colour (lack of dark blue in comparison to all other tracks) of the innermost track in both plots reflects this. When looking at Rialto, Utmost and Truman varietal SNP plots (tracks 3-5) the Rialto varietal SNP plot (track 4) appears to have a SNP dense region, compared to the other chromosomes and other varieties, at the beginning of chromosome 1. This trend was conserved in the within chromosome analysis where chromosome 1 appears to be very SNP sparse on the whole with little or no SNP dense regions in Truman and Utmost plots (tracks 3 and 5) but the same comparatively SNP dense region appears in the Rialto varietal SNP plot (track 4) at the beginning to middle of chromosome 1. This SNP dense region could account for the elevated numbers of varietal Rialto specific SNPs in comparison to Utmost and Truman.

Chapter 4. Mutant identification combined with target enrichment in wheat

Using the gene capture probe set for target enrichment followed by next generation sequencing; an early flowering locus was mapped in the diploid wheat *Triticum monococcum* and in hexaploid bread wheat *Triticum aestivum,* the stripe rust resistance gene was located. A bespoke pipeline and algorithm was developed for mutant loci identification and the pseudo-chromosome reference was implemented. This novel method will allow widespread application of sliding window mapping-by-sequencing analyses to datasets that are; enriched, lacking a finished reference genome or polyploid.

4.1 Introduction

In chapter 2 a bespoke mapping, SNP calling and allele frequency analysis pipeline (BWA, SAMtools and VarScan) was successfully implemented to identify regions that were potentially harbouring a causal SNP in various diploid *Arabidopsis* mutant plants. Generation of artificial sequencing datasets that were created using the *Arabidopsis* Col-0 reference genome allowed a successful simulation of mutant identification for a diploid, tetraploid and finally a hexaploid mutant with a full reference genome using a sliding window mapping-by-sequencing analysis.

Mapping-by-sequencing analyses typically require the generation of shotgun sequence for a scored F2 mapping population. In chapter 3 a NimbleGen gene capture array in solution (120Mb) was introduced and validated. Utilization of this enrichment array allows effective enrichment and subsequent sequencing of the gene rich regions of hexaploid wheat. This allows high coverage to be generated for target regions by eliminating much of the repetitive sequence from the analysis and greatly reduces the need for whole genome re-sequencing and the great cost associated with it. Mapping-by-sequencing analysis also benefits from a reference genome for mapping of the sequence data. For wheat, like many crop species, no finished genome reference sequence is available and as such the gene capture array design-space contigs were arranged into a long-range order on the finished *Brachypodium* genome and then using synteny between *Brachypodium* and wheat, they could be ordered and concatenated into wheat based pseudo-chromosome sequences that are used here directly as a reference genome for sliding window analyses (section 3.3). This extends a proof of principle approach where *Arabidopsis* cDNA sequence was assembled into *Brassica rapa* based pseudo-chromosomes using synteny between the two. In a mapping-by-sequencing analysis two mutant intervals were identified as positions in *B. rapa* using allele frequency estimates at 4375 marker positions in an enriched subset of *Arabidopsis.* These *B. rapa* intervals were later translated back to a single *Arabidopsis* position (Galvão *et al.*, 2012).

This adds a new level of complexity to the overall aim here; to identify a mutant region in hexaploid wheat using enriched sequencing data i.e. a fragmented reference genome. Since the intention here is to ultimately apply this methodology to polyploid wheat this necessitates the development of a novel pipeline that prioritizes adaptability to polyploid species since current tools such as SHOREmap are tailored to diploid species. This analysis required further development of the current bespoke pipeline and mutant identification algorithm (chapter 2) to firstly map an early flowering locus, a deletion, in the gene enriched genome sequence of the diploid einkorn wheat mutant *Eps-3A^{m}* (*Triticum monococcum* L.) (section 4.2). *Eps-3A^{m}*, is an early flowering mutant with an altered circadian clock phenotype, with previous mapping suggesting that the phenotype is due to the deletion of a circadian clock gene, an ortholog of the Arabidopsis circadian clock gene *LUX ARRHYTHMO/ PHYTOCLOCK 1 (LUX),* that is thought to play an important role in the evening complex within the circadian clock (Campoli *et al*., 2013; Mizuno *et al*., 2012; Gawroński *et al.,* 2014; Hazen *et al.,* 2005). This mutant dataset was developed and phenotyped by Gawroński *et al.* and sequenced at the CGR (Gawroński *et al.,* 2014).

Bread wheat is made up of the A, B, and D genomes. Based on Acc-1 gene evolution the bread wheat A genome donor *T. urartu* and *T. monococcum* were estimated to have diverged only 0.5 to 1 million years ago (Huang et al., 2002). As such here the genome of *T. monococcum* has been used successfully as a model for the A genome of hexaploid wheat (Wicker *et al.,* 2003). Although originally designed based on hexaploid wheat the gene capture array was shown here to effectively enrich a divergent diploid wheat with high synteny to one of the three wheat genomes. This analysis allows a mutant identification trial in an enriched diploid dataset before adding the extra complication of a hexaploid genome.

This mutant identification analysis involves the use of parental Recombinant Inbred Lines (RILs). These are immortal populations in which recombinant chromosomes have been fixed through inbreeding. RIL lines are advantageous, as they only have to be genotyped once; they are mainly homozygous and the remaining small percentage of the genome that is heterozygous ensures that only a small portion of the genome segregates for the two parental alleles (Weigel, 2012). Figure 4.1 demonstrates an example of the construction of a RIL line in which a parental cross produces F1 offspring that are intercrossed to produce F2 offspring. Random F2 crosses are performed and then random pair matings of offspring (two from each cross) in each generation for multiple generations (inbreeding) (Pollard, 2012). Mapping-by-sequencing relies on a local skewing of allelic frequency close to the site of the loci responsible for the mutant phenotype when you bulk therefore genotype-by-sequencing a phenotypically scored F2 mapping population. Such a bulk segregant F2 mapping population is also employed within this analysis.

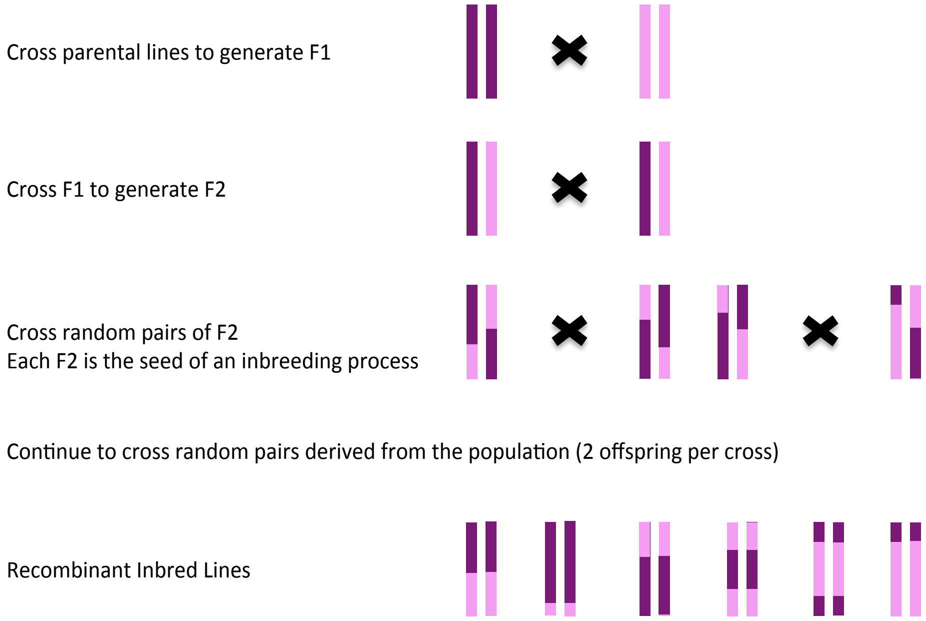

Cross parental lines to generate F1
Cross F1 to generate F2
Cross random pairs of F2
Each F2 is the seed of an inbreeding process
Continue to cross random pairs derived from the population (2 offspring per cross)
Recombinant Inbred Lines

Figure 4.1. Construction of a RIL line. Example of the construction of a Recombinant inbred line (RIL) employed by Pollard where a set of diploid chromosomes represents an individual and each parental genotype is represented by either pink or purple.

In section 4.3 a further slight adjustment of the algorithm that was developed in section 4.2 was made to enable identification of gene regions that are associated with stripe rust resistance or susceptibility in two pools of enriched mutant hexaploid wheat plants. These pools were developed and phenotyped by a group at Reading University led by Donal O'Sullivan and sequenced at the CGR in Liverpool. The main objective was to re-map stripe rust resistance genes in the parental lines and in doing so, increase the density of polymorphisms associated with the intervals. Stripe rust or yellow rust, caused by *Puccinia striiformis*, is one of the most important diseases of bread wheat with epidemics often leading to severe wheat yield losses. The most efficient methods of combating stripe rust disease involve the utilization of resistant cultivars in affected regions. Frequent emergence of novel stripe rust races results in resistant wheat cultivars often becoming susceptible after being grown for some periods of time (Wellings *et al.*, 1990; Chen *et al.*, 2002). As a result, the search for new stripe rust-resistance genes and breeding of new resistant wheat varieties is carried out on a continued basis (Deng *et al.*, 2004).

The analysis in section 4.3 involved the use of wheat lines that were bred specifically by Donal O'Sullivan's group to be double haploid. Such lines are created, in this case, when two parental purebred lines of the wheat varieties Avalon and Cadenza are crossed to produce F1 progeny. The F1 progeny haploid cells are then allowed to undergo chromosome doubling to produce offspring that are a mosaic of the parental lines with both chromosomes largely identical i.e. widespread homozygosity between chromosome pairs within each of the 3 diploid wheat genomes. Here two genotypes were to be studied after bulk segregant analysis of F1 offspring; those with the Yr7 stripe rust resistance locus originating from the Cadenza parent and those with Yr7 stripe rust susceptibility as per the Avalon parent.

Here, target enrichment in wheat, combined with a sliding window mapping-by-sequencing mutant identification approach, is demonstrated using a pseudo genome reference, derived from wheat-*Brachypodium* synteny in; a diploid wheat, mapping the *Eps-3A*^{m} mutation to a small deletion on chromosome 3 in *T. monococcum* and in a hexaploid wheat to enable identification of regions that are associated with stripe rust resistance/susceptibility.

4.2 Mutant identification in the diploid wheat *T. monococcum*

4.2.1 Sample preparation and mapping

An early flowering *T. monococcum* mutant KT3-5 was crossed with a wild accession KT1-1 of *Triticum boeoticum*; The F1 progeny self-pollinated and 1-2 seeds per plant were grown. This was repeated for ten or eleven generations to obtain the RILs. To eliminate another QTL for flowering time on chromosome 5 markers linked to the eps5 and Vrn2 were used and RIL25 (early flowering) and RIL71 (wild type) lines were selected. RIL25 and RIL71 were then used as the parental lines for this analysis and crossed accordingly to produce an F2. This F2 mapping population was phenotypically classified and bulk segregated into two groups; wild type (parent RIL71) phenotype named Bulk A and early flowering (parent RIL25) phenotype named Bulk B. Each bulk contained approximately 250 individual plants (Gawroński *et al.,* 2014). The full details of the breeding and phenotyping of the lines that were analyzed here is as previously reported (Gawroński *et al.*, 2014).

The two bulk segregated populations, along with the parental RIL25 and RIL71 lines were enriched using the gene capture array and sequenced at the University of Liverpool's CGR as follows; Genomic DNA was purified using Agencourt AMPure XP beads (Beckman Coulter). Samples were quantified using a Qubit double-stranded DNA Broad Range Assay Kit and Qubit fluorometer (Life Technologies). Genomic DNA was sheared for 3×60s using a Covaris S2 focused-ultrasonicator and the size distribution of the fragmented DNA was assessed on a Bioanalyzer High Sensitivity DNA chip (Agilent). End-repair, 3′-adenylation, and adapter ligation were performed according to the Illumina TruSeq DNA Sample Preparation Guide (Revision B, April 2012) without in-line control DNA and without size-selection. Amplification of adapter-ligated DNA (to generate pre-capture libraries), hybridization to custom wheat NimbleGen sequence capture probes, and washing, recovery and amplification of captured DNA were all carried out according to the NimbleGen Illumina Optimized Plant Sequence Capture User's Guide (version 2, March 2012), with the exception that purification steps were carried out using Agencourt AMPure XP beads instead of spin columns. Final libraries were quantified by Qubit double-stranded DNA High Sensitivity Assay and the size distribution ascertained on a Bioanalyzer High Sensitivity DNA chip. The 4 libraries were then pooled in equimolar amounts based on the aforementioned Qubit and Bioanalyzer data. Sequencing was carried out on two lanes of an Illumina HiSeq 2000, using version 3 chemistry, generating 2 x 100bp paired end reads.

As seen in figure 4.2a sequence datasets for all 4 samples were mapped to the pseudo-chromosome sequences, which were generated from the gene capture array design-space, using BWA (v 0.6.2) fragment short read mapping. Indexing of the reference sequence implemented the 'IS' algorithm and during alignment of reads to the reference 4 mismatches were allowed per sequencing read. All unmapped, non-uniquely mapped and duplicate reads were later removed using SAMtools. Finally SAMtools mpileup (v 0.1.18) (Li, H. *et al.*, 2009) was implemented on the 4 datasets and SNP calls were filtered out using VarScan (VarScan.v2.2.11.jar) (Koboldt *et al.,* 2012) with the following parameters: discard SNPs covered by 10 or fewer reads, discard sequencing reads with a quality less than 15 and if the alternate allele has less than 2 supporting reads passing the quality filter discard it. For this SNP analysis the tool awk was implemented to remove indels from the VarScan output. The analysis steps are shown in figure 1.6 and in the command outline appendix sections 1 and 4.

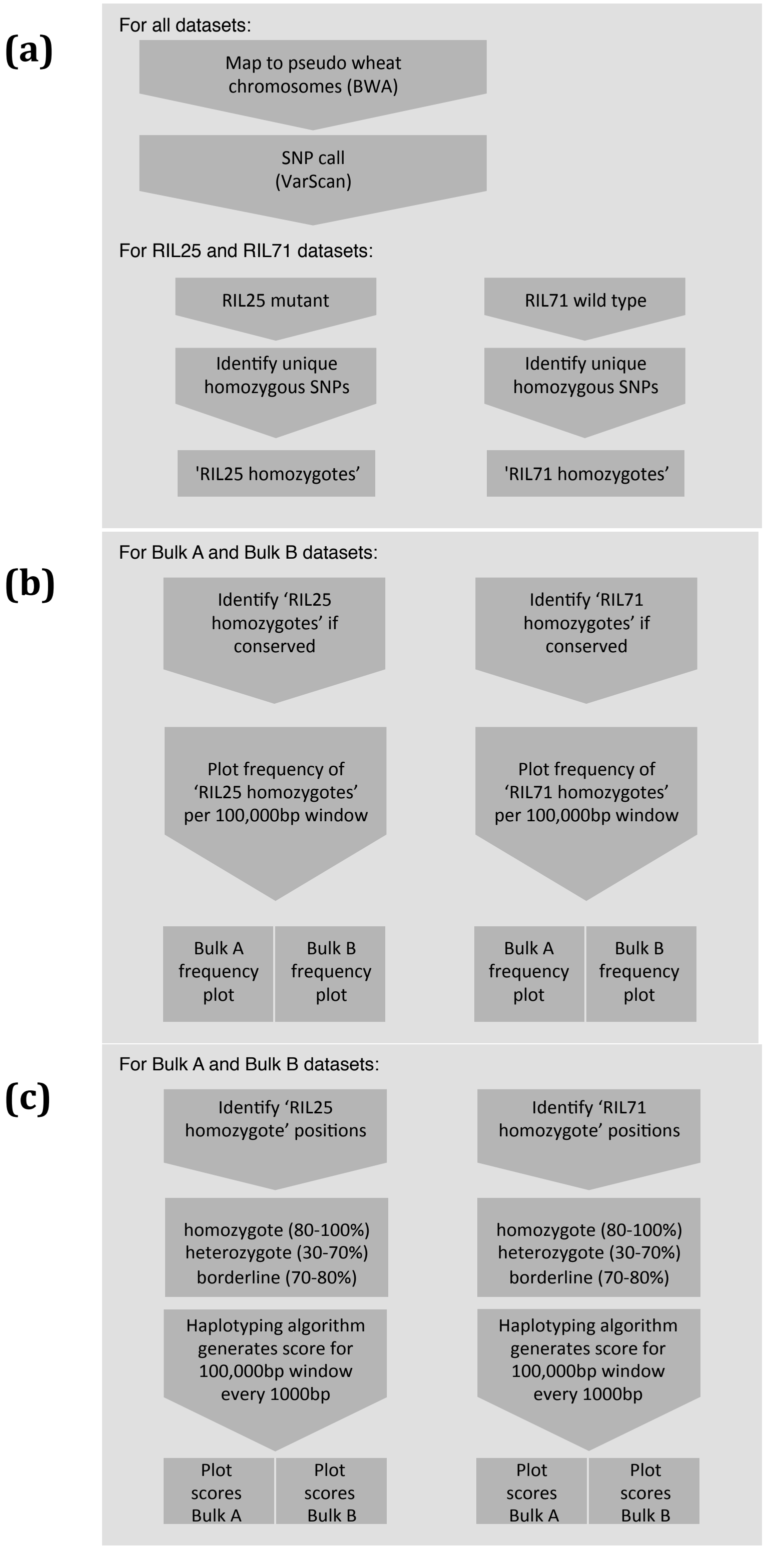

(a)
For all datasets:
Map to pseudo wheat chromosomes (BWA)
SNP call (VarScan)
For RIL25 and RIL71 datasets:
RIL25 mutant
RIL71 wild type
Identify unique homozygous SNPs
Identify unique homozygous SNPs
'RIL25 homozygotes'
'RIL71 homozygotes'
(b)
For Bulk A and Bulk B datasets:
Identify 'RIL25 homozygotes' if conserved
Identify 'RIL71 homozygotes' if conserved
Plot frequency of 'RIL25 homozygotes' per 100,000bp window
Plot frequency of 'RIL71 homozygotes' per 100,000bp window
Bulk A frequency plot
Bulk B frequency plot
Bulk A frequency plot
Bulk B frequency plot
(c)
For Bulk A and Bulk B datasets:
Identify 'RIL25 homozygote' positions
Identify 'RIL71 homozygote' positions
homozygote (80-100%) heterozygote (30-70%) borderline (70-80%)
homozygote (80-100%) heterozygote (30-70%) borderline (70-80%)
Haplotyping algorithm generates score for 100,000bp window every 1000bp
Haplotyping algorithm generates score for 100,000bp window every 1000bp
Plot scores Bulk A
Plot scores Bulk B
Plot scores Bulk A
Plot scores Bulk B

Figure 4.2. Processing 4 sets of enriched sequencing data to identify a mapping interval containing the deletion that is inducing the phenotype of interest. **(a)** Standard mapping and SNP calling pipeline **(b)** Initial homozygote allele frequency determination method for Bulk A and Bulk B samples **(c)** Final allele frequency algorithm or Bulk A and Bulk B samples to identify the interval of interest

Initially low mapping coverage was seen across the 4 datasets (54% of reference mapped to on average at a depth of ~59x). This was anticipated due to mapping the diploid wheat *T. monococcum,* that has since diverged from the hexaploid wheat A genome donor, to a reference sequence that was designed to represent any/all of the 3 genomes of the hexaploid wheat Chinese Spring. Further optimization of this mapping analysis enabled identification of parameters that allowed increased mapping of this divergent dataset to the reference sequence; firstly the mapping seed by default has 2 mismatches within it that was increased to 3, the seed length was reduced from 32 to 30 and the quality threshold for trimming reads was set at 20. Due to local re-arrangements in sequence between the diverged species *T. monococcum* and hexaploid wheat the 100bp raw sequencing reads were split into two 50bp reads and mapped separately. The 4 datasets were re-mapped using these parameters and SNP calling was repeated. Coverage was highly conserved between the 4 samples and increased with on average 70% of the reference mapped to at ~70x coverage (table 4.1).

Sample	**Average % coverage of pseudo-chromosome base space**	**Average depth of coverage**	**Number of homozygous SNPs identified**	**Number of heterozygous SNPs identified**
RIL25	69.8	63.8	978,511	118,330
RIL71	69.7	72.4	1,013,269	119,883
Bulk A	69.8	69.4	159,822	143,147
Bulk B	69.0	67.7	188,363	155,570

Table 4.1. Mapping Statistics for the 4 enriched wheat DNA samples in relation to the pseudo-chromosome reference sequence. Statistics calculated across the pseudo-chromosome assembly base space. Homozygous SNPs in 80-100% of sequencing reads and heterozygotes in 30-70% sequencing reads.

In RIL25 and RIL71 mapping extended across 108,218 and 108,263 of the design-space contigs that made up the pseudo-chromosomes respectively. Of the 7,031 and 6,986 unmapped target sequence contigs in RIL25 and RIL71, the majority, 6,072 were unmapped in both samples. In a related study mapping the hexaploid wheat Chinese Spring to the pseudo-chromosome sequences 98% of the target sequence was mapped. This encompassed

114,981 design-space contigs with only 268 unmapped, 186 of these were also unmapped by the two RIL *T. monococcum* lines.

Figure 4.2a also details the identification of RIL25 and RIL71 specific homozygous SNPs. Homozygous SNPs were classified as SNPs in 80% or more of the sequencing reads and 881,860 homozygous SNPs were identified in relation to the Chinese Spring reference that were shared between the two RIL lines. These were removed leaving 96,651 RIL25 and 131,409 RIL71 specific homozygous SNPs. The RIL specific SNPs could then be used to generate a RIL25 and RIL71 'reference genome'.

4.2.2 Initial homozygote allele frequency determination for mutant identification

Two mapping-by-sequencing mutant identification pipelines were developed from the bespoke mutant identification pipeline in chapter 2 and were successfully used for this analysis (Figure 4.2b and 4.2c). The first pipeline was used to identify regions with increased homozygous frequency compared to the parent genome and was closely based on the methodologies that were demonstrated by SHOREmap. Of the homozygote SNPs that were specific to the RIL25 parent 34,125 SNPs could be found as conserved homozygous alleles in the Bulk A data and 46,154 in the Bulk B data i.e. also in 80-100% of the sequencing reads. Of the homozygote SNPs that were specific to the RIL71 parent 54,376 were conserved as homozygotes in the Bulk A data and 64,700 in the Bulk B data.

Frequencies of the RIL71 and RIL25 SNPs were calculated per 100,000bp window along each chromosome for Bulk A and Bulk B datasets and displayed graphically. To minimalize noise the data for all RIL25 specific SNPs could be combined with the following calculation; BulkB homozygote frequency per interval – BulkA homozygote frequency for the same interval, highlighting shared homozygosity between Bulk B and the mutant RIL25 parent. A clear peak of conserved RIL25 homozygous SNP frequency was observed at the end of chromosome 3 in Bulk B (Figure 4.3a). The same data was displayed for the RIL71 specific homozygous SNPs with the opposite calculation; BulkA homozygote frequency per interval – BulkB homozygote frequency per corresponding interval, highlighting shared homozygosity between Bulk A and RIL71. The same clear peak was observed at the end of chromosome 3 in Bulk A (Figure 4.3b). All raw data for figure 4.3a and 4.3b can be found in Appendix 3, table 1 and 2 respectively. The interval that the peak occurred within translated to the window 10,000,000bp-10,600,000bp of the pseudo wheat chromosome 3. There were 748 probes concatenated in order to form this region and these probes aligned to the region 58,063,918-59,004,348 in *Brachypodium* chromosome 2 (~940Kbp) including the genes Bradi2g60780-62310 (~160 genes).

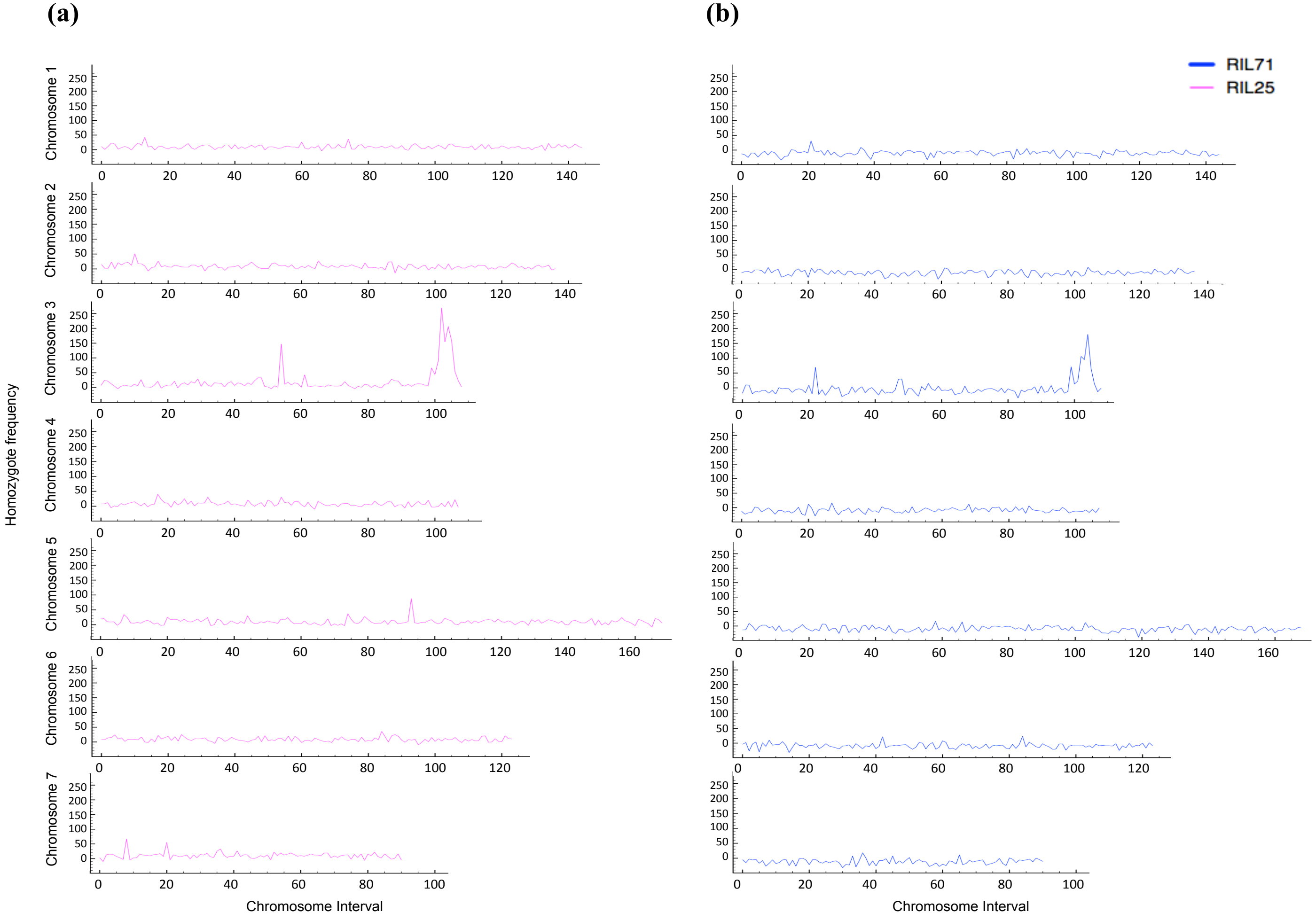


Figure 4.3. Frequencies of Bulk A and Bulk B homozygotes calculated along each pseudo-chromosome. (a) Frequency of 'RIL25 homozygous' SNPs per window; Bulk B frequency minus Bulk A frequency per 100,000bp window **(b)** Frequency of 'RIL71 homozygous' SNPs; Bulk A frequency minus Bulk B frequency per 100,000bp window.

4.2.3 Final haplotyping algorithm for mutant identification

Figure 4.2c details the final improved algorithm for analysis that has been developed with downstream implementation on hexaploid wheat in mind. This method scores regions of interest by prioritizing long homozygous parental haplotypes, the longer the length and the more homozygous the region, the higher the score generated. Scores are calculated per user-defined window in both analyses and in the final algorithm an additional 1000bp window overlap was included. To enable this analysis again the RIL25 and RIL71 specific homozygote lists were implemented. However rather than looking for conserved homozygote positions in the Bulk A and Bulk B datasets the SNP alleles were simply identified in the Bulk A and Bulk B datasets regardless of homozygous or heterozygous status.

Of the homozygote SNPs that were specific to the RIL25 parent, 49,126 of the SNP alleles were conserved in the Bulk A data regardless of homozygous or heterozygous status, and 62,388 in the Bulk B data. Similarly of the homozygote SNPs that were specific to the RIL71 parent, 54,377 of these were found in the Bulk A data and 83,401 in the Bulk B data. These SNPs were categorized into homozygous, heterozygous or borderline (see Figure 4.2c for categorizing limits) and a scoring system was developed to calculate a homozygote score per 100,000bp window along the pseudo-chromosomes at 1000bp intervals (Figure 4.4).

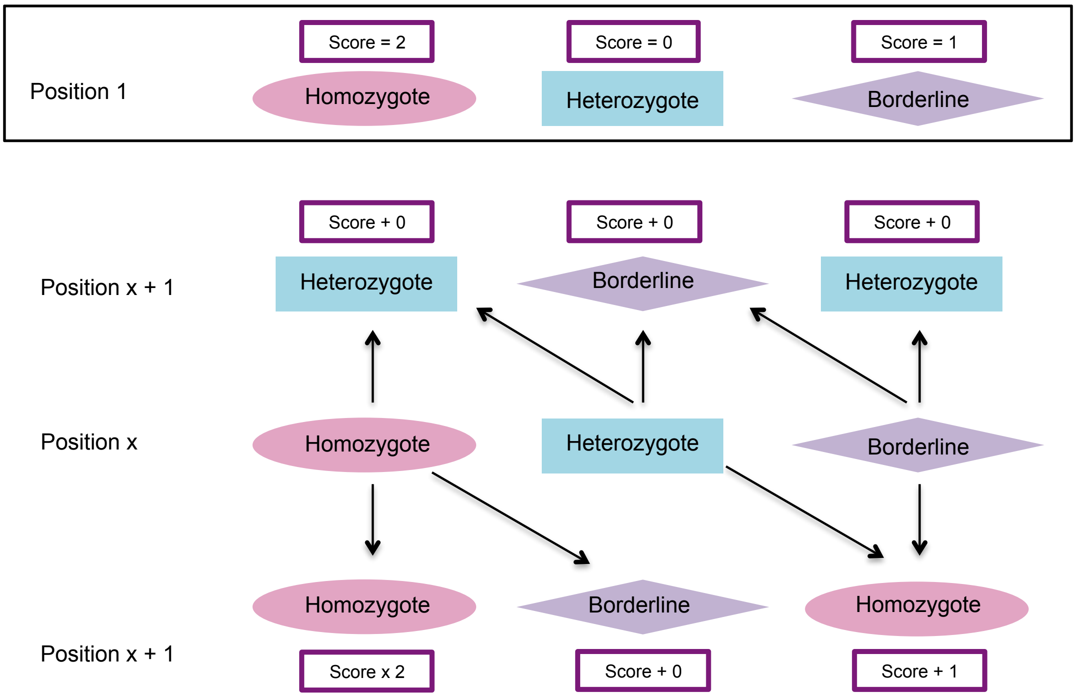


Figure 4.4. Outline of mutant identification algorithm. A unique scoring system used to calculate a homozygote score per user-defined window (100,000bp used) along the pseudo-chromosomes at 1000bp intervals. Position 1 defines the first SNP position in the 100,000bp window for which a score of 0, 1 or 2 is determined. For SNP position 2 the current score is amended taking into account the position 1 SNP call (position x) in relation to the position 2 SNP call (position x + 1). Further iterations ensue for every SNP position i.e. position x will become position 2, 3, 4 up to the end of the 100,000bp window and the score is amended each time until a final value is recorded. The window for analysis is then re-set 1000bp downstream and the score re-set to 0 before the analysis will be repeated.

The scores that were defined using the new algorithm for the Bulk A and Bulk B datasets in relation to both parents were plotted in figure 4.5 respectively and magnification of the peak regions confirm the same peak has been defined that was seen in the previous analysis. This method generates a much greater volume of analyzable data that its predecessor calculating a score every 1000bp rather than every 100,000bp. It amplifies the interval of interest at the end of chromosome 3 in comparison to the initial method whilst reducing background noise. The additional step of subtracting the Bulk A data from Bulk B and vice versa is no longer necessary to reduce noise.

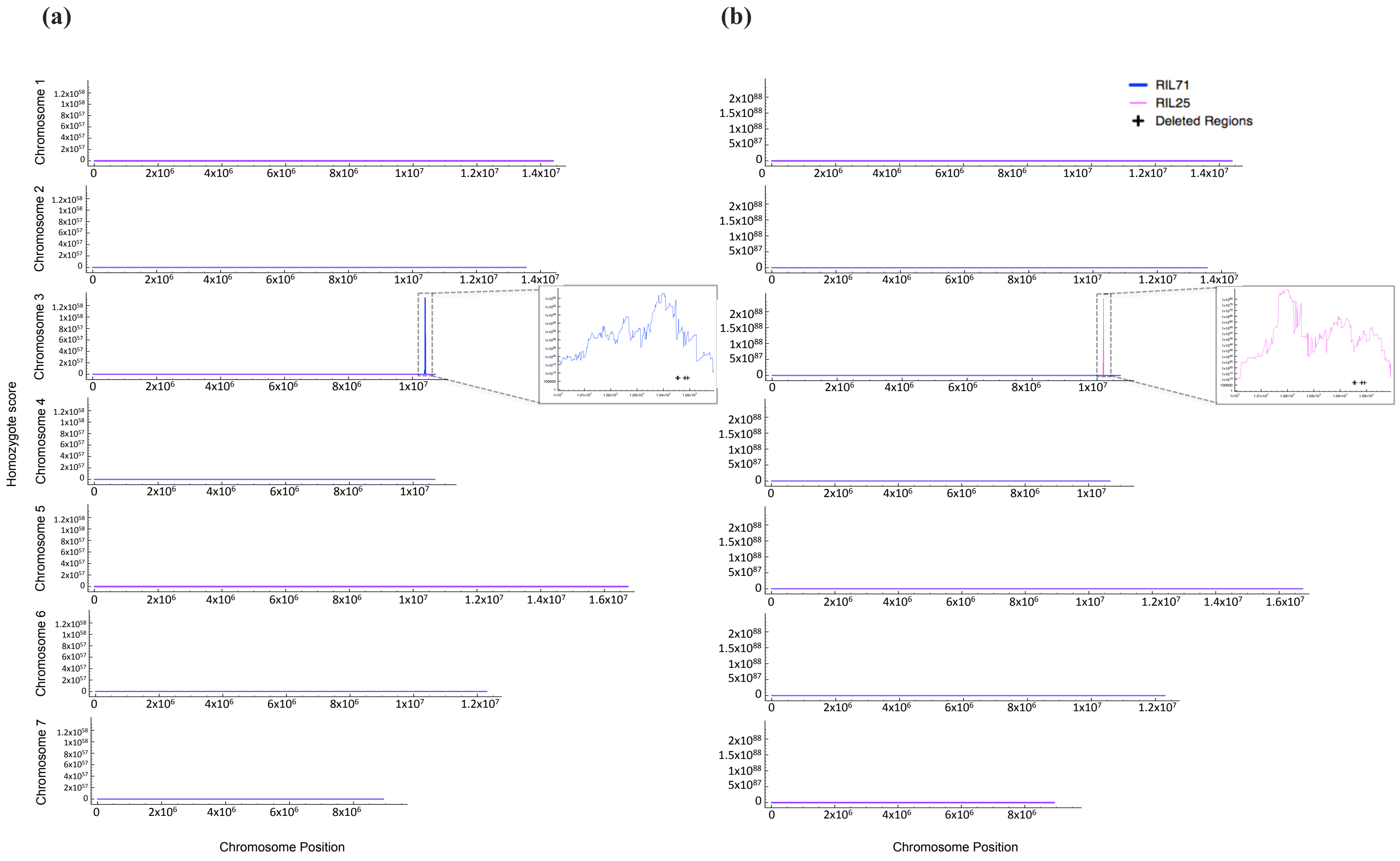


Figure 4.5. Homozygosity scores calculated for Bulk A and Bulk B datasets. Scores plotted along each pseudo-chromosome. Haplotypes conserved with the RIL25 (magenta line) and RIL71 (blue line) parental unique homozygote SNPs. Scores calculated per 100,000bp window and calculated at 1000bp intervals. **(a)** Scores for Bulk A dataset **(b)** Scores for the Bulk B dataset. Peak interval magnified and regions harbouring a deletion > 100bp are highlighted.

4.2.4 Conclusions from mutant identification in *T. monococcum*

The enrichment approach not only allows identification of SNPs it also allows the identification of copy number variation and deletions. Therefore, the "reference genome" was scanned for deletions. To define a deleted region that was potentially inducing the mutant phenotype it required mappeds reads in the wild type RIL71 and/or the Bulk A datasets, whilst being unmapped in the mutant RIL25 and the Bulk B datasets. Deleted regions were defined that were longer than 100bp, and that fitted this mapping expectation. Using this approach 163 deleted regions were identified (Appendix 3, table 3), 11 were within the interval that was identified at the end of chromosome 3.

Although the enrichment array is predicted to contain the majority of wheat's genic sequence, as only ~70% of the reference sequence has been mapped to it is possible that the causal deletion could be partially excluded from the analysis. As such, we may not see a full segment deletion but a concentrated region of smaller deleted regions using this approach. The 163 deletions were further filtered for regions where two or more deletions could be seen within a 1000bp window, resulting in 18 deleted regions that are detailed in table 4.2, the majority of which lie underneath the defined interval of interest on pseudo-chromosome 3.

Deletions within the peak interval of interest are in bold type in table 4.2 and they are also plotted in figure 4.5. This highlights sequence covering ~40Kbp and particularly a wheat gene region (10,481,196-10,482,558bp) that previously was found to align to the *Brachypodium* gene Bradi2g62067 with similarity to the *Arabidopsis* LUX gene. In a BLASTN alignment of this wheat gene region to the BLAST nr nucleotide database (Altschul *et al.,* 1990), strong similarity was seen to the *T. aestivum* cultivar Chinese Spring LUX gene (e value < 1e-5, length ~859bp and sequence identity 99%). The LUX gene is known to affect both the circadian clock and flowering time in *Arabidopsis* (Hazen *et al.,* 2005) and therefore is a strong candidate for the mutation responsible for the *Eps-3A^{m}* mutation in *T. monococcum* (Gawroński *et al.,* 2014).

Chrom	Position	Length of hit	Associated gene	Function
3	2205655	121	Bradi2g50140	1,3-beta-D-glucan synthase activity
3	2205890	441	Bradi2g50140	1,3-beta-D-glucan synthase activity
3	2206382	163	Bradi2g50140	1,3-beta-D-glucan synthase activity
3	**10452241**	**129**	**Bradi2g61960**	**DEAD box ATP dependent RNA helicase activity**
3	**10453160**	**105**	**Bradi2g61960**	**DEAD box ATP dependent RNA helicase activity**
3	**10456240**	**313**	**Bradi2g61960**	**DEAD box ATP dependent RNA helicase activity**
3	**10456774**	**103**	**Bradi2g61960**	**DEAD box ATP dependent RNA helicase activity**
3	**10481196**	**131**	**Bradi2g62067**	**Similar to LUX gene G2-like (Myb-like domain)**
3	**10481946**	**271**	**Bradi2g62067**	**Similar to LUX gene G2-like (Myb-like domain)**
3	**10482446**	**112**	**Bradi2g62067**	**Similar to LUX gene G2-like (Myb-like domain)**
3	**10491979**	**221**	**Bradi2g62093**	**(Upstream of) gene contains F-box domain**
3	**10492675**	**179**	**Bradi2g62093**	**(Upstream of) gene contains F-box domain**
4	2761361	195	Bradi1g69930	Putative digalactosyldiacylglycerol synthase
4	2761813	182	Bradi1g69930	Putative digalactosyldiacylglycerol synthase
5	577462	132	Bradi2g39240	RNA binding
5	577604	239	Bradi2g39240	RNA binding
5	4042813	110	Bradi4g04880	Protein binding
5	4042987	102	Bradi4g04880	Protein binding

Table 4.2. Detailing the pseudo-chromosome regions that harbour potential deletions. Deletions are defined as regions longer than 100bp mapped to by the RIL71 data and Bulk A data and also unmapped by the RIL25 data and the Bulk B data. Only when 2 or more deleted segments are found within a 1000bp window are they included. Regions associated with the candidate gene are underlined and regions within our peak interval of interest are in bold type.

With no prior knowledge regarding the deleted region of interest this method would still be able to reduce the number of deletions that would be taken forward for further analysis to a small manageable number. Particularly as any analysis would focus on the regions underneath the defined interval only. This region encompasses ~40Kbp and 9 identified potential deleted regions that could point towards the longer deletion within the region.

The mutant identification analysis that was carried out in chapter 2 involved calculation of a homozygote to heterozygote ratio in a bulk segregant mutant pool against a wild type reference sequence. Here, it did not identify an interval of interest. Instead analyses within this chapter adapted these principles and used algorithms based on conserved homozygote frequency of the bulk segregant mutant pool with the mutant parental line to locate an interval. Prioritizing conserved homozygosity between the mutant parent and bulk segregant mutant F2 pool; excludes most noise, utilizes only those homozygotes that have in effect had a double validation-appearing in two independent datasets and still retains and highlights the interval of interest. This is a general improvement on the methodology used in chapter 2 that is likely to decrease noise and identify an interval even in a divergent dataset. It is of particular use if there is no defined reference sequence for the wild type parental line that is used and therefore the final homozygote haplotyping algorithm that was developed here will be the analysis of choice for further mutant identification analysis within this project.

Both of the pipelines featured in figure 4.2b and 4.2c are attached as user-friendly Perl scripts allowing command line input of required files and various parameters to allow user definition of windows for analysis and ploidy level of the organism under analysis. They output data files of scores per user-defined window along with ready-made raw plots of these values per chromosome using R. The original raw algorithm detailed in figure 4.2b (Homozygote_frequency_plus_plot.pl) requires simply a list of homozygote SNPs that are conserved between the mutant parent and mutant bulk segregant offspring; a tab separated file [chromosome <tab> position] and filtered VarScan and GATK outputs are therefore acceptable. A tab separated file specifying a line for each chromosome [chromosome <tab> length] is also required. Otherwise the user defines the window size for analysis (typically 200,000bp) and output file prefix. The resultant output has been updated to be a text file containing [Chromosome <tab> Start position of interval on chromosome <tab> homozygote number] rather than the original [Chromosome <tab> Window number <tab> homozygote number] output that was used in figure 4.3. A pdf file is also generated of the raw frequency plot in R using this text file. An example pdf plot file is shown below in figure 4.6a for Bulk B homozygote frequency in relation to the RIL25 mutant parent.

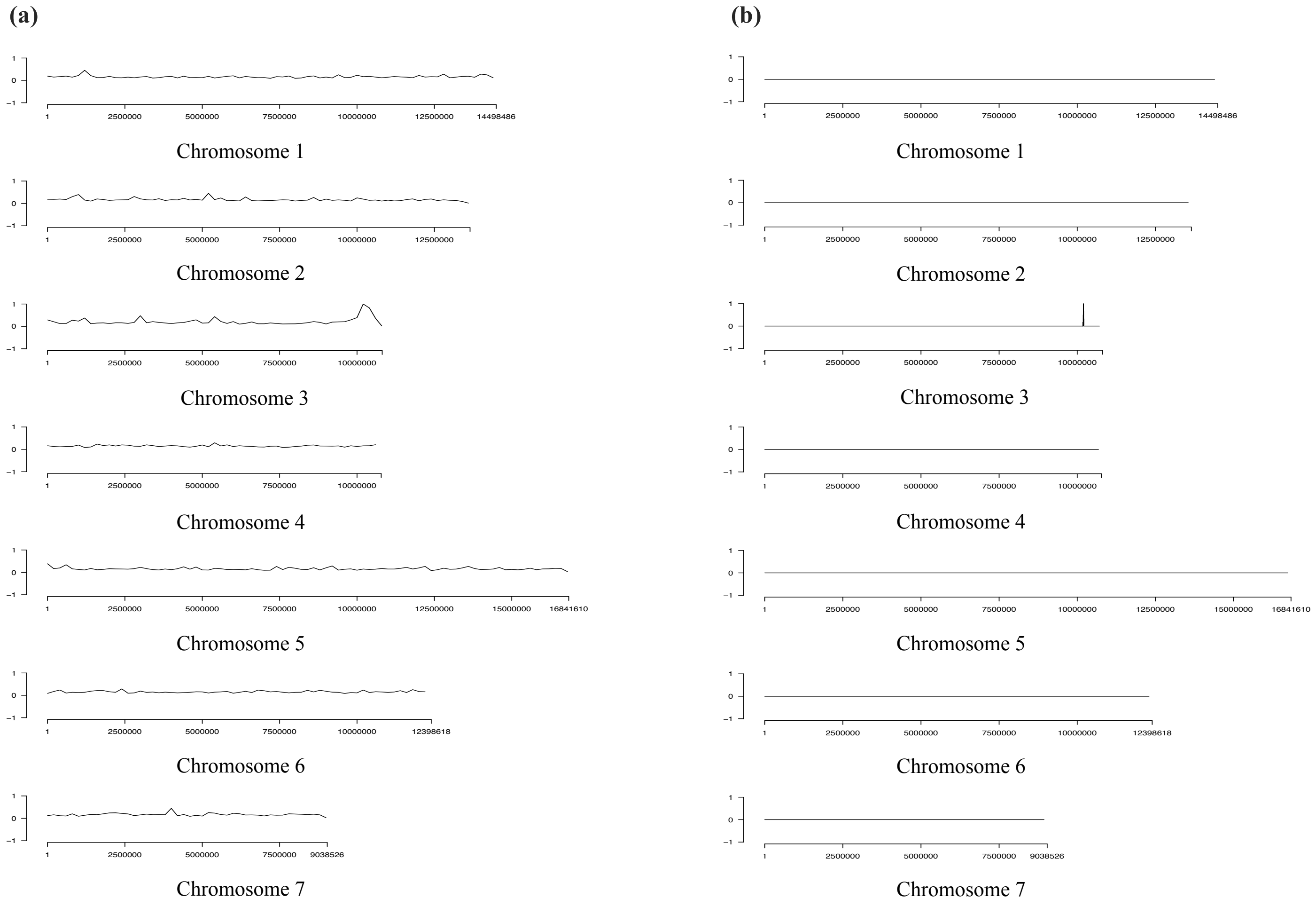


Figure 4.6. Homozygosity frequencies/scores calculated for Bulk A and Bulk B datasets along each pseudo-chromosome. Automated R plots for; **(a)** Frequency of ‘RIL25 homozygous’ SNPs per window; Bulk B frequency per 100,000bp window **(b)** Scores plotted using final algorithm in relation to Bulk B SNP positions that are conserved with the RIL25 parental unique homozygote SNPs. Scores calculated per 100,000bp window and calculated at 1000bp intervals.

The final haplotyping algorithm detailed in figure 4.2c (Haplotyping_hex_wheat_plus_plot.pl) requires a list of SNP positions where SNP alleles that are in the mutant bulk segregant offspring are seen as homozygotes in the mutant parent that are unique to it; a tab separated file [chromosome <tab> position <tab> Alternate allele <tab> % reads with alternate allele] this can be filtered more easily from the VarScan output. A tab separated file specifying a line for each chromosome [chromosome <tab> length] is also required. Otherwise the user defines the window size for analysis (typically 200,000bp) and output file prefix. Additionally if the user does not define homozygote, heterozygote and borderline limits those that are defined in figure 4.2c are implemented i.e. diploid settings. However if analysis of a polyploid is to be carried out these limits can be user defined at the command line according to ploidy number. Such methodology is employed in section 4.3 for a hexaploid e.g. a homozygote defined as in ~33% of sequencing reads. The resultant output is a text file containing [Chromosome <tab> Start position of interval on chromosome <tab> homozygote score] plus a pdf file of the resultant raw frequency plot in R. An example pdf plot file is shown in figure 4.6b for Bulk B homozygote scores in relation to the RIL25 mutant parent.

The complete mapping/SNP calling pipeline plus final haplotyping algorithm is also available on iPlant (The iPlant Collaborative, 2011) as two workflows within the Discovery Environment; ‘Mapping Illumina seq data Part 1’ and ‘SNP calling Illumina seq data Part 2’. These workflows map and SNP call in Illumina sequencing datasets, ideally requiring a mutant parental line, a wild type parental line and a bulk segregant mutant F2 pool as input as was used within this study. The workflows allow user definition of parameters but the parameters that were used for this study are implemented by default and the 2 parental SNP lists generated as output are used as input for the workflow; ‘Identification of unique homozygous SNPs in mutant’ to identify mutant parental specific SNPs. Finally the workflow; ‘Mutant Identification 1’ takes this mutant parent specific SNP list and the bulk segregant mutant F2 population SNP list as input, finds conserved SNP alleles between the 2 and implements the homozygote haplotyping algorithm to output the pdf file R plot seen in figure 4.6b identifying the mutant interval of interest. The text file of homozygote scores used to plot this is also generated as output.

4.3 Identification of genes linked to stripe rust resistance in 2 hexaploid wheat mutants

4.3.1 Mapping and SNP identification pipelines

Two parental purebred lines of the wheat varieties Avalon and Cadenza were crossed to produce F1 progeny. The F1 progeny haploid cells were then allowed to undergo chromosome doubling to produce double haploid offspring. Two distinct pools of F1 double haploid offspring were created; Pool one (P2) contained 53 individuals with a score of less than or equal to 2 of seedling reaction to an AVRYr7 isolate i.e. yellow stripe rust resistant lines. Pool two (P3) contained 50 individuals with a score of greater than or equal to 7 of seedling reaction to an AVRYr7 isolate i.e. yellow stripe rust susceptible lines.

The two bulk segregated populations, that were equivalent to F2 mapping populations, P2 and P3, along with purebred parental Avalon and Cadenza lines were developed by Donal O'Sullivan's group and enriched using the NimbleGen wheat gene capture array and sequenced using the Illumina HiSeq technology at the CGR, with the same methodology as section 4.2, generating paired end reads (2 x 100bp). The pipeline that was developed for processing of this sequencing data is summarized in figure 4.7 and is a derivative of that shown in figure 4.2a plus 4.2c. The pipeline shown in figure 4.2b was depreciated and is not used here. The theory behind figure 4.7a/4.7b and figure 4.2a/4.2c is identical; the main changes being the replacement of RIL25 and RIL71 parental lines with Avalon and Cadenza hexaploid wheat lines, the replacement of Bulk A/B bulk segregant datasets with 2 pools P2 and P3 and the limits for homozygote, heterozygote and borderline SNP categorizing have been adjusted accordingly for a hexaploid dataset.

The sequence datasets for the 2 bulk segregated populations P2 and P3, along with purebred Avalon and Cadenza were all mapped to the pseudo-chromosome sequences, generated from the gene capture array design-space, using BWA-short fragment mapping; Indexing of the reference sequence involved use of the 'IS' algorithm and 4 mismatches were allowed per mapped read with a Q score of 20 to allow read trimming in areas of low quality. All unmapped, non-uniquely mapped and duplicate reads were later removed using SAMtools. Finally SAMtools mpileup (v 0.1.18) (Li, H. *et al.*, 2009) was implemented on the 4 datasets and SNP calls were filtered out using VarScan (VarScan.v2.2.11.jar) (Koboldt, D. *et al.,* 2012) with the same parameters that were implemented in section 4.2.1. The steps involved in this analysis are shown in section figure 1.6 and in the command outline appendix sections 1 and 4.

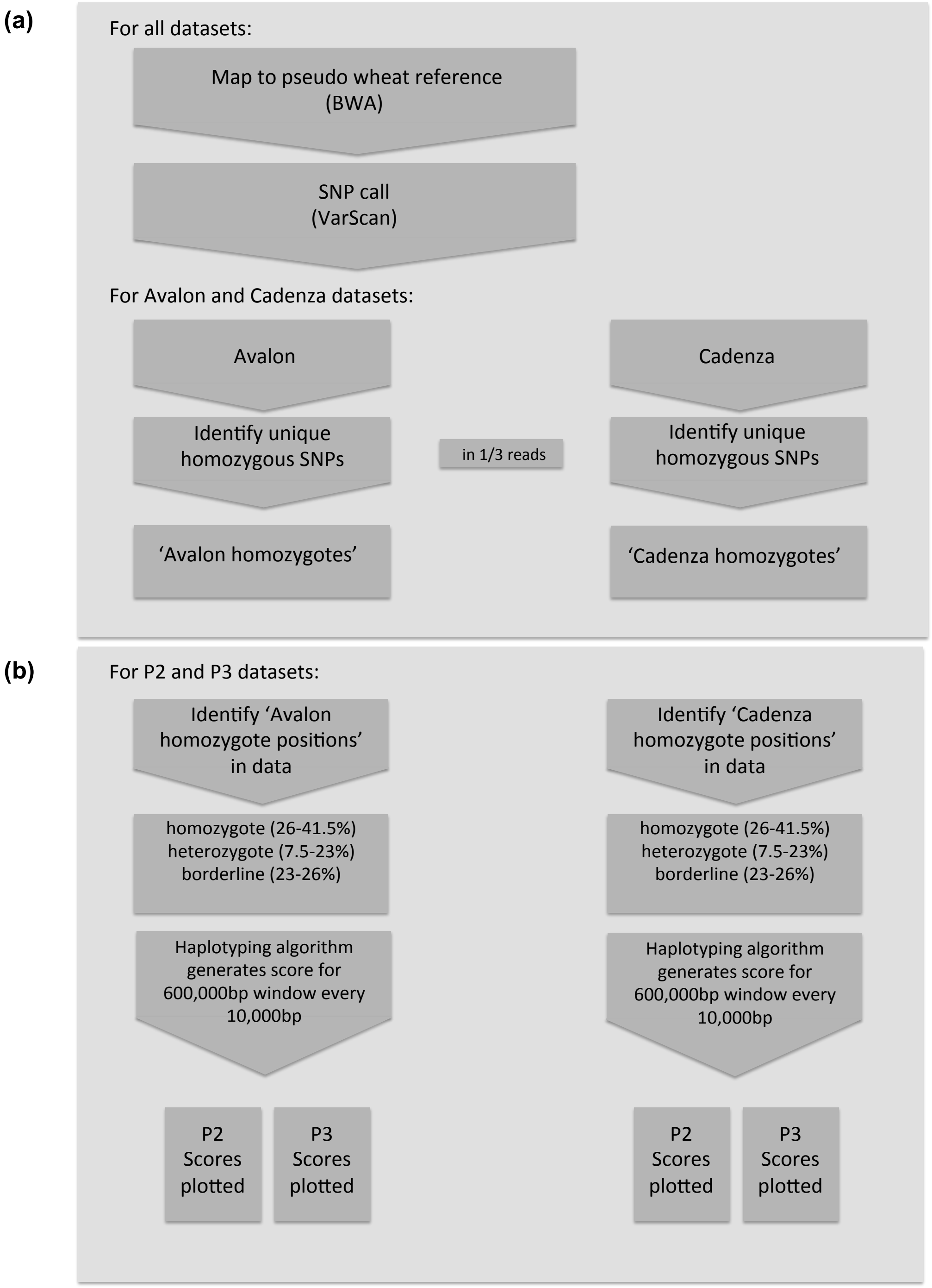

(a)
For all datasets:
Map to pseudo wheat reference (BWA)
SNP call (VarScan)
For Avalon and Cadenza datasets:
Avalon
Cadenza
Identify unique homozygous SNPs
in 1/3 reads
Identify unique homozygous SNPs
'Avalon homozygotes'
'Cadenza homozygotes'
(b)
For P2 and P3 datasets:
Identify 'Avalon homozygote positions' in data
Identify 'Cadenza homozygote positions' in data
homozygote (26-41.5%) heterozygote (7.5-23%) borderline (23-26%)
homozygote (26-41.5%) heterozygote (7.5-23%) borderline (23-26%)
Haplotyping algorithm generates score for 600,000bp window every 10,000bp
Haplotyping algorithm generates score for 600,000bp window every 10,000bp
P2 Scores plotted
P3 Scores plotted
P2 Scores plotted
P3 Scores plotted

Figure 4.7. Processing 4 sets of enriched sequencing data to identify a mapping interval containing the gene that is inducing the phenotype of interest. (a) Standard mapping and SNP calling pipeline **(b)** Pipeline implementing the final allele haplotype frequency algorithm (utilized in figure 4.2c) for P2 and P3 bulk segregant samples to identify the interval of interest

Approximately 94% of this pseudo wheat reference was mapped across all 4 datasets (P2, P3, Avalon and Cadenza) with an average depth of coverage of approximately 45 in Avalon and Cadenza and 75 in P2 and P3. Further parsing of the VarScan SNP output allowed homozygous SNPs for one of the three genomes between the Chinese Spring reference and the two parental lines (Avalon and Cadenza) to be identified (SNPs in 23-43% of the sequencing reads with the reference allele in 57-77% of reads). This resulted in 2 lists 219,550 Avalon specific homozygotes and 254,101 Cadenza specific homozygotes. The high likelihood that a SNP allele that was in ~33% of the sequencing reads had come from one homozygous genome rather than 2 identical heterozygote genomes ensured the success of this methodology. Any deviations or locations with multiple alternate alleles were so small in number that they did not have an effect on the analysis outcome.

Mapping positions from Avalon specific homozygotes list were located in each of the P2/P3 datasets individually and if they had a depth of coverage greater than or equal to 50 and if the SNP allele that was identified in Avalon could be found in the P2/P3 data then they were added to a P2 and a P3 'Avalon unique homozygous SNP list' respectively (P2; 14,868 SNPs and P3; 16,336 SNPs). These lists detailed the mapping position of the SNP; its alternate allele in Avalon and the percentage of sequencing reads that this alternate allele could be seen in in the respective bulk segregated dataset (P2 or P3). The exercise was repeated using the Cadenza specific homozygotes to generate lists for P2 and P3 if the SNP allele was conserved (P2; 17,235 SNPs and P3; 17,137 SNPs).

Thus 4 lists were created; for each of the P2 and P3 datasets one each of; conserved Avalon unique homozygous SNPs and conserved Cadenza unique homozygous SNPs. These files were analyzed using the script Haplotyping_hex_wheat_plus_plot.pl. A tab separated file specifying a line for each chromosome [chromosome <tab> length] was also required. Otherwise the user defined window size for analysis was altered as necessary between datasets and the user-defined homozygote, heterozygote and borderline limits were altered to enable effective SNP calling in a hexaploid as per figure 4.7b. The relevant scores per window along each chromosome were outputted for each dataset in relation to both parents.

4.3.2 P2 and P3 datasets; Mutant identification

Homozygosity scores, were outputted every 10,000bp for each 600,000bp chromosomal window for the P2 and P3 datasets and plotted in relation to both parents. A larger window size than the typical 200,000bp window was used due to the expectation that the defined interval would be larger because of; the lower number of pooled plants combined with the hexaploid nature of wheat. These window sizes were determined to be best for each dataset individually and the plots are detailed in figure 4.8 where the Avalon and Cadenza relevant scores have both been plotted together.

Figure 4.8a shows the result gained from the analysis in the sample P2. It shows one main peak that is seen in Cadenza on chromosome 2 (~7.5Mbp) and not seen in Avalon. Since plants in this pool were bulk segregated on the basis of shared stripe rust resistance to an AVRYr7 isolate, that was likely to be inherited from the Cadenza parent, they are expected to show a high degree of homozygosity that is shared with Cadenza around the Yr-7 locus. The Yr-7 locus is documented to be on wheat chromosome 2 so this peak could be indicative of the gene region (McIntosh, R. A. *et al.,* 1998). The approximately plateaued tip of this peak on chromosome 2 was determined to be between 7,200,001 and 7,880,001bp. This region was extended to include the entire peak for downstream analyses i.e. 6,870,001-8,380,001bp and was used in a BLASTN alignment to the BLAST nr nucleotide database (Altschul *et al.,* 1990) (e-value < 1e-3). The region encompassed; 247 homozygous SNPs (122 in tip of peak interval) that are unique to sample P2 and detailed along with their associated genes in appendix 3, table 4. It has been previously observed that most disease resistance genes in plants encode nucleotide binding site leucine-rich repeat proteins (NBS-LRR) (McHale *et al.,* 2006). It was noted from appendix 3, table 4 that 1 SNP was associated with the NBS-LRR disease resistance protein homologue and that 6 SNPs were associated with regions similar to the *Brachypodium* RGA4-like disease resistance protein that is also known to be of the NBS-LRR protein family (Ratnaparkhe *et al.,* 2011). The 7 SNPs were all unique to P2 and, in the region 7,518,606–7,556,492bp, they were approximately central to the tip of the peak interval of interest that was defined for sample P2 on chromosome 2. Analysis of pooled plants based on shared stripe rust resistance due to presence of the Yr-7 stripe rust resistance locus has allowed the location of SNPs that are associated with a gene that is likely to be linked to such disease resistance.

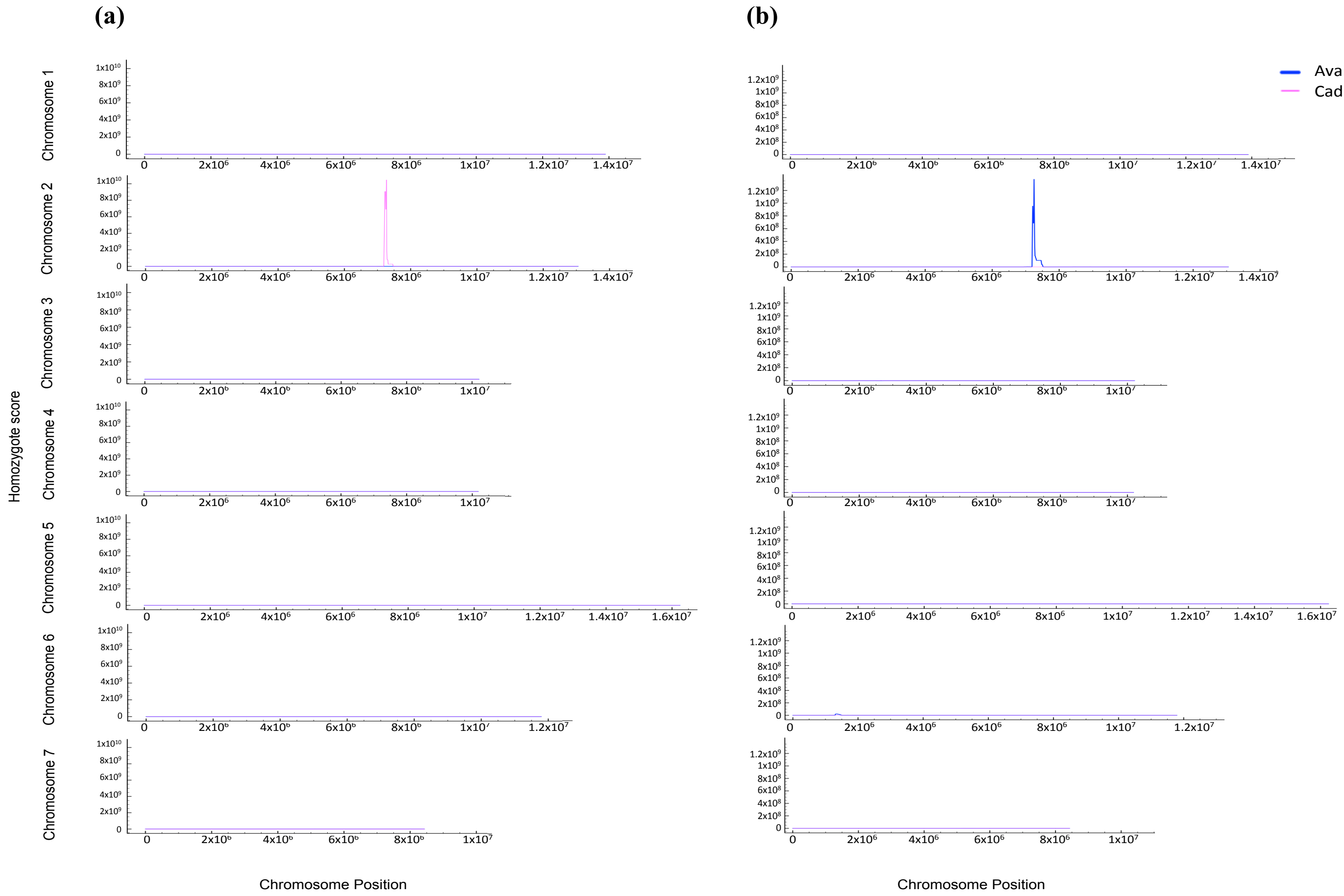


Figure 4.8. Homozygosity scores calculated for the P2 and P3 bulk segregant datasets along each pseudo-chromosome. (a) Scores calculated per 600,000bp window along each chromosome at 10,000bp intervals. **Magenta line;** Scores plotted for 'Cadenza unique homozygote SNPs' found in the P2 bulk segregated dataset. **Blue line;** Scores plotted for 'Avalon unique homozygote SNPs' found in the P2 bulk segregated dataset. **(b)** See **(a)** but scores derived from P3 dataset.

Figure 4.8b shows the result gained from the analysis in the sample P3. It shows 1 main peak region in Avalon on chromosome 2 (~7.5M bp) that is not seen in Cadenza. Plants in this pool were bulk segregated on the basis of shared stripe rust susceptibility to an AVRYr7 isolate due to absence/disruption of the Yr-7 locus that was likely to be inherited from the Avalon parent. They are expected to share a high degree of homozygosity with Avalon in this gene region that is documented to be on wheat chromosome 2 (McIntosh, R. A. *et al.,* 1998). This peak could therefore be indicative of the Yr-7 locus. We expect P3 to be an approximate mirror image of P2 due to one being resistant and one being susceptible to the same AVRYr7 isolate; this mirror image of plots has been seen as the peak for P2 was in the chromosome 2 Cadenza plot and the peak in P3 is approximately at the same position in chromosome 2 but in Avalon. This acts as a reinforcement of the accuracy of the analysis. The approximately plateaued tip of the peak that was seen in P3 was determined to be between 7,200,001 and 8,060,001bp. This region was extended to include the entire peak for downstream analyses i.e. 6,940,001-8,190,001bp and was used in a BLASTN alignment to the BLAST nr nucleotide database (Altschul *et al.,* 1990) (e-value < 1e-3). The region encompassed 197 homozygous SNPs (143 in tip of peak interval) that are unique to sample P3 and detailed along with their associated genes in appendix 3, table 5. 11 SNPs were associated with the NBS-LRR disease resistance protein homologue and 10 further SNPs were associated with regions similar to the *Brachypodium* RGA4-like disease resistance protein that is also of the NBS-LRR protein family (Ratnaparkhe *et al.,* 2011). 21 SNPs were identified in P3, in this region that showed homology to the NBS-LRR disease resistance proteins, while only 7 SNPs were identified in a similar sized and located interval in P2. Pooling and analysis of samples based on shared stripe rust susceptibility has allowed definition of a group of homozygous SNPs that are in gene regions that are likely to be linked to stripe rust resistance. The elevated level of SNPs found in P3 compared to P2 could be responsible for the disruption of transcription of the disease resistance gene and disease susceptibility in this sample. The 21 SNPs are therefore candidates for further investigation; they were unique to P3 and in the region 7,508,311–7,553,295bp approximately central to the tip of the peak interval of interest that was defined for sample P3 on chromosome 2.

The 9000 SNP Infinium assay (iSelect array) includes SNPs that were discovered in transcriptomes generated from multiple wheat lines. Largely annotated, the array can be used to genotype a diverse set of polyploid wheat lines (Cavanagh *et al.,* 2013). 296 Yr-7 linked iSelect SNP sequences were used in a BLAST search against the entire wheat enrichment array to find their relative positions. 193 had hits (65%) and 159 could be allocated to a pseudo-chromosome sequence position. Of these SNPs 144 (90%) were correctly anchored to pseudo-chromosome 2 with ~80% of these within the region 6,000,00-9,000,000bp. The relative scores (Cadenza/Avalon unique homozygote SNP scores per window) for the 144 iSelect Yr-7 SNP positions were extracted for each of the P2/P3 datasets and plotted in figure 4.9 (see Appendix 3, table 6 for list) to identify those that were situated within figure 4.8's peak interval. If multiple windows hit the SNP the average score was taken.

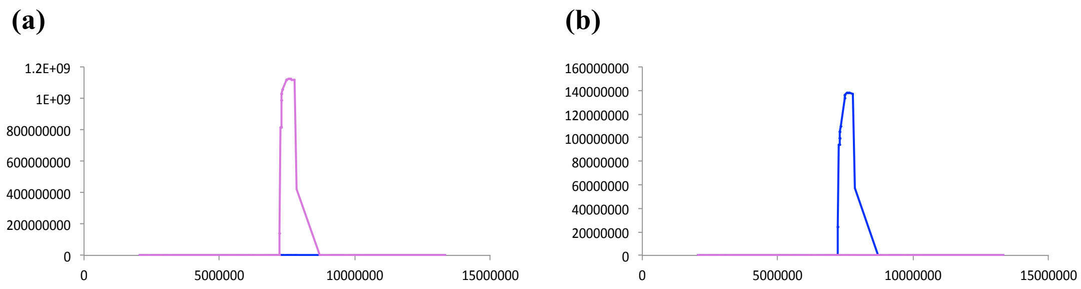


Figure 4.9. Homozygosity scores calculated from the P2 and P3 bulk segregant datasets for the iSelect Yr-7 linked SNP positions. Scores originally calculated per 600,000bp window along each chromosome at 10,000bp intervals but reported here only for windows on chromosome 2 containing iSelect Yr-7 linked SNPs. **(a) Magenta line;** Scores plotted for 'Cadenza unique homozygote SNPs' found in the P2 bulk segregated dataset at iSelect Yr-7 linked SNP positions. **Blue line;** Scores plotted for 'Avalon unique homozygote SNPs' found in the P2 bulk segregated dataset at iSelect Yr-7 linked SNP positions. **(b)** See **(a)** but scores derived from P3 dataset.

In sample P2 the 30 Yr-7 linked iSelect SNP positions that gained a score above 1x10^{9} (~10% of graph limit) are detailed in table 4.3. These SNP scores made up the peak in figure 4.9a and were derived from the Cadenza specific list as anticipated. The SNPs mapped to the region 7,293,769-7,783,887bp on chromosome 2 i.e. tip of the peak interval in figure 4.8a. In sample P3 an identical list of 30 Yr-7 linked iSelect SNP's that gained a score above 1x10^{9} are detailed in table 4.3. These SNP scores made up the peak in figure 4.9b, are found at the tip of the peak interval in figure 4.8b, and were derived from the Avalon specific list as anticipated. Of the 30 iSelect SNP positions 4 and 1 were conserved with the P2 and P3 homozygous SNP lists respectively i.e. alternate allele found in the population. In the BLASTN search these SNPs showed no homology with disease resistance proteins (not in the list of 7 P2 and 21 P3 candidate SNPs) and therefore may demonstrate proximity to the Yr-7 locus only. Locating Yr-7 linked SNP positions within and around the defined peak in the P2/P3 datasets suggests that the Yr-7 locus has been correctly located here even if the datasets do not have SNP alleles at many of these positions.

iSelect SNP name	Position on Chromosome 2	P2 Cadenza Score	P3 Avalon Score
RFL_Contig337_1645	7293769	1027635761	104896474
Tdurum_contig54925_202	7293769	1027635761	104896474
Tdurum_contig54925_225	7293792	1027635761	104896474
RFL_Contig337_1432	7293982	1027635761	104896474
Tdurum_contig54925_415	7293982	1027635761	104896474
BobWhite_c19554_544	7294114	1027635761	104896474
BS00088489_51	7294306	1027635761	104896474
GENE-1125_32	7294306	1027635761	104896474
Tdurum_contig46389_1838	7318237	1054095226	109368420
Tdurum_contig46389_1540	7318535	1054095226	109368420
Tdurum_contig46389_1459	7318616	1054095226	109368420
Excalibur_c5557_201	7456131	1116094943	132971055
BS00022717_51	7456454	1116094943	132971055
BS00023060_51	7456706	1116094943	132971055
Excalibur_rep_c68985_110	7462316	1120523493	134656957
RAC875_rep_c118667_79	7472424	1120826498	135792788
RAC875_c28108_144	7472838	1120826498	135792788
TA005830-0667	7472956	1120826498	135792788
RAC875_c28108_400	7475389	1120826498	135792788
BS00011825_51	7475593	1120826498	135792788
IACX8470	7475580	1120826498	135792788
Kukri_c18058_764	7476327	1120826498	135792788
RAC875_rep_c85788_180	7564051	1121253728	137829007
CAP8_rep_c8162_101	7598207	1121283358	137821194
wsnp_Ex_c12922_20472434	7606991	1121227500	137779505
wsnp_Ex_c12922_20473104	7609812	1121200238	137765496
Kukri_c25716_284	7614022	1121200238	137765496
Kukri_c25716_445	7614182	1121087765	137751488
Excalibur_c10071_213	7655834	1121087765	137751488
wsnp_BE490267A_Ta_2_1	7783887	1120637065	137667033

Table 4.3. Yr-7 iSelect SNPs mapping to wheat pseudo-chromosome 2. Scores have been extracted if greater than 1x10^9 and averaged over the 60 windows hit per SNP. P2 Cadenza scores and P3 Avalon scores detailed.

4.4 Conclusions

In section 4.2 a mutant bulk segregant F2 population of the non-model grass *T. monococcum* was developed from parental RILs, it was target enriched for genic regions and a region was identified on chromosome 3 that is likely to contain the *Eps-3A^m* mutation. A region of

~600Kbp was initially identified within the pseudo-chromosomes and within this region a ~40Kbp region could be pinpointed based on the identification of deletion hotspots. Finally, by assessing gene annotation, the candidate gene for the phenotype itself could be narrowed to a single capture design-space contig of 3693bp that had a high deletion frequency and showed a high degree of similarity to the *T. aestivum* cultivar Chinese Spring LUX gene. The LUX gene is known to affect both the circadian clock and flowering time in *Arabidopsis* (Hazen *et al.,* 2005). This was all made possible with the development of a bespoke mapping/SNP calling and mutant identification algorithm.

Additionally the use of a target enrichment strategy using capture probes that have been designed against the hexaploid wheat Chinese Spring to enrich the genic portion of a closely related plant has been demonstrated gaining on average 70x mapping coverage across 70% of the pseudo-chromosome reference sequence. This highlights the possibility for other capture probe sets to be used for close relatives with little or no resources available e.g. the soybean NimbleGen SeqCap EZ in solution exome enrichment probe set could be applied for study of the pea.

This study extends a proof of concept approach where enrichment of a subset of a phenotyped *Arabidopsis* F2 mapping population was performed in combination with a mapping-by-synteny approach to order *Arabidopsis* cDNA into *B. rapa* pseudo-chromosomes based on synteny. Two mutant intervals were defined in *B. rapa* using allele frequency analysis at marker positions. This translated to one position in *Arabidopsis* (Galvão et al 2012). Here, the full genic sequence of wheat was enriched and ordered into wheat pseudo-chromosomes based on synteny with the closely related *Brachypodium* to allow sliding window mapping-by-sequencing analyses. The mutant deletion could be identified directly as a position in wheat. The combination of sliding window analyses and mapping-by-synteny, implementing a pseudo genome directly, has not yet been documented. By targeting the majority of wheat's genic sequence the concerns expressed by Galvão et al, that the causal mutation would be unlikely to be targeted with enrichment, are addressed. Here the likelihood of enrichment of the region of interest is increased and not only has a more divergent species been used to order the fragmented mapping reference, the mapping reference and enrichment capture probe set are both divergent from the analyzed species.

A group at Reading University led by Donal O'Sullivan provided the P2/P3 Yr-7 stripe rust resistant/susceptible wheat datasets. Analysis of these datasets yielded confident definition of almost identical peak intervals and a small group of novel SNPs in common disease resistance genes that the group will go on to investigate further. In the P2 dataset 7 SNPs were associated with the NBS-LRR disease resistance protein family; these SNPs were all unique to P2 and approximately central to the tip of the peak interval of interest that was defined for this dataset on chromosome 2. This SNP region is likely to indicate the location of the Yr-7 locus for stripe rust resistance. In the P3 dataset 21 SNPs were associated with the NBS-LRR disease resistance protein family. This SNP increase in the P3 dataset compared to P2, in the same peak region on chromosome 2 that showed homology to the NBS-LRR disease resistance proteins, could be responsible for the disruption of the disease resistance gene and therefore disease susceptibility in this sample and as such these 21 novel SNPs (not iSelect positions) are candidates for further study.

113 Yr-7 linked iSelect SNP sequences could be correctly anchored to pseudo-chromosome 2 within the region 6,000,00-9,000,000bp. 30 iSelect SNP sequences were located at the tip of the peak interval in both the P2 and P3 datasets. 5 of these 30 were also found in the P2/P3 homozygous SNP lists, i.e. alternate allele seen in the populations, though these positions could not be associated with disease resistance genes and are unlikely to be candidates for stripe rust resistance/susceptibility. The ability to find Yr-7 linked SNP positions within the defined peak in P2 and P3 suggests that the Yr-7 locus has been correctly located here even if the P2/P3 datasets do not have SNP alleles at many of the iSelect SNP positions. In theory all of the Yr-7 linked SNPs should relatively closely associate with the peak regions in the BLASTN search. Cases where this was not true are likely to be a result of local inaccuracies in the contig ordering that was used to make up the pseudo-chromosomes. The full peak region that is identified can encompass over 2Mbp of genic material. This can be tentatively narrowed down to peak plateaus that tend to be less than 1Mbp. Homing in on the peak tip could be more confidently relied upon if the pseudo-chromosomes contained all genic material and contig order could be fully confirmed. James *et al.* noted that, at a minimum of 15x coverage, larger pools of ~200 plants generated interval sizes 159-603Kbp, while smaller pools of ~50 plants generated interval sizes 216-1350Kbp (Velikkakam *et al.*, 2013). Here the average pool size was ~66 therefore the defined interval of ~1Mbp is approximately within the anticipated range.

In chapter 3 utilization of the IWGSC wheat chromosome assemblies showed that 80% of the design-space contigs that were concatenated into the pseudo-chromosomes were found to have been associated with the correct chromosome. All figures in this study have been re-plotted retrospectively after the removal of any probes that had a chromosomal position that could not be validated. Almost identical results were observed when this is compared to plots prior to removal with no noteworthy deviations.

Chapter 5. A comprehensive genome wide analysis of methylation patterns in wheat

Here a study of methylation patterns in wheat is outlined that utilized sodium bisulfite treatment combined with target enrichment. An enrichment system was specifically designed, developed, validated and implemented here to perform one of the first studies of methylation patterns in hexaploid bread wheat across the 3 genomes that used a genome-wide subset of genes and can thus be used to infer genome-wide methylation patterns and observations. This investigation confirmed that differential methylation exists between the A, B and D genomes of wheat and that temperature is capable of altering methylation states.

5.1 Introduction

Methylation of the cytosine residues in eukaryote DNA is thought to act as a mechanism of gene expression control. As outlined in section 1.8 it is clear that the location of methylation within or around a gene is important, however, the reasoning for this is, as of yet, poorly understood (Brenet *et al.,* 2011) and as such the various predicted effects of methylation, depending on gene location, remain controversial and largely without clarification. Rabinowicz *et al*. analyzed a small subset of whole genome sequencing data for hexaploid wheat and identified a high number of genes. They predicted high levels of methylated pseudogenes in wheat (recently amplified and then silenced), reducing the number of active genes to a level closer to that, which was expected (Rabinowicz *et al,* 2005).

The 3 main methods used in the laboratory for the study of methylation patterns include; bisulfite treatment, differential enzymatic cleavage and affinity based methods that use antibodies or proteins to pull down methylated DNA. Here the study of methylation patterns in wheat was tested using bisulfite treatment. Sodium bisulfite treatment is an increasingly popular method for epigenetic profiling and allows effective discrimination of the methylation status at every cytosine residue making this method the gold standard in methylation studies (Darst *et al.,* 2010). The clear significance of the impact of methylation on the genome makes it an obvious area for research but to study methylation in the large hexaploid wheat genome, without encountering the problems previously detailed due to the large size of the wheat genome, the combination of a bisulfite treatment with target enrichment was the best way forward. This technique was used to test: firstly, if differential methylation exists between the A, B and D genomes; secondly, using two growth temperatures for the Chinese Spring to test if temperature is capable of altering the methylation state and to see if this is both genome specific and genome independent; finally, to investigate if it is this underlying methylation that can control both genome specific and temperature dependent changes in gene expression.

Previously detailed capture probe sets i.e. the gene and exome capture arrays, were designed with, and made by, NimbleGen (detailed in section 1.6). Such capture probes are typically less than 100bp in length, DNA based and tiled across the design-space for the array. Here an Agilent Sure select custom capture probe set was designed and used for the methylation study in preference to a NimbleGen probe set since at this time NimbleGen did not support the use of their capture probe sets for the study of methylation patterns using bisulfite treatment. Agilent capture probe sets utilize 120bp RNA baits that, here, do not overlap.
Agilent's Sure Select Methyl-Seq Target Enrichment System allows the study of methylation patterns in target regions. Such methodology opens up the possibility of cost effective epigenetic profiling in large genomes and here it is demonstrated that enrichment can be used to give a genome-wide overview of methylation patterns across a subset of genic regions in the wheat genome. As an initial proof of principle an enrichment capture probe set was designed to target a 6Mbp subset of the genic regions of wheat. Here this will be referred to as the methylation capture probe set or array. To design this methylation capture probe set, a subset of regions were selected from the wheat gene capture array design-space that was introduced in figure 1.11 and validated in chapter 3. This design-space had already been through extensive validation and this ensured that probe sequences that were derived from it, similarly to those in the gene capture array, were; unique, non-repetitive, gene-rich, represent all 3 genomes, exclude chloroplast and mitochondrial sequence and are evenly distributed across the wheat genome. Use of this methodology ensured a successful methylation array probe set with little additional validation needed. The methylation array's enrichment performance is detailed in section 5.4.

The methylation array was used to test a number of hypotheses in wheat (section 5.6-5.8); firstly that Chinese Spring hexaploid wheat DNA could be enriched using the array to see if differential methylation exists between the A, B and D genomes. A list of naturally occurring homeologous SNP positions within the array bait sequences would allow identification of differential methylation between the A, B and D genomes in this analysis. Such SNPs make it possible to associate sequencing reads with a homeologous SNP allele and ultimately a particular wheat genome and this methodology is detailed in figure 5.1 and section 5.3. This limited the analyzable dataset to those regions that could be associated with homeologous SNPs (see section 5.5 for details). Secondly, two growth temperatures were used for the Chinese Spring (12°C to represent a lower more ambient temperature for wheat growth in the UK and 27 °C to represent a contrasting high temperature for wheat growth) such DNA was enriched using the array and bisulfite treated to test if temperature is capable of altering the methylation state and to see if this is both genome specific and genome independent. Finally, gene expression was compared between the two growth temperatures to test whether this differential methylation correlated with changes in gene expression (section 5.8). The two Chinese Spring datasets grown at 12°C and 27°C were developed, enriched, bisulfite treated and sequenced by the CGR.

5.2 Design of the methylation array

The design-space (~110Mbp) of the gene capture array was used as a starting point for design of the methylation array. 50,000 120bp fragments were selected from this design-space to form the RNA baits that would be used in the 6 Mbp methylation array. Fragments were selected with 3 main properties; previous good mapping coverage of the region, the presence of homeologous SNPs in the region from previous mapping analyses (chapter 3) and good genome-wide representation. Regions were ranked on the basis of homeologous SNP presence and coverage and 120bp sequences were distributed evenly across the more 'desirable' base-space. The 120bp baits were uploaded onto Agilent's EArray online (custom array design tool) to allow submission for manufacture. Bait 'boosting' was selected to allow excess unused design-space (less than 1Mb in this case) to be filled with repeat sequences of baits that are predicted to perform less efficiently i.e. those with an above average GC content are 'boosted' to ultimately gain even depth of sequence coverage across the array.

In a BLAST search the methylation array baits hit 47% of the genes in the most closely related sequenced grass, *Brachypodium* and 34% of the 97481 full length wheat cDNA contigs that were identified by Brenchley, R. *et al* when only top hits for each probe with an e-value less than 1e-5 were considered. Moreover, aligned regions with 1e-5 and over 90% sequence identity to the wheat cDNA contigs that were identified by Brenchley, R. *et al* were used to determine the transcribed regions of the bait sequences. Approximately 37% of the array probe set sequence was identified as transcribed. Thus, both methylation patterns in transcribed and untranscribed regions could be analyzed (Brenchley *et al*., 2012).

For all mapping analyses in this study rather than mapping directly to the 6Mb 120bp Agilent probe sequences, unless otherwise stated, data was mapped to the 120bp probes plus any contiguous DNA sequence surrounding the probes that was available. These will be referred to as the extended methylation bait sequence and reference contigs ranged from 121bp-8835bp with a median length of 698bp. The total size of the mapping reference was therefore approximately 44Mb (35% transcribed).

5.3 Identification of a reference list of homeologous SNPs

Publicly available sequencing datasets representing the closest available diploid ancestors for genome A (*T. monococcum*), B (*Ae. speltoides*) and D (*Ae. tauschii*) were mapped individually to the extended methylation bait sequences to identify homeologous SNPs.

The sequencing datasets were mapped to the extended methylation bait reference sequence using BWA-short fragment mapping (version 0.7.4). The "IS" algorithm was implemented to index the reference sequence. Genome A data was generated externally on the Illumina GAIIx and the ~30bp reads that were generated were mapped using 1 mismatch per read. Genome B data was generated externally on the Illumina GAIIx (Brenchley *et al.* 2012) and the ~100bp reads were mapped using 4 mismatches per read. Genome D data was generated externally using SOLiD sequencing and the ~30bp reads that were generated were mapped with 4 mismatches with use of parameters to allow mapping of reads in colour-space (Brenchley *et al.* 2012). Mapping results were processed using SAMtools; any non-uniquely mapping reads, unmapped reads and duplicate reads were removed. SNP calling was carried out using the GATK pipeline due to diploid datasets (all steps in figure 1.6 and commands in the command outline appendix sections 1, 2 and 3). When the GATK Unified Genotyper was implemented for SNP calling; a minimum quality of 50 and coverage of 6 was used, SNPs were filtered using standard GATK parameters and homozygous SNPs only were selected.

A list of positions was identified at which all three genome's alleles (taken from the three ancestral genomes) were unambiguous, known, and at least one differed from the reference base and/or the other two genomes i.e. homeologous SNPs. All C/T or G/A SNPs were excluded from this list to avoid future confusion between genuine SNP sites and C/T conversions of un-methylated cytosines as a result of the bisulfite treatment. 38,384 homeologous SNPs were identified (Homeologous_SNP_list.txt). These SNP alleles were used to associate cytosine residue methylation status with a wheat genome (figure 5.1).

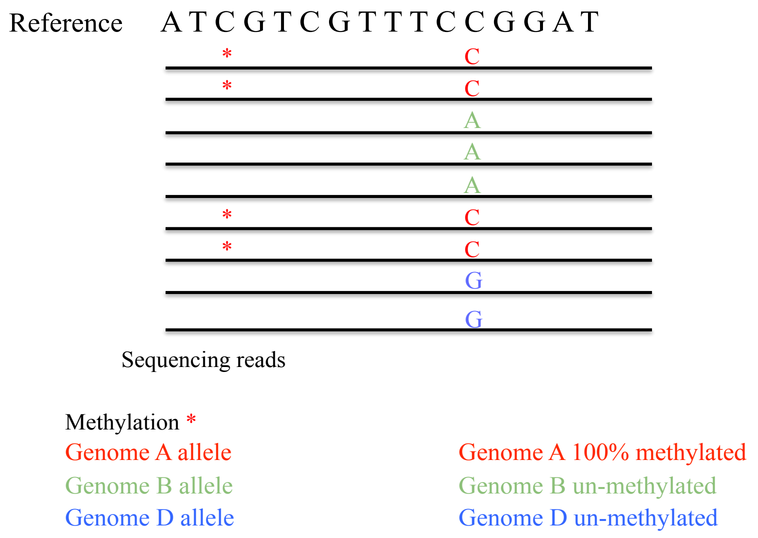


Figure 5.1. Theory behind association of methylation sites with the 3 wheat genomes. Illustrating the association of a homeologous SNP allele with a methylation site within the same sequencing read to allow determination of its genome of origin. Here genome A specific methylation is shown.

The full methodology that is utilized here to associate the methylation status of cytosine residues with a wheat genome using homeologous SNPs is outlined in figure 5.2. The pipeline generates a final output file detailing every cytosine residue, its associated reads, the genomes they have been associated with plus % methylation for each genome. Here cytosines are used in downstream analyses if all three genomes are mapped to by a minimum of 5 reads each.

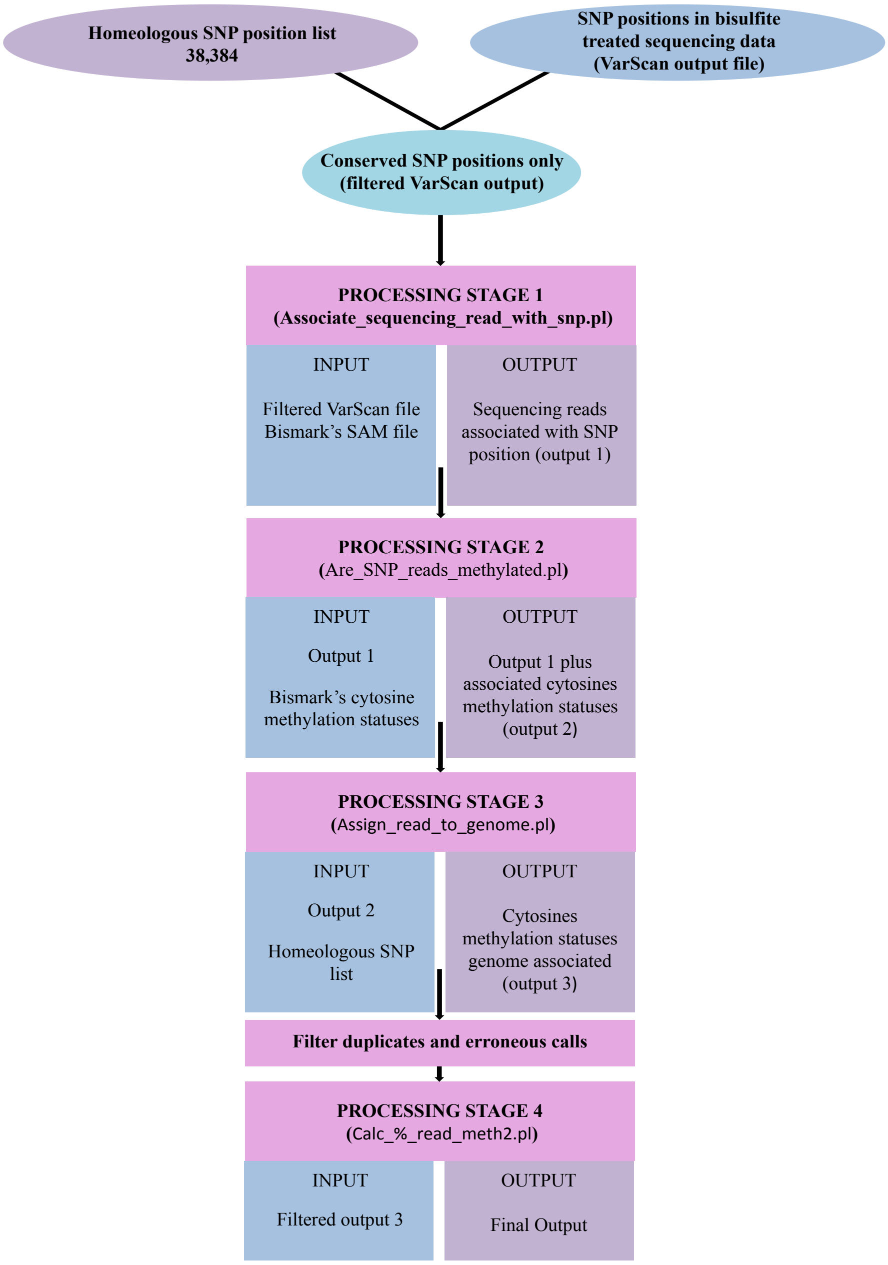

Homeologous SNP position list
38,384
SNP positions in bisulfite treated sequencing data (VarScan output file)
Conserved SNP positions only (filtered VarScan output)
PROCESSING STAGE 1
(Associate_sequencing_read_with_snp.pl)
INPUT
Filtered VarScan file
Bismark's SAM file
OUTPUT
Sequencing reads associated with SNP position (output 1)
PROCESSING STAGE 2
(Are_SNP_reads_methylated.pl)
INPUT
Output 1
Bismark's cytosine methylation statuses
OUTPUT
Output 1 plus associated cytosines methylation statuses (output 2)
PROCESSING STAGE 3
(Assign_read_to_genome.pl)
INPUT
Output 2
Homeologous SNP list
OUTPUT
Cytosines methylation statuses genome associated (output 3)
Filter duplicates and erroneous calls
PROCESSING STAGE 4
(Calc_%_read_meth2.pl)
INPUT
Filtered output 3
OUTPUT
Final Output

Figure 5.2. Pipeline for association of methylation sites with the 3 wheat genomes. SNP positions were identified in the enriched wheat bisulfite treated sequencing dataset using VarScan (reads mapping to these SNP positions have sufficient average mapping quality, sufficient depth overall and one or more alternate allele present). This SNP list is filtered for positions that are conserved in the homeologous SNP list. **Stage 1** takes this filtered VarScan list as input and Bismark's mapping output SAM file and outputs any sequencing reads of sufficient quality with an associated SNP plus the SNP allele represented in the read (output 1). **Stage 2** takes output 1 plus Bismark's 3 files of all CpG/CHH/CHG cytosine sites methylation statuses to generate output 2 with a line for each sequencing read/SNP association outlining how many cytosines in the read are methylated/un-methylated and

translating their positions to those from extended methylation bait sequences. **Stage 3** output 2 sequencing reads are associated with the appropriate wheat genome using their SNP allele plus the full homeologous SNP list as input to generate output 3. Duplicate lines, if a sequencing read has more than one associated SNP, were collapsed into 1 and SNPs associating one read with different genomes are known erroneous calls and filtered out from the file. **Stage 4** uses the filtered output 3 to produce a final output file detailing methylation status at every genome associated cytosine residue across the 3 genomes (all Perl scripts are available as supplementary data).

5.4 Enrichment performance and validation of the array

5.4.1 Non-Bisulfite treated samples

An initial trial was carried out using the methylation capture probe set for wheat enrichment prior to sequencing without the use of bisulfite treatment. This analysis acted as a control for comparison to determine if the array could efficiently enrich without the added complication of bisulfite treatment. Four genomic DNA samples were enriched using the array and sequenced by the CGR. These samples were all extracted from the areal tissue of 7 day old seedlings of the wheat variety Chinese Spring and included; two replicate plants (known as 12B and 12C) grown at 12°C and an additional two replicate plants (known as 27B and 27C) grown at 27°C. The genomic DNA was quantified and sheared for 6×60s using the Covaris S2 focused-ultrasonicator. Fragmented DNA quality and quantity were assessed on a Bioanalyzer High Sensitivity DNA chip (Agilent) prior to purification using 1.8 × Agencourt AMPure XP beads (Beckman Coulter). End-repair, 3′-adenylation, adapter ligation, enrichment and PCR were carried out. Amplified libraries were then indexed using 6 PCR cycles. Final libraries were quantified and pooled in equimolar amounts. Sequencing was carried out on an Illumina HiSeq 2000, using version 3 chemistry, generating 2 x 100bp paired end reads.

The four sequencing datasets for the samples were mapped to the extended methylation bait reference sequence using BWA-short (version 0.7.4) plus 4 mismatches per read. Mapping results were processed using SAMtools; any non-uniquely mapping reads, unmapped reads and duplicate reads were removed. SNP calling was carried out using SAMtools mpileup that was implemented prior to VarScan, to loosely identify positions containing an alternate allele, with a minimum coverage of 6, an average mapping quality above 15 and a minor allele frequency (MAF) of greater than 0.1. The steps involved in this analysis are shown in figure 1.6 and in the command outline appendix sections 1 and 4.

The sequence that was generated from the four wheat genomic DNA samples, sample 12 B and C and sample 27 replicate B and C had an average depth of coverage of 43.3 across 94.8% of the 6Mb array probe sequence. Mapping statistics between the repeat samples were comparable and a SNP comparison between 12B and 12C, in regions that were mapped by both datasets at a minimum depth of 15, revealed ~100% SNP conservation (27B and 27C also yielded ~100% SNP conservation). As such, the data was merged for the repeat samples 12B and 12C and also for 27B and 27C generating overall data for the 12°C and 27°C samples and resulting in an average depth of coverage of ~84x with ~96.4% of the 6Mb array probe sequence being mapped (see table 5.1 for full details). Notably an average depth of coverage of ~43x was observed with ~54% of the ~44Mb mapping reference being mapped across both samples i.e. the mapped region extended into surrounding next-to-target regions covering almost 4x the 6Mb capture probe space.

Sample	Mean % coverage per reference probe	Mean depth of coverage per reference probe	Number of Probes mapped (50000 total)	% of Reference probes mapped
12B	98.8	43	48104	96
12C	98.9	45.8	47985	96
27B	98.9	42.2	48043	96
27C	98.8	42.2	48034	96
12	99.4	86.3	48316	97
27	99.3	82.2	48339	97

Table 5.1. Mapping Statistics for enriched wheat DNA samples (non-bisulfite treated). All mapping statistics in relation to the 6Mb array probe base space. Mapping Statistics included for 4 original enriched wheat samples (12 B/C and 27 B/C) plus the data when replicate datasets were merged (12°C and 27°C)

The methylation array was seen to effectively enrich ~97% of the target 6Mb plus up to an additional 18Mb of sequence in its immediately surrounding regions. The depth of coverage across the probe set was largely consistent with less than 0.2% of baits exceeding 10x the average depth of coverage and only ~3% of probes unmapped. The extended methylation bait sequences were ordered and concatenated into pseudo-chromosome sequences to allow visualization of per chromosome mapping coverage that is shown for the 12°C sample in figure 5.3. These pseudo-chromosomes were designed as per the methodology in section 3.3 for the gene capture array design-space and allowed inclusion and approximation of the order of >80% of the methylation array probes.

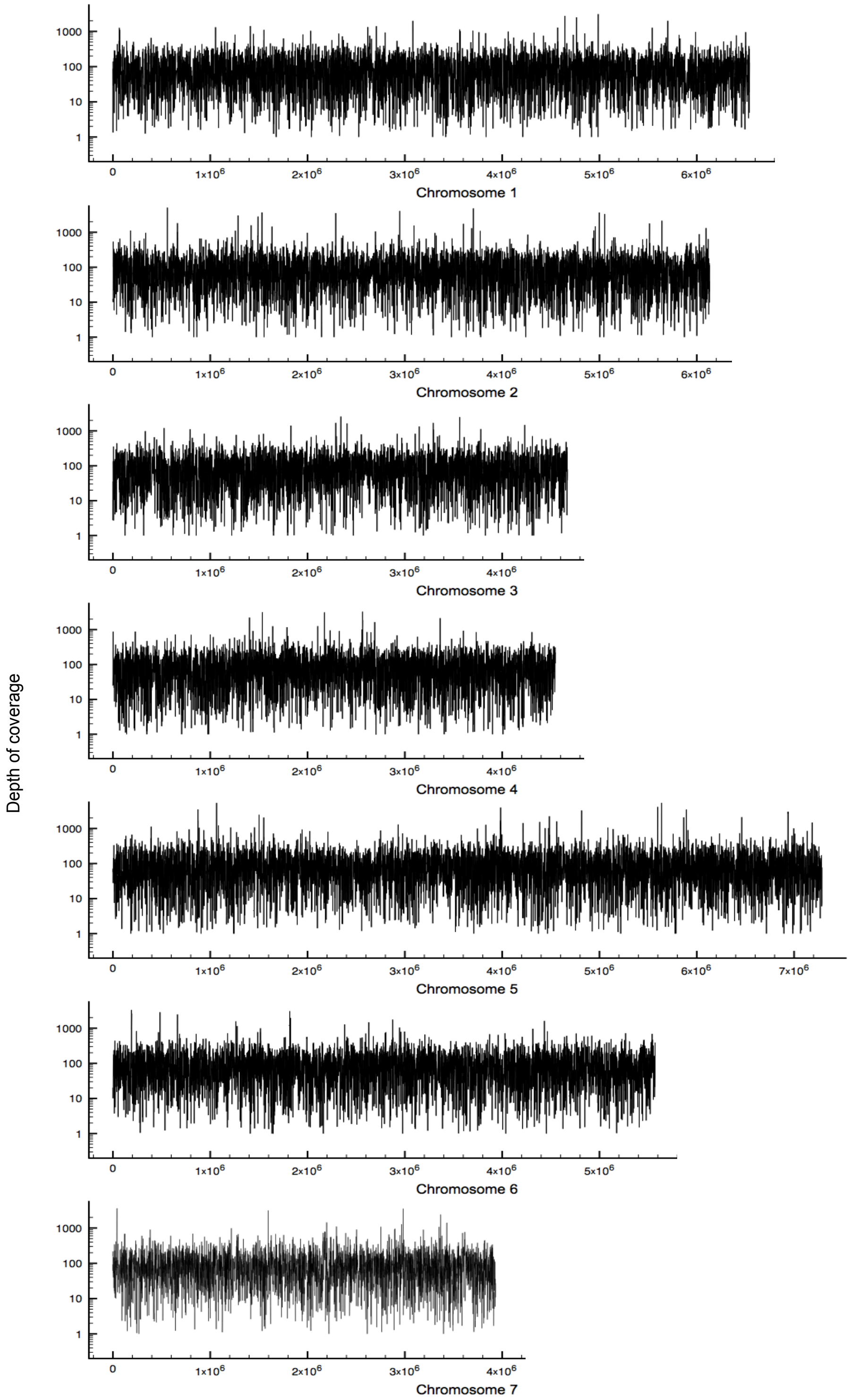


Figure 5.3. Average depth of coverage per bait probe plotted for the 12°C sample. Coverage plotted along each pseudo-chromosome construct after ordering and concatenation of extended methylation bait sequences.

5.4.2 Setting of thresholds for methylation

The standard thresholds used within this study to classify residues allow clear distinction of methylated and un-methylated regions. At each cytosine residue site, where three genomes can be identified, the percentage of the reads mapping to each genome that were methylated can be calculated using Bismark's categorization of sequencing reads as methylated or un-methylated at each cytosine residue. Thresholds of 100% (>= 75%), 50% (< 75% and > 25%) and 0% (<= 25%) methylation were used to categorize the data for easier comparison. It was noted that the vast majority of cytosine residues could be classified as methylated/un-methylated with less than 0.05% of residues classified as intermediate. Constant thresholds were utilized across CpG, CHH or CHG methylation sites to allow comparison between CpG and non-CpG site methylation (Harris *et al.,* 2010).

5.4.3 Bisulfite treated samples

The analysis in section 5.4.1 was repeated using six genomic DNA samples, plants (known as 12B, 12C and 12D) grown at 12°C and plants (known as 27B, 27C and 27D) grown at 27°C. This time the six samples were enriched using the array and after end-repair and 3'-adenylation; methylated adapter ligation, hybridization, bisulfite conversion and PCR were carried out according to the SureSelect^{XT} Methyl-Seq Illumina Multiplexed Sequencing Protocol (version B, January 2013). Amplified libraries were then indexed using 6 PCR cycles, quantified and pooled as per the standard protocol. Sequencing was again carried out on an Illumina HiSeq 2000, using version 3 chemistry, generating 2 x 100bp paired end reads (one strand only). Sample growth and sequencing was performed by the CGR.

The sequencing datasets that were generated for the 6 samples were mapped to the extended methylation bait sequences using Bismark. Bismark is an aligner and methylation caller designed specifically for bisulfite treated sequence data (Krueger and Andrews, 2011). During mapping of sequencing reads a mismatch number of 3 was used per read and the non-directional nature of the library was specified. The Bismark methylation extractor tool was then used to identify all cytosine residues within the mapping and categorize the reads mapping to them as un-methylated or methylated at that position while also detailing which type of potential methylation site was present (CHH, CHG or CpG). The mapping results generated by Bismark also come in the form of a SAM file. This allowed mapping results also to be processed using SAMtools to remove: any non-uniquely mapping reads, unmapped reads and duplicate reads. SNP calling could then be carried out using SAMtools mpileup that was implemented prior to VarScan, to loosely identify positions containing an alternate allele, with a minimum coverage of 6, an average mapping quality above 15 and a MAF of greater than 0.1. This SNP calling plus Bismark analysis allows sequencing reads and therefore cytosine residues to be assigned to genomes for downstream methylation analyses.

The six samples (sample 12B, 12C, 12D, 27B, 27C and 27D) that were enriched using the array, bisulfite treated and then sequenced had an average depth of coverage of ~102 across 96.3% of the 6Mb array probe sequence (see table 5.2 for full details).

Sample	Mean % coverage per reference probe	Mean depth of coverage per reference probe	Number of Probes mapped (50000 total)	% Of reference probes mapped
12B	97.2	89.8	49838	99
12C	96.9	70.4	49692	99
12D	97.9	138.5	49928	99
27B	97.1	80.9	49798	99
27C	97.2	96.3	49807	99
27D	97.9	135.8	49917	99

Table 5.2. Mapping Statistics for six enriched and bisulfite treated wheat DNA samples. All mapping statistics in relation to the 6Mb array probe base-space

Mapping statistic and SNP comparisons between the repeat samples again revealed extensive conservation between 12B, 12C and 12D and between 27B, 27C and 27D (~100% conservation), in addition methylation statuses were highly conserved (100% conservation using threshold values). Therefore sequence datasets for all of the 12°C sample replicates and also for the 27°C sample replicates were merged and used in downstream analyses. This resulted in an average depth of coverage of 297.6x with ~97.5% of the 6Mb array probe sequences being mapped to (detailed in table 5.3). An average depth of coverage of ~128x was observed with ~62% of the ~44Mb mapping reference being mapped to across both samples i.e. the mapped region extended into surrounding next-to-target regions covering over 4x the 6Mb capture probe space. If, at a particular cytosine residue, the percentage methylation of a genome was >=15% different between repeat samples then this site was recorded as a possible site of ‘background methylation’ i.e. sites containing noise that could therefore easily be flagged in downstream analyses. This ‘background methylation’ accounted for less than 2% of the cytosine residues that were analyzed and could reflect noise i.e. poorer quality reads/mapping or regions showing tissue specific methylation.

Sample (°C)	Average % coverage of reference probe	Average depth of coverage per reference probe	Number of Probes mapped	% Of reference probes mapped
12	98.4	290.8	49986	99
27	98.5	304.4	49982	99

Table 5.3. Overall Mapping Statistics in Bisulfite treated data. Detailing the mapping output statistics for 2 enriched and bisulfite treated wheat DNA samples in relation to the 6Mb array probe base space

The methylation array was again seen to effectively enrich ~98% of the target 6Mb plus an additional ~18Mb of sequence in its immediately surrounding regions even with use of the bisulfite treated DNA that is more difficult to map and due to the nature of the treatment is highly degraded. Looking at all analyzed sites, 86,192 for the 12°C sample plus 94,363 for the 27°C sample (detailed in section 5.5), the average number of reads per genome could be calculated across the 2 samples; 32-33% for genome A, 34-35% for genome B and 33% for genome D. We see approximately 1/3 of the reads assigned to each genome so the bait probes enrich all 3 genomes effectively and consistently.

Over the 2 samples on average 31,939,028 (~20%) sequencing reads in each case were mapped to the reference sequence. It is estimated that ~63% of off target sequencing reads include repetitive sequence. The array probe set was re-designed based on this analysis to remove baits exceeding 10x the average depth of coverage i.e. potentially enriching off-target material. Future enrichment analyses using the re-designed baits are predicted to allow generation of an additional ~15% on target sequencing reads and enrichment protocol development is likely to increase this figure further.

5.5 Determination of the available dataset for analysis

In the 12°C sample Bismark identified 7,813,105 cytosine residues (methylated and un-methylated) in the sequencing data; 8,069,906 were identified in the 27°C sample. These numbers vary between datasets due to slight differences in the mapping coverage of the reference. In the 12°C sample ~37% of cytosine residues were predicted to be transcribed; this is in proportion to the ~35% of the full 44Mb mapping reference that is predicted to be transcribed. Considering fully or partially methylated cytosine residues only, a similar proportion were seen in transcribed/un-transcribed regions with ~32% thought to be transcribed.

To assign a cytosine residue to the A, B or D wheat genome an individual sequencing read must contain the cytosine residue plus a homeologous SNP allele and 38,384 homeologous SNP positions were identified for this purpose in section 5.3. The workflow detailed in figure 5.4 outlines how the 7,813,105 and 8,069,906 cytosines in the 12°C and 27°C samples respectively could be interrogated to produce a list of 86,192 residues for the 12°C sample and 94,363 for the 27°C sample where; using homeologous SNPs all 3 genomes could be identified at a depth of 5 or greater per genome in the mapped reads and where the 3 genomes' methylation patterns were identical across the 3 repeat (B, C and D) samples making up the dataset i.e. no 'background methylation'. In both cases alleles representing each of the 3 genomes could be identified in ~66% of methylation sites with a depth of coverage over 20. In both samples ~72% of these 86,192 and 94,363 residues were found to be in transcribed regions. This ~72% is high, considering that ~35% of the full 44Mb mapping reference is predicted to be transcribed, and most likely reflects lower mapping coverage in non-transcribed regions, as such transcribed regions are more likely to gain sufficient coverage for identification of homeologous SNP alleles and be preferentially selected to allow methylation site association with them. Lower mapping coverage tends to be gained in non-transcribed regions due to their comparatively repetitive nature causing sequencing reads to map non-uniquely i.e. multiple times and be removed from analyses (Jiang and Goertzen, 2011).

Stage	Feature	Proportion Transcribed/Non-Transcribed
Stage 1 Reference & SNP list ↓	Reference sequence 44Mb 38,384 Homeologous SNPs	35%/65% 61%/39%
Stage 2 Within mapped sequencing data ↓	7,813,105 cytosine residues (8,069,906) 11,877 Homeologous SNPs (12,092)	37%/63% (37%/63%) 67%/33% (67%/33%)
Stage 3 Associated data ↓	510,619 cytosine residues associated with a homeologous SNP (522,358)	65%/35% (65%/35%)
Stage 4 Final dataset: -3 genomes mapped 5x or more -methylation identical in B, C and D repeat samples	86,192 cytosine residues (94,363)	72%/28% (72%/28%)

Figure 5.4. Determination of the subset of data that was available for detailed analysis in the 12°C sample. Corresponding numbers in the 27°C sample shown in brackets. At each stage the percentages of residues/SNPs that are transcribed/non-transcribed are detailed. **Stage 1** describes the mapping reference and the reference homeologous SNP list developed in section 5.3. **Stage 2** Shows all cytosine residues and reference homeologous SNP positions within the mapped 12°C /27°C sample datasets. In **Stage 3** the outputs from **stage 2** are combined to identify all cytosines that could be associated with homeologous SNP locations in each of the datasets. In **stage 4** the final analyzable dataset; cytosine residues were selected if all 3 genomes could each be identified at a depth of 5x or more in the mapped reads and if the 3 genomes' methylation status were identical in the 3 repeat (B, C and D) samples making up the dataset.

The list of 86,192 residues for the 12°C sample and 94,363 for the 27°C sample were produced as follows (figure 5.4); Cytosine residues that could be associated with homeologous SNP locations included 510,619 residues in the 12°C sample (65% in transcribed regions) and 522,358 residues in the 27°C sample (65% in transcribed regions). 11,877 homeologous SNP positions were conserved in the 12°C sample sequencing data and used for this cytosine association (12,092 positions in the 27°C sample). In both samples ~67% of the SNP positions were predicted to be in transcribed regions. Finally, cytosine residues with sufficient per genome mapping coverage and identical methylation patterns across the 3 repeat (B, C and D) samples could be selected leaving the 86,192 residues for the 12°C sample and 94,363 for the 27°C sample.

5.6 Identification of global methylation patterns

The subset of wheat sequence that was analyzed here, representing an unbiased selection of gene rich genomic DNA sequence, should represent the patterns seen across the entirety of the wheat genome's gene regions and is potentially comparable with methylation patterns seen in other organisms with high genic content such as the ~90% non-repetitive gene rich *Arabidopsis thaliana* (Lister *et al.,* 2008). Widman, N. *et al.* performed a comprehensive study of methylation patterns in the model plant genome *Arabidopsis.* This study implemented detailed knowledge of the plant to define a cytosine residue as methylated if 80%, 25% or 10% or more of the total number of sequencing reads were methylated for CpG, CHG and CHH sites respectively. Here, looking at all cytosine residues and defining methylation using the same cutoffs as Widman, N. *et al.* 32% of the residues that were CpG sites were methylated, 15% of CHG sites were methylated and 11% of CHH sites were methylated. Widman, N. *et al.* saw a similar pattern, at slightly lower levels, in Arabidopsis where 24% of CpG sites, 7% of CHG sites and 2% of CHH sites were methylated. If the under-represented vast repetitive regions of wheat were also taken into account we are likely to see a higher proportion of methylated sites overall since methylation associates strongly with repetitive regions (Zhang *et al.,* 2006; Widman *et al*., 2009).

Although the implementation of Widman, N. *et al.*'s methodology is useful within this study constant standard thresholds for methylation identification were utilized across CpG, CHH or CHG methylation sites given the relatively unknown expectations of the wheat methylome and to allow comparison between CpG and non-CpG site methylation (Harris *et al.,* 2010; Widman *et al*., 2009). Thresholds were implemented on a genome-by-genome basis to allow discrimination of sites where only one genome was methylated. Using these standard thresholds and looking at all analyzed residues (86,192 in the 12°C sample and

94,363 in the 27 °C sample); ~10% of residues under analysis were CpG sites with 52% that showed one or more genome to be methylated, ~22% of residues were CHG sites with 3% methylated and ~68% of residues were CHH sites with ~3% methylated. These percentages of CpG/CHG/CHH residues follow a similar pattern to that observed across the entire dataset, as such, the subset for analysis are representative of the whole dataset in this respect.

This same analysis was repeated focusing on sites that showed one or more genome to be methylated; distribution varies slightly between transcribed regions (~64% CpG methylation, ~27% CHH and ~9% CHG) and non-transcribed regions (~55% CpG methylation, ~34.5% CHH and ~10.5% CHG) within the subset. CpG sites consistently account for the majority of methylated sites, although there is a slight reduction in CpG site methylation and an increase in CHH/CHG site methylation in non-transcribed sites compared to transcribed sites. In previous studies in plants such as *Arabidopsis* (Lister *et al.,* 2008; Glaus *et al.,* 2012) CpG sites tend to be seen almost exclusively in coding regions while CpG along with CHG and CHH sites typically only seen in non-coding regions. Here, all three types of methylation are a significant presence in both transcribed and non-transcribed regions although an increase in CHH/CHG site methylation was observed in non-transcribed sites.

5.6.1 Identification of genome specific methylation/non-methylation in wheat

~5% of the cytosine residues that were analyzed overall contained methylation (100% threshold) in 1 or more genomes. Differential methylation has been observed between the A, B and D genomes in 27% of analyzed methylated cytosine residues, in each of the 12°C and 27°C samples, and was recorded at a minimum difference of 50% (for full list of differentially methylated residues for sample 12 see Appendix 4, table 1).

Table 5.4 details the breakdown of the genome specific methylation into relevant genomes, transcription status and finally type of methylation (CpG, CHH or CHG) for the 12°C and 27°C samples while figure 5.5 shows a visual representation of this data using the data averaged between the 12°C and 27°C datasets due to consistently high similarity overall between them across the three genomes.

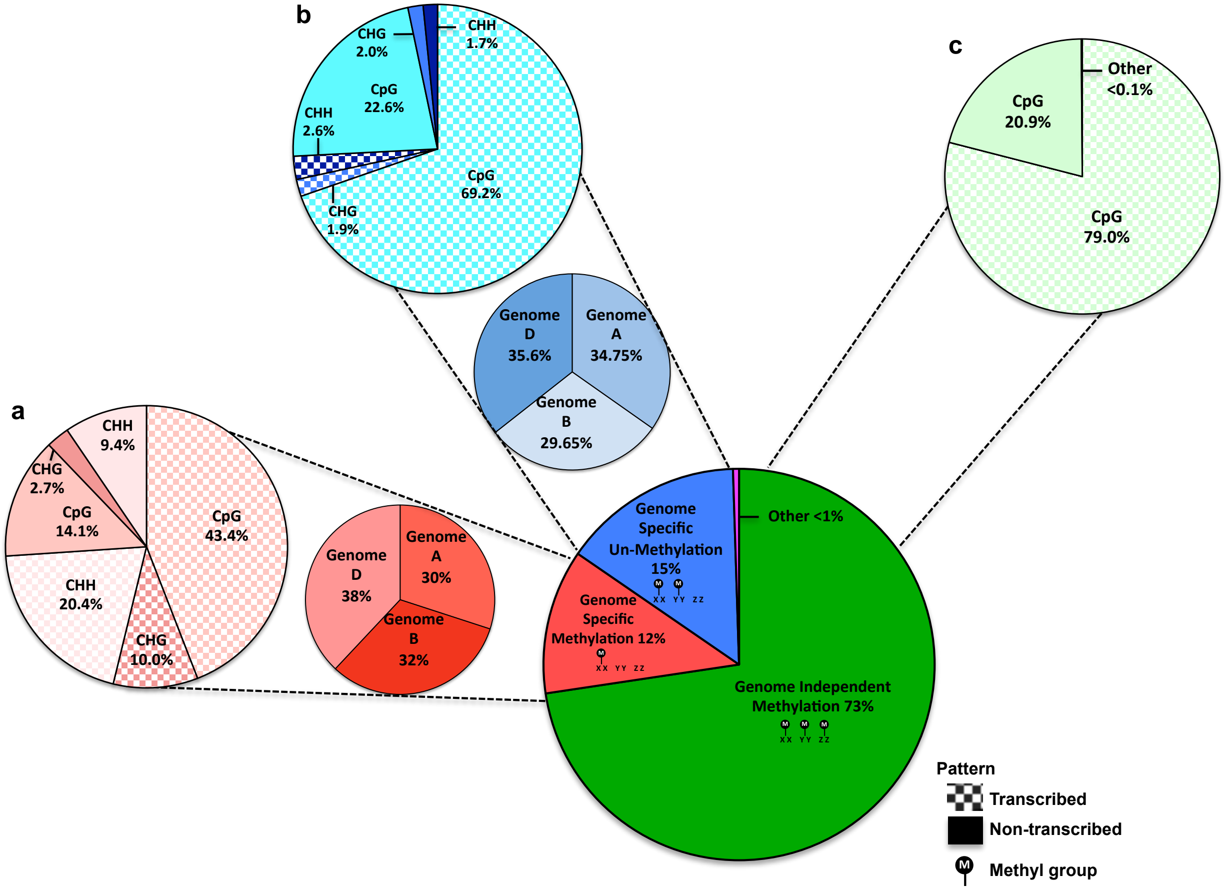


Figure 5.5. Categorizing observed methylation averaged across the 12°C and 27°C sample datasets. **(a)** Genome specific methylation occurrences broken down into distribution between the 3 genomes and finally an overview of its division between CpG, CHG and CHH sites in transcribed and non-transcribed regions (averaged over all 3 genomes due to high similarity). **(b)** Genome specific non-methylation occurrences broken down identically to **(a)**. **(c)** Genome independent methylation and its division into CpG, CHG and CHH sites in transcribed and non-transcribed regions.

	Sample 12			Sample 27		
Genome Specific Methylation	**A**	**B**	**D**	**A**	**B**	**D**
Transcribed (%)	60	76	81	54	78	74
% CpG methylation	61.5	58.0	67.8	47.6	56.2	73.3
% CHH methylation	27.0	29.0	18.6	42.9	27.6	20.0
% CHG methylation	11.5	13.0	13.6	9.5	16.2	6.7
Not Transcribed (%)	40	24	19	46	22	26
% CpG methylation	43.0	57.0	57.1	47.0	55.4	62.0
% CHH methylation	51.0	34.0	14.3	47.0	33.7	19.0
% CHG methylation	6.0	9.0	28.6	6.0	10.9	19.0
Genome Specific Non-Methylation	**A**	**B**	**D**	**A**	**B**	**D**
Transcribed (%)	61	78	76	64	77	69
% CpG methylation	92.1	95.2	89.2	94.0	95.0	87.1
% CHH methylation	6.6	2.4	6.2	5.0	3.0	5.7
% CHG methylation	1.3	2.4	4.6	1.0	2.0	7.2
Not Transcribed (%)	39	22	24	36	23	31
% CpG methylation	81.6	92.7	76.2	82.2	91.5	64.5
% CHH methylation	6.1	5.2	9.5	4.5	5.7	12.9

Table 5.4. Summary of orientation of the methylation sites that were analyzed in sample 12 (12°C) and sample 27 (27°C). Breakdown of methylation into genome specific and independent sites; transcribed/non-transcribed regions and CpG/CHH/CHG sites.

In the 12°C sample; 31.8% of genome specific differential methylation was from genome A, 31.9% from genome B and 36.3% from genome D. In the 27°C sample; 28.1% of the differential methylation was from genome A, 32.1% from genome B and 39.8% from genome D. In both cases the three percentages were found to be significantly different to the expected percentage of 1/3 (Sample 12; $X^2 = 6.579$, $p = 0.0373$ and Sample 27; $X^2 = 34.793$, $p < 0.0001$). Genomes A and B show genome specific methylation in relatively similar proportions whilst Genome D contains the most overall. All figures are normalized to account for bias in the SNP numbers used to associate sequencing reads with each genome. The same pattern could not be distinguished in the genome specific non-methylated group (one genome is un-methylated while the other two are methylated) where, across the two samples, distribution between the three genomes was without a visible significant trend.

Genome specific methylation is mostly at CpG sites (~57%) although CHH and CHG sites are still a significant presence (~30% and ~13% respectively) across the three genomes, across the 12°C and 27°C samples and whether transcribed or non-transcribed. These figures closely resemble those seen over all methylated sites and there does not appear to be a bias in the regions selected for genome specific methylation. Genome specific non-methylation yields similar results with a less prominent CHH and CHG site presence. Across the three genomes, in both samples and over transcribed and non-transcribed regions genome specific non-methylation was more prominently at CpG sites (~87%) with a lower but significant presence at CHG (~7%) and CHH (~6%) sites. These figures do deviate from those expected within the dataset looking at all methylated sites. There does seem to be a stronger CpG bias and CHH/CHG reduction in methylation sites within this dataset when considering genome specific non-methylation and this is conserved across transcribed and non-transcribed regions.

No difference is seen consistently between transcribed and non-transcribed regions when looking at genome specific patterns as a whole. Both samples behave in a similar way, as do the three genomes. CpG genome specific methylation is preferred while CHG and CHH methylation sites are a significant presence. Genome specific non-methylation has a stronger bias for CpG methylation and a lower level of CHH/CHG sites. Figure 5.6 represents the distribution of methylated residues along the pseudo-chromosome sequences for (a) the 12°C sample and (b) the 27°C sample. Sites are relatively evenly distributed across this, the genic portion of the genome, with most gaps due to missing information for analysis in a region rather than a break in methylation.

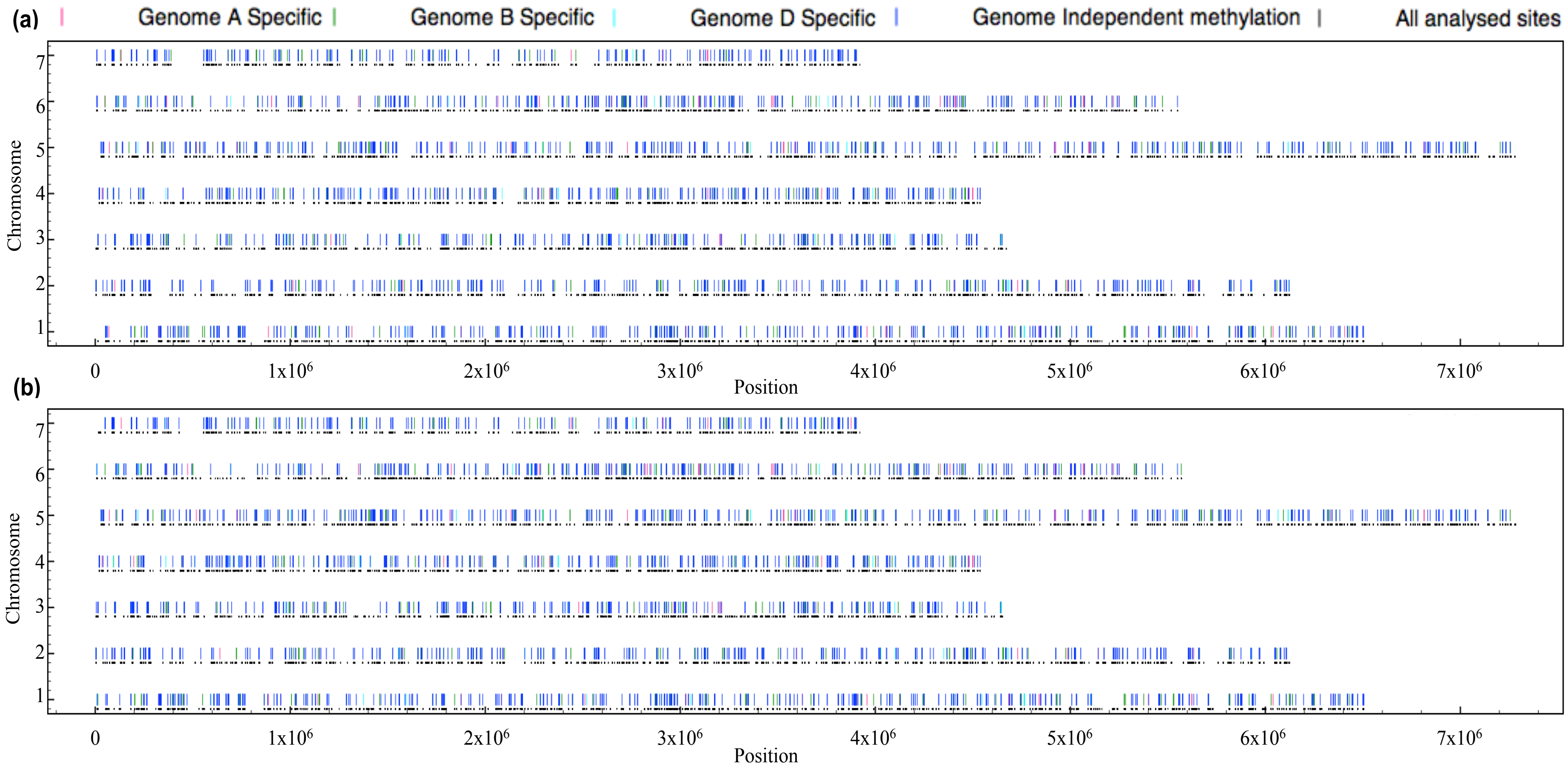


Figure 5.6. Positional information for methylation sites. Incidences of genome specific methylation/genome specific non-methylation between the 3 wheat genomes and genome independent methylation are detailed relative to all analyzed sites. Data is shown along each of the pseudo-chromosomes using threshold values. **(a)** the 12°C sample **(b)** the 27°C sample

GO enrichment using the program GOEAST was performed on these differentially methylated 12°C sample residues by Jonathan Price. This highlighted common functional terms associated with the differentially methylated sites of the A, B and D genome's (p <0.01) (Appendix 4, table 2). The A genome's enriched genes tended to relate to signaling pathways, metabolic processes and response to water stimulus. The B genome showed enrichment for terms such as: biosynthetic/metabolic processes, RNA splicing and protein de-phosphorylation. Finally the D genome's enrichment profile contained terms involved in chromatin silencing, histone modification and methylation/regulation of gene expression and chromosome organization. Extended bait probe sequences were ranked based on the number of differentially methylated sites per bp across each, this identified highly differentially methylated regions. All of the GO terms that were associated with the top 10% of contigs in this list correlated with the GOEAST result highlighting its capability to identify highly differentially methylated genes. The GO enrichment analysis was repeated for the 27°C sample and the enriched GO terms were compared with those in the 12°C sample; 78% of terms found in the 12°C sample genome B were found in the 27°C sample genome B and 100% of those found in the 12°C sample genome D were found in the 27°C sample genome D. No significant enrichment could be confirmed in genome A to allow comparison. This demonstrates highly conserved genome specific methylation between the two samples.

For each instance of genome specific methylation (A, B and D) the average depth of coverage of the methylated genome across the dataset was calculated and compared to the average coverage of the other two non-methylated genomes at each position. This analysis was conducted across the two samples and depths at each genome were normalized to account for the minor deviation from 1/3 reads per genome that was seen in the dataset. The mean coverage's in the non-methylated group (mean=99.2, SD=16.3) were compared to those of the methylated group (mean=122.7, SD=18.5) and no significant difference was found (p=0.164, two-tailed t test). Therefore there was no issue with bias towards enrichment of methylated or un-methylated DNA sequence.

5.6.2 Identification of genome independent methylation

Across the 12°C and 27°C sample datasets over 98% of the cytosine residues that were analyzed displayed genome independent behavior i.e. methylated or un-methylated status conserved across the three genomes. Of these genome independently methylated or un-methylated residues ~4% in each sample could be defined as showing genome independent methylation i.e. all three genomes methylated. This ~4% accounted for ~73-74% of all analyzed residues that contained methylation in one or more genomes (detailed for the 12°C sample in Appendix 4, table 3). In table 5.4 and figure 5.5 it can be seen that in the 12°C and

27°C samples, across all three genomes, most of the genome independent methylated residues are, irrespective of transcription status, almost exclusively CpG sites (>99%). This differs from the observation looking at genome specific associations. Figure 5.6 represents this genome independent subset of methylated residues for (a) the 12°C sample and (b) the 27°C sample and outlines their relative positions along the pseudo-chromosomes; again they are approximately evenly distributed along the chromosomes.

Furthermore, for both samples, most residues displaying genome specific methylation are in transcribed regions (~74%) as are those displaying genome independent methylation (~79%) and those displaying genome specific non-methylation (~71%) (table 5.4). None of these three proportions deviate by more than 10% from the ~72% of residues that were found to be in transcribed regions in the full list that they were derived from of 86,192 residues for the 12°C sample and 94,363 for the 27°C sample. We can therefore conclude that genome specific/genome independent methylation does not appear to target transcribed/non-transcribed regions and is found in these regions in the anticipated proportions.

5.7 Transposon and chloroplast methylation state assessment through the analysis of off target sequence

In theory the majority of the sequencing reads in the enriched dataset should map to the extended bait reference sequence. Elimination of repetitive sequence from the capture probes and high specificity of the long 120bp sequences should stop enrichment of off target repetitive regions. It is however possible to carry over contamination of enriched sequence with non-enriched repetitive sequence that can then appear in the sequencing data. This is a particular problem in wheat where, due to its high repetitive sequence content, most of the contaminating non-enriched sequencing reads are likely to represent repetitive sequence. If this is the case the contamination carry-over will be random and as such the repetitive sequence diversity should be representative of that seen in the sequencing analysis of non-enriched wheat by Brenchley *et al. (*Brenchley *et al.,* 2012).

The unmapped sequencing data (non-bisulfite treated) for the 12°C sample, ~80% of the sequencing data, was analyzed to see if repetitive sequence could be identified/categorized. The unmapped sequencing reads were used in a BLAST alignment to the wheat TREP repetitive sequence database (e-value of 1e-5). This allowed comparison of it to the repetitive sequencing data that was seen in non-enriched wheat by Brenchley *et al.* Jonathan Price carried out this comparison and no notable proportional deviations were observed between the 2 datasets (see table 5.5) i.e. there was no bias introduced for specific repetitive regions by enrichment. The percentage of sites in each transposon type showing 100% threshold methylation in one or more genomes was also calculated using Bismark to map the 12°C sample's previously un-mapped bisulfite treated data to the TREP database sequences. It was noted that transposons in general were hyper-methylated, typically >25% sites showed 100% methylation in one or more genomes, in comparison to gene regions where we see such methylation in ~5% of the sites that were analyzed. The retro transposon group SINE was the only exception to this. In an additional comparison, using Bismark to map the 12°C sample's un-mapped bisulfite treated data to the wheat chloroplast genome (Middleton *et al.,* 2014), that is thought to be generally un-methylated (Kovarik *et al.,* 2001); less than 0.0% of sites showed methylation at the 100% threshold i.e. the bisulfite treatment was effective and the rate of incomplete conversion was low allowing discrimination of this known feature.

Type	No. of reads	Percentage of total (%)	Brenchley *et al.* Percentage of total (%)	Average Methylation (%)
DNA transposons	**1408434**	**23.669**	**18.691**	**26.78**
Helitron	2319	0.039	0.303	23.81
TIR	1403061	23.579	18.311	25.93
HAT	1034	0.017	0.052	24.11
Harbinger	73350	1.233	0.427	25.29
Marnier	140677	2.364	0.128	30.68
CACTA	1082075	18.185	15.995	24.94
Mutator	105743	1.777	0.557	24.04
Unknown	191	0.003	0.011	26.54
Unknown	3054	0.051	0.077	26.78
Retro transposons	**4235099**	**71.172**	**79.779**	**22.01**
SINE	151	0.003	0.005	0
LINE	297416	4.998	1.026	23.73
LTR	3937532	66.172	78.748	23.07
Gypsy	2815602	47.317	44.034	21.66
Copia	940479	15.805	17.394	21.5
Unknown	181451	3.049	1.490	26.06
Unknown	**306946**	**5.158**	**1.530**	27.21
	16934614			

Table 5.5. Repeat composition of the 12°C sample's unmapped sequencing reads. Enriched sequencing reads from the 12°C sample that did not map to the reference sequence (non-bisulfite treated) were aligned using BLASTN to the TREP repeat database and the number of read matches are detailed here with the % of the total transposon database hit. The same analysis performed by Brenchley *et al.* on non-enriched wheat is also shown. Average % methylation was calculated for each transposon type using Bismark to map the bisulfite treated data to it i.e. the % of sites that showed 100% methylation in one or more genomes.

5.8 Investigating temperature dependent differential methylation and gene expression

78,628 cytosine residue sites, from the subset of sites for which three genomes could be confidently identified, are conserved between the 12°C and 27°C samples. Using threshold values the methylation statuses of these sites were compared to identify any differential methylation between the two samples in one, two or all of the three genomes and no sites could be found. Methylation status between the two samples was highly conserved. 24 sites were identified where differential methylation of 15% or more could be seen between the two samples in one or more of the three genomes. This number was reduced to 23, as sites also which showed 15% or more difference between the three repeats within either the 12°C or 27°C sample were removed i.e. 'background methylation'. This analysis only considered differential methylation between the 12°C and 27°C samples that was conserved over the three replicate samples that each dataset was made up of and sites are detailed in full in Appendix 4, table 4. Of the 23 sites identified, genome independent differences between the two samples were exceptionally rare with none seen in this analysis. Two cases were seen where two genomes were differentially methylated between the two samples leading us to believe that genome independent differential methylation may be possible, however, unusual, and most likely missed in the small dataset observed. Genome specific methylation variation between the two samples was observed the most commonly with 6 cases affecting genome A, 13 cases affecting genome B and 2 cases affecting genome D.

In an independent analysis by Mark Quinton-Tulloch RNA-seq data was generated and analyzed for the 12°C and 27°C samples to allow determination of gene expression across the dataset and to allow comparison with methylation i.e. cDNA for the two wheat samples was generated and sequenced using the Illumina HiSeq. This RNA-seq data was run through the BitSeq pipeline (mapping using Bowtie2) that estimates transcript expression level using a Bayesian approach (Glaus *et al.,* 2012). The PPLR values that were generated detailed changes in gene expression levels between the 12°C and 27°C samples per extended methylation array bait sequence. Information was also generated per sample on each genomes percentage contribution to overall gene expression per extended methylation array bait sequence using homeologous SNPs. The expected average contribution per genome was ~1/3 unless genome specific differences in gene expression were seen. Here, for all cytosine residues under analysis for the 12°C sample the per genome percentage contributions to gene expression, from the extended bait sequence that the residue was found in, i.e. the gene associated with the residue, could be identified and are detailed in figure 5.7 (frequency of residues per 5% contribution interval are calculated).

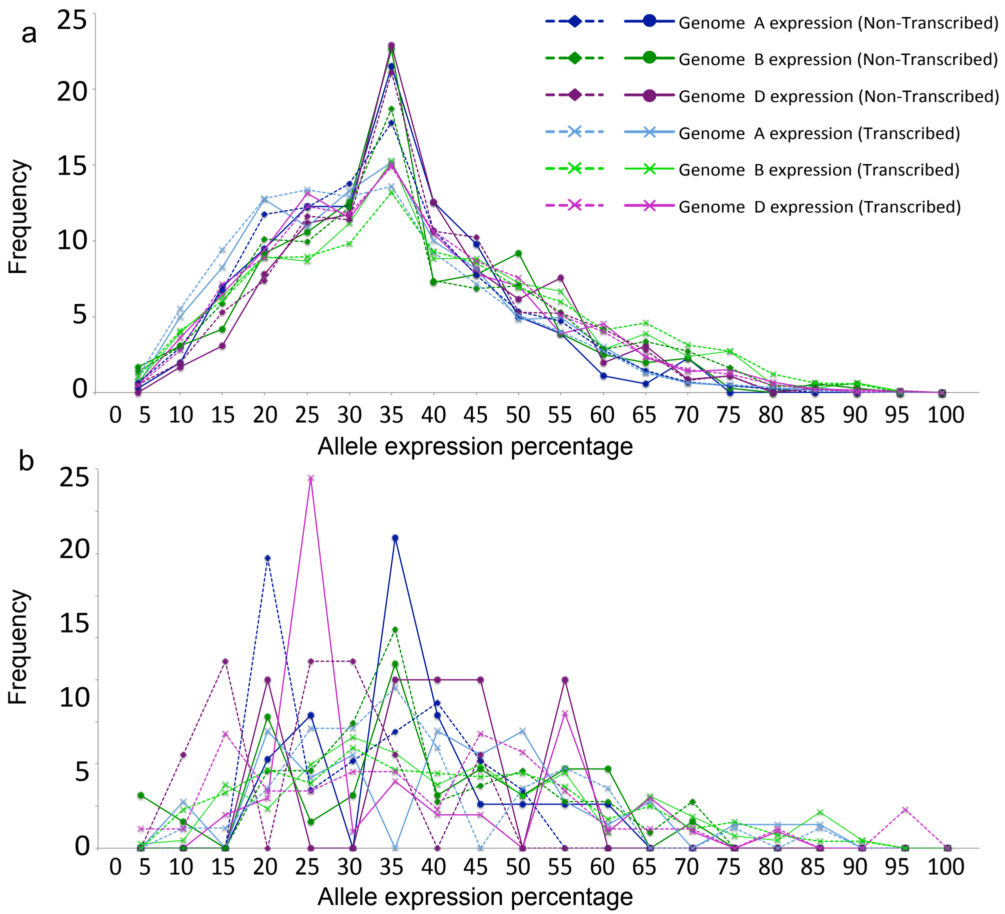


Figure 5.7. Frequency plot of percentage per genome contribution to overall gene expression. Percentage contribution per genome to gene expression overall for the contig that the cytosine residue originated from. Frequency of sites calculated within each 5% interval in the range 0-100%. **(a)** For every genome independently methylated site (dashed line) and genome independently un-methylated site (non-dashed line) the percentage contributions of each of the 3 genomes to gene expression overall were retrieved for the extended bait sequence that the site originated from and for each 5% interval the number of sites were calculated. **(b)** For each site showing either genome specific methylation (non-dashed line) or genome specific non-methylation (dashed line), for the genome concerned only, the percentage contribution of that genome to gene expression overall was retrieved for the extended bait sequence that the site originated from. For each 5% interval the number of sites were calculated and are plotted (normalized numbers of A, B and D specific methylation are used to allow comparison).

It was noted from figure 5.7 that the number of residues represented in figure 5.7a (over 3000 genome independently methylated residues and over 80,000 genome independently un-methylated residues) was significantly higher than those represented in figure 5.7b (~1100 genome specific sites overall). In order to create a fairer comparison of contribution to gene expression between genome specific and genome independent sites figure 5.8 was developed. This figure demonstrates the same data using a comparable sized subset of methylated/un-methylated genome independent residues to the genome specific subset of residues (all genome specific and genome independent residues and their raw % contributions to gene expression for the gene that the residue is associated with can be found in Appendix 4, table 5 and 6 respectively).

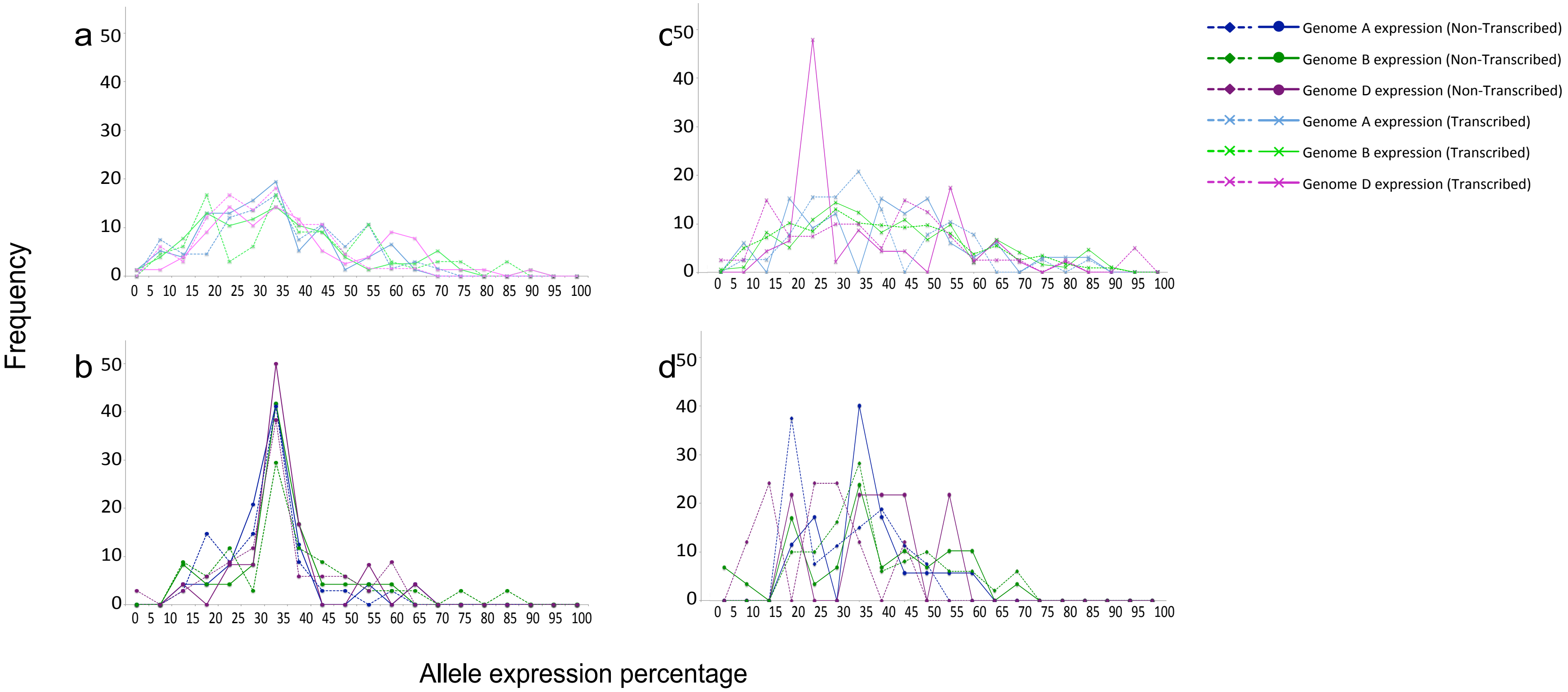


Figure 5.8. Frequency plot of percentage per genome contribution to overall gene expression (subset). Percentage contribution per genome to gene expression overall for the contig that the cytosine residue originated from. Frequency of sites calculated within each 5% interval in the range 0-100%. See figure 5.7 for full annotation. **(a)** Plots for each of a subset of 100 genome independently methylated sites (dashed line) and genome independently un-methylated sites (non-dashed line) transcribed residues only. **(b)** See (a) but for non-transcribed regions only. **(c)** Plots for each site showing either genome specific methylation (non-dashed line) or genome specific non-methylation (dashed line), for the genome concerned only (normalized numbers) and for transcribed residues only. **(d)** See (c) but for non-transcribed regions only.

Figure 5.8b demonstrates that gene expression percentages that were associated with genome independently methylated/non-methylated non-transcribed sites in the region tend, with an approximately normal distribution, towards a peak in the anticipated 30-35% interval. Here methylation in a genome independent manner, if affecting gene expression, appears to affect all three genomes so that they contribute equally to the overall expression profile at that position. In contrast to this, gene expression that is associated with a cytosine residue in genome specific methylated or non-methylated non-transcribed regions (figure 5.8d) is far more variable with no consistent trends seen and mainly multimodal distributions. It is likely that genome specific changes in methylation status in non-transcribed regions may skew that genomes contribution to overall gene expression from the expected 30-35% interval. For these skewed peaks, where possible, associated GO terms that deviate from those seen across all sites in the 30-35% interval are detailed in Appendix 4, table 7 and in genome A are linked to low expression of genes linked to phosphorylation, carboxylase, synthase, O-acyltransferase, ATPase, protein kinase and transglucosylase activity and high expression of genes with arabionsyltransferase and auxin transporter activity. In genome B low expression is linked to genes involved in transmembrane transporter activity and finally in genome D low expression is linked to genes with ligase/catalytic activity and damaged DNA binding.

In transcribed regions gene expression data tends towards a more platykurtic set of distributions when associated with both genome independent methylated/non-methylated cytosines (figure 5.8a) and genome specific methylated or non-methylated cytosines (figure 5.8c) except for the addition of a clear peak in the genome specifically methylated cytosine subset for genome D showing an average associated gene expression level with a left skew tending towards ~20-25% i.e. a conserved decrease in expression correlating with methylation of this genome and a smaller right skewed peak tending towards ~50-55%. Again for these 2 peaks, where possible, associated GO terms that deviate from those seen across all sites in the 30-35% interval are detailed in Appendix 4, table 7. The peak in the 20-25% interval was associated with monooxygenase activity and the peak in the 50-55% interval associated with ATP-dependent DNA helicase activity.

The poorly understood downstream effect of methylation of the different gene body regions makes interpretation of such data difficult, particularly with the added complication of a poorly annotated genome. With only a reliable classification of analyzed sites into transcribed and non-transcribed regions and, as of yet, no further information regarding promoter, first exons etc. we could reliably see both increased and decreased expression profiles associated with methylation in transcribed regions (in the first exon decreasing expression or in a downstream exon potentially associating with increased expression) and likewise in non-transcribed regions (in promoter regions decreasing expression or in downstream introns possibly associating with increased expression).

Mark Quinton-Tulloch's RNA-seq data for the 27°C sample was used for a comprehensive comparison of gene expression between the two samples i.e. investigate if temperature dependent differential methylation could be correlated with gene expression. The PPLR values that were generated from his gene expression analysis detailed changes in expression levels between 12°C and 27°C samples for each genome individually. A low PPLR tending towards 0 indicates down regulation of a gene in the 27°C sample compared to the 12°C sample and a high PPLR value tending towards 1 indicated up-regulation of the 27°C sample compared to the 12°C sample. These values allowed confident analysis of the gene expression status of the 23 sites that had sufficient mapping of RNA-seq data and also showed differential methylation between samples. 16 sites showed PPLR values for the genome that was differentially methylated, that deviated from the baseline 0.5 by +/- 20%; these show gene expression changes that could potentially be caused by differential methylation and are detailed in table 5.6 with syntenic genes and GO annotations.

Methylation in promoter and first exon regions has been associated with silencing of a gene, therefore hyper-methylation in transcribed and non-transcribed regions leading to a decrease in expression and hypo-methylation leading to an increase in expression could be indicative of first exon or promoter regions. Methylation increase in internal introns and exons has been closely correlated and linked to increased gene expression (Zhang *et al.,* 2006; Brenet *et al.,* 2011). In table 5.6 9 up-regulated genes in the 27°C sample are linked to differential methylation; 3 were of unknown function and likely to be involved in protein binding, 1 was linked to stress response due to the presence of stress response elements (2 sites correlated with this gene), 1 gene with amine receptor activity thought to be involved in binding proteins, 1 gene was found to be similar to the PRP5 heat shock protein in a BLAST alignment, 1 gene had ATPase activity, 1 gene had potassium ion transmembrane transporter activity and finally 1 gene was related to putative glycine hydroxymethyltransferase. 7 down-regulated genes in the 27°C sample were linked to differential methylation; 3 of unknown function and likely to be involved in protein binding, 1 with similarity to the vacuolar ATPase B subunit that has interestingly been linked to tissue specific transcript level decrease under heat stress (Kluge *et al.,* 2003), a gene likely to encode a frigida-like protein found typically in winter accessions to prevent flowering until after winter (Risk *et al.,* 2010), a gene with glutamate synthase activity and finally a residue that was related to the F-box protein that has been up-regulated under cold stress (Jain *et al*., 2007).

Contig: Position	Associated *Brachypodium* gene	Associated GO term	PPLR value	Gene expression (sample 27°C)	Methylation status (sample 27°C)	Region*
CONTIG216081_150-888-608:726	Bradi2g46350 F-box protein	-	0.242	Down-regulation	Hypo-methylated genome B	T
CONTIG26428_1-2754-2097:2211	Bradi2g40770 amine receptor activity	GO:0005515 protein binding GO:0035091 phosphatidylinositol binding	0.801	Up-regulation	Hyper-methylated genome D	T
CONTIG3867046_1-404-130:254	Bradi1g66670 PRP5 heat shock related protein	-	0.676	Up-regulation	Hyper-methylated genome B	T
CONTIG606372_1-981-385:485	Bradi1g15830 Stress response	-	0.607	Up-regulation	Hypo-methylated genome B	T
CONTIG606372_1-981-385:510	Bradi1g15830 Stress response linked	-	0.607	Up-regulation	Hypo-methylated genome B	T
CONTIG79256_1-2552-86:249	Bradi2g48082 similarity to putative vacuolar ATPase B subunit	-	0.277	Down-regulation	Hyper-methylated genome A	T
CONTIG829414_1-1110-466:560	Bradi1g55530	GO 0003755 stress related peptidyl-prolyl cis-trans isomerase activity ras GTPase binding	0.38	Down-regulation	Hypo-methylated genome A	T
CONTIG1004297_1-1020-93:210	Bradi4g08780 Frigida like protein	-	0.101	Down-regulation	Hyper-methylated genome B	T
CONTIG1254038_1-910-552:542	Bradi4g14900	GO 0003723 Motor activity and RNA binding	0.695	Up-regulation	Hyper-methylated genome B	NT
CONTIG150583_1-2126-197:157	Bradi4g06350	GO:0005515 protein binding	0.112	Down-regulation	Hyper-methylated genome B	NT
CONTIG1541763_1-805-627:635	Bradi1g19080 glutamate synthase (NADPH) activity	GO:0005515 protein binding	0.073	Down-regulation	Hyper-methylated genome B	T
CONTIG171513_232-1141-280:399	Bradi1g09327 putative glycine hydroxymethyltransferase	-	0.624	Up-regulation	Hypo-methylated genome B	T
CONTIG21278_1149-3562-622:629	Bradi3g04350	GO:0005515 protein binding GO:0001653 peptide receptor GO:0004888 transmembrane receptor activity GO:0004672 protein kinase activity	0.834	Up-regulation	Hyper-methylated genome A	T

CONTIG2710387_1-564-171:303	Bradi1g15610 Potassium ion transmembrane transporter activity	GO:0009674 potassium:sodium symporter activity	0.677	Up-regulation	Hyper-methylated genome A	T
CONTIG2862936_1-548-255:308	Bradi3g59920 ATPase activity	GO:0008017 microtubule binding	0.766	Up-regulation	Hypo-methylated genome A	T
CONTIG336185_1-1612-297:381	Bradi2g61980	GO:0009672 auxin:hydrogen symporter activity	0.694	Up-regulation	Hypo-methylated genome B	T
CONTIG75035_1-2272-1646:1765	Bradi1g63433	GO:0005524 ATP binding GO:0042626 ATPase activity, transmembrane movement	0.035	Down-regulation	Hypo-methylated genome A	T

Table 5.6. Annotation of differentially methylated sites. *Brachypodium* genes and GO annotations associated with each of the contigs containing a differential methylation site (>= 15%) between the 12°C and 27°C samples plus a PPLR value for the differentially methylated genome that deviated from 0.5 by +/- 20%. *T: transcribed, NT: non-transcribed.

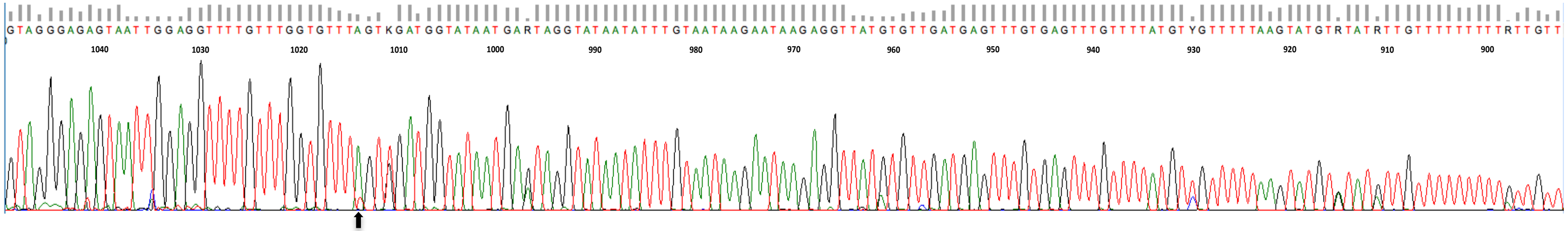
5.9 Validation of homeologous SNP calls and methylation status

The DNA for the 12°C and 27°C samples was bisulfite treated using the EZ DNA Methylation-Gold kit (Zymo research group). 23 SNP sites were selected at random for validation and primers were designed to capture 150-400bp regions surrounding the SNP sites in bisulfite treated DNA. These regions contained a total of 337 analyzed cytosine sites (304 un-methylated and 33 partially or fully methylated) that could also be used for methylation site status validation. For primer design all C's were treated as Y's (C/T) and no more than 2 Y's were included in a primer sequence. PCR amplification of the DNA followed using KAPA HiFi HotStart Uracil+ ReadyMix and finally samples were sequenced using Sanger sequencing (by Source Bioscience). If SNP alleles that had been seen in the next generation sequencing data could be seen in the Sanger sequencing data in approximately the same proportions (within ~20%) they were said to be validated. Similarly a methylation call was deemed to be correct if the Sanger sequencing data showed a proportion of methylation that was within approximately 20% of that seen previously i.e. 50% methylation would mean approximately equal peaks of C and T in the Sanger sequencing data.

In this independent validation of methylation and SNP calls ~80% of SNPs analyzed from both samples were validated, >99% of sites that had been previously determined to be un-methylated were confirmed in both samples and finally of those sites that had been previously determined to be fully or partially methylated 88% were found to be accurate calls in the 12°C sample (91% in 27°C sample). An example of one such SNP validation in the 12°C sample is shown in figure 5.9 (all additional SNP validations in Appendix 4, figure 1). This SNP analysis was coupled to the analysis of the cytosine residues surrounding the SNP call. Table 5.7 details the positions of these residues and their expected and observed methylation statuses, demonstrating the high degree of accuracy of calls generated within this study (all additional cytosine residue validations in Appendix 4, table 8).

Methylation site (893-1115bp)	Status in next generation sequencing data Sample 12	Peak in Sanger sequencing data Sample 12	Methylation site	Status in next generation sequencing data Sample 12	Status in Sanger sequencing data Sample 12
894	98.4% un-methylated	T (un-methylated)	964	98.9% un-methylated	T (un-methylated)
899	100% un-methylated	T (un-methylated)	965	98.9% un-methylated	T (un-methylated)
900	99.7% un-methylated	T (un-methylated)	972	99.2% un-methylated	T (un-methylated)
901	99.3% un-methylated	T (un-methylated)	981	99.5% un-methylated	T (un-methylated)
906	99.7% un-methylated	T (un-methylated)	983	98.2% un-methylated	T (un-methylated)
907	99.3% un-methylated	T (un-methylated)	984	100% un-methylated	T (un-methylated)
912	97.3% un-methylated	T (un-methylated)	985	98.4% un-methylated	T (un-methylated)
924	98.7% un-methylated	T (un-methylated)	987	96.1% un-methylated	T (un-methylated)
926	99.4% un-methylated	T (un-methylated)	990	99.5% un-methylated	T (un-methylated)
928	98.7% un-methylated	T (un-methylated)	992	99.0% un-methylated	T (un-methylated)
930	56.6% un-methylated	~60% T / ~40% C	996	99.2% un-methylated	T (un-methylated)
931	99.4% un-methylated	T (un-methylated)	1003	98.3% un-methylated	T (un-methylated)
937	98.7% un-methylated	T (un-methylated)	1015	98.0% un-methylated	T (un-methylated)
940	95.7% un-methylated	T (un-methylated)	1017	98.4% un-methylated	T (un-methylated)
941	98.1% un-methylated	T (un-methylated)	1019	98.8% un-methylated	T (un-methylated)
942	99.1% un-methylated	T (un-methylated)	1024	98.4% un-methylated	T (un-methylated)
946	97.8% un-methylated	T (un-methylated)	1026	99.2% un-methylated	T (un-methylated)
949	99.8% un-methylated	T (un-methylated)	1036	98.1% un-methylated	T (un-methylated)
958	99.3% un-methylated	T (un-methylated)	1039	98.5% un-methylated	T (un-methylated)
960	97.8% un-methylated	T (un-methylated)	1048	98.2% un-methylated	T (un-methylated)
962	99.3% un-methylated	T (un-methylated)			

Table 5.7. Methylation site validation data. Observations from the Sanger sequencing output generated for sample 12 (12°C) in Contig462845_1-1429-894 between positions 894 and 1049 (figure 5.9). Data used for the validation of methylation sites at individual cytosine residues.



SNP position 1014
In next generation sequencing see; T: 33% A: 67%
We see here; T: ~30% A: ~70%

Figure 5.9. Sanger sequencing output trace. Raw sequencing output generated for sample 12 (12°C) in Contig462845_1-1429-894 between positions 894 and 1049. Data used for the validation of SNP and methylation sites.

5.10 Conclusions

The methylation array enriched the 3 genomes consistently without bias for methylated and un-methylated regions (section 5.4). When mapped, enriched data extended out from target regions covering over 4 times the original 6Mb capture target sequence. Validation of up to 88% of homeologous SNPs and over 99% of methylation sites that were analyzed in section 5.9 supports enrichment combined with bisulfite treatment as a sound method to accurately identify SNPs and methylation sites in enriched hexaploid wheat data. There was no bias introduced for specific repetitive regions by enrichment and it was noted that transposons in general were hyper-methylated in comparison to gene regions.

~5% of the cytosine residues that were analyzed overall contained methylation in 1 or more genomes. Genome specific methylation did exist between the A, B and D genomes with the D genome showing preferential methylation at a significant level. The enriched genes that were linked to the D genomes preferential methylation, unlike those linked to the A and B genome, were involved in; chromatin silencing, histone modification, methylation/regulation of gene expression and chromosome organization; adding weight to the hypothesis that the expression of the D genome may be controlled by methylation. Genome independent methylation existed more commonly across the genome than genome specific methylation (~73-74% of all analyzed residues that contained methylation in one or more genomes). Bias for CpG sites was stronger in genome independent sites, while genome specific sites had a consistently higher proportion of CHG/CHH methylation in predicted transcribed and non-transcribed regions, but still a bias towards CpG methylation (section 5.6).

Genome independent methylation or non-methylation resulted in consistent expression profiles across the 3 genomes i.e. similar levels of expression in associated genes. In contrast to this, genome specific methylation was seen to effect the expression of that genome, in relation to the expression profiles of the other 2 genomes, in non-transcribed regions; such genome specific methylation in genome A was linked to low expression of genes linked to phosphorylation, carboxylase, synthase, O-acyltransferase, ATPase, protein kinase and transglucosylase activity and high expression of genes with arabionsyltransferase and auxin transporter activity; in genome B genome specific methylation was linked to low expression of genes involved in transmembrane transporter activity; in genome D genome specific methylation was linked to low expression in genes with ligase activity, catalytic activity and damaged DNA binding activity. In transcribed regions genome specific methylation was seen to effect the expression of the D genome only, in relation to the expression profiles of the other 2 genomes. A large decrease in expression was seen when genome D was specifically methylated in transcribed regions in a gene associated with monooxygenase activity. This could be indicative of first exon methylation that is strongly associated with gene silencing (Brenet *et al.,* 2011). Also a small increase in expression was seen when genome D was specifically methylated in transcribed regions in a gene associated with ATP-dependent DNA helicase activity. This supports the hypothesis that methylation could be directly affecting gene expression in wheat in general across the genome (section 5.8).

Temperature was found to have an effect on methylation; 23 differentially methylated sites were identified between the 12°C and 27°C samples. These sites were mainly genome specific but differences were also seen in 2/3 genomes indicating that it is likely to be possible for methylation differences to occur genome independently between samples, however, much more rarely. The adaptation of this study to a larger region of the genome would be likely to confirm this. Some of the temperature dependent differential methylation was found to be likely to be linked to the increased or decreased expression of affected genes with high biological significance (table 5.6); 1 up-regulated gene in the 27°C sample was linked to stress response due to the presence of stress response elements (2 sites correlated with this gene), 1 up-regulated gene was found to be similar to the PRP5 heat shock protein in a BLAST alignment and finally 1 up-regulated gene was related to putative glycine hydroxymethyltransferase. 1 down-regulated gene in the 27°C sample had similarity to the vacuolar ATPase B subunit that has interestingly been linked to tissue specific transcript level decrease under heat stress (Kluge *et al.,* 2003), 1 down-regulated gene was likely to encode a frigida-like protein found typically in winter accessions to prevent flowering until after winter (Risk *et al.,* 2010) and finally a down-regulated gene was related to the F-box protein that has been up-regulated under cold stress (Jain *et al*., 2007) (section 5.8).

6. Discussion

The main aims of this project included the development of high throughput pipelines for mapping, SNP calling and mapping-by-sequencing mutant identification analyses in complex species. Mapping and SNP calling analyses were first implemented in the model diploid organism *Arabidopsis.* Analysis of the *Arabidopsis* clock mutant *early bird* (*ebi-1*), that was developed using BWA and GATK, was successfully utilized to identify a large proportion of *ebi-1* specific SNPs that were identified by Ashelford *et al*. in 2010 (~90%) and the SNP that is responsible for causing the *ebi-1* phenotype. Mutant identification analyses were also performed in various known bulk segregant *Arabidopsis* mutants using SHOREmap (Schneeberger *et al.,* 2009). All analyses gained clear intervals of interest in anticipated mapping intervals, and a short list of potential phenotype inducing EMS SNPs, demonstrating the functionality of this methodology. These analyses outlined the theory demonstrated by James *et al*. in 2013 that a higher number of pooled plants used in the analysis would increase our ability to define an interval of interest.

Initial mapping, SNP calling and mutant identification in a diploid organism with a complete reference sequence allowed pipeline development prior to the end goal of application in the target enriched hexaploid bread wheat. In addition to mutant identification in SHOREmap these results were also replicated using a bespoke mapping, SNP calling and mutant identification pipeline, to assist in its development. Bespoke pipeline development was considered necessary in anticipation of the downstream polyploid application issues that were encountered with use of methodologies such as SHOREmap that is intended for diploid organisms only. The bespoke mutant identification pipeline was successfully implemented on a simulated hexaploid dataset that was created using the *Arabidopsis* genome.

Wheat's large genome size makes the whole genome re-sequencing, that mapping-by-sequencing analyses benefit from, a costly option. The NimbleGen gene capture array was implemented to eliminate repetitive sequence and to target enrich the majority of gene rich regions of wheat (~110Mb) covering ~99% of the wheat cDNA sequence. Here, this gene capture array's efficiency was validated and it was used to enrich 4 wheat varieties. Use of a polyploid mapping and SNP calling pipeline allowed determination of varietal SNPs in comparison to the reference dataset, Chinese Spring, in each of the other 3 varieties. This enabled varietal comparative analyses. Varietal SNP numbers were loosely conserved across Truman, Utmost and Rialto and more strongly conserved across Utmost/Truman with Rialto having the most overall. No two varieties appeared to have more or less similarity to each other in comparison to the 3rd variety and Chinese Spring. Looking at SNP distribution smaller regions of varietal SNP density or scarcity were numerous and scattered consistently across the genome. Rialto appeared to have a SNP dense region, compared to the other chromosomes and other varieties, at the beginning of chromosome 1 that could account for the elevated numbers of varietal Rialto specific SNPs in comparison to Utmost and Truman. Since Rialto is reported to contain the 1B/1R wheat/rye translocation on chromosome 1 this could account for increased SNP diversity in this region (Burnett *et al.,* 1995).

In addition to a manageable size for cost effective re-sequencing, mapping-by-sequencing largely benefits from a complete reference genome, and as such many of the current mapping-by-sequencing pipelines require such a sequence (Nordstrom *et al.,* 2013). Here, the full genic sequence of wheat i.e. the target sequence for the gene capture array, was assembled into pseudo-chromosomes where possible based on synteny with the closely related *Brachypodium*. 68% of the target sequence contigs could be ordered and concatenated into 7 pseudo wheat chromosomes using this methodology. 800 *Brachypodium*-wheat markers were used for this assembly that was then implemented successfully in several mutant identification analyses. This successful application demonstrates that in relatively un-characterized species mutant identification analyses are possible with limited resources. With the release of the wheat genome zipper to the public in March 2014 by the IWGSC (available at: https://urgi.versailles.inra.fr/download/iwgsc/) the number of available wheat markers has been increased and future pseudo-chromosome constructs are likely to benefit from a more accurate structure. Furthermore, when full wheat chromosome assemblies are completed, this will allow a further improved set of pseudo wheat chromosomes with all of the genic sequence included and arranged confidently, without the limitation of including only those contigs that could align to barley or *Brachypodium.* A full wheat genome chromosomal assembly would allow the development of the enrichment array itself to ensure inclusion of all of the relevant genic material for wheat.

Using gene capture target enrichment and the pseudo-chromosome reference sequence, in combination with an evolved version of the bespoke mutant identification pipeline that was previously implemented in *Arabidopsis,* mapping intervals were successfully identified in a series of diploid and hexaploid wheat mutant bulk segregant F2 mapping populations.

A region was identified on chromosome 3 that is likely to contain the *Eps-3A^{m}* mutation in the early flowering diploid mutant *T. monococcum*. A region of ~40Kbp region could be pinpointed based on the identification of deletion hotspots. Finally, by assessing gene annotation, the candidate gene for the phenotype itself could be narrowed to a single capture target sequence contig of 3693bp that had a high deletion frequency and showed a high degree of similarity to the *T. aestivum* cultivar Chinese Spring LUX gene. The LUX gene is known to affect both the circadian clock and flowering time in *Arabidopsis* and had previously been associated with this deletion (Hazen *et al.,* 2005). This analysis highlights the capability of capture probe sets to be used effectively for close relatives with little or no resources available. This developed a proof of concept approach where enrichment of a subset of a phenotyped *Arabidopsis* F2 mapping population was performed in combination with a mapping-by-synteny approach to order *Arabidopsis* cDNA into *B. rapa* pseudo-chromosomes based on synteny. Two mutant intervals were defined in *B. rapa* using allele frequency analysis at marker positions that translated to one position in *Arabidopsis* (Galvão et al 2012). With use of an enrichment system targeting the majority of the wheat genic sequence the concerns expressed by Galvão et al, that the causal mutation would be unlikely to be targeted with enrichment, were addressed and the likelihood of enrichment of the region of interest were increased. Here, not only had a divergent species been used to order the fragmented mapping reference, additionally the mapping reference itself and enrichment capture probe set were both divergent from the species under analysis. This analysis also uniquely demonstrates the combination of sliding window analyses with mapping-by-synteny by implementing a pseudo genome.

Highly similar ~1Mb intervals were defined in the P2 and P3 Yr-7 stripe rust resistant and susceptible wheat datasets on chromosome 2. A small group of novel SNPs, within this interval, that were found in common disease resistance genes, were defined as candidates for further investigation. In the P2 dataset 7 SNPs were associated with the NBS-LRR disease resistance protein family; the 7 SNPs were all unique to P2 and approximately central to the tip of the peak interval of interest that was defined on chromosome 2. This SNP region is likely to indicate the location of the Yr-7 locus for stripe rust resistance. In the P3 dataset 21 SNPs were associated with the NBS-LRR disease resistance protein family. This SNP increase in the P3 dataset compared to P2, in the same peak region on chromosome 2 that showed homology to the NBS-LRR disease resistance proteins, could be responsible for the disruption of the disease resistance gene and therefore disease susceptibility in this sample. Therefore these 21 novel SNPs are candidates for further study. 90% of the 159 iSelect Yr-7 linked SNPs could be correctly anchored to the pseudo-chromosome 2. ~80% fell within the defined intervals or in close proximity to them (6,000,00-9,000,000bp) supporting correct interval identification although none were in the list of 21 candidate SNPs. In theory Yr-7 linked SNPs should all relatively closely associate with peak regions. Cases where this was not true are likely to result from local inaccuracies in pseudo-chromosome order. As such, this analysis would benefit from increased numbers of markers in *Brachypodium* for improved ordering of the pseudo-chromosomes, and/or an increased number of SNPs that are linked to the gene of interest. Only 65% of Yr-7 linked SNP sequences could be confidently located in the gene capture array design-space. It is anticipated that those sequences that could not be confidently placed overlapped the ends of assembled contigs, potentially hitting more than one design-space contig or extended into repetitive regions that were not included in the design-space. A full wheat genome assembly would allow scaffolding of the gene capture array design-space contigs and relative positional information regarding the SNP positions to be gained to confirm this.

The initial main aim of mapping, SNP calling and mutant identification in an enriched hexaploid wheat dataset using a pseudo-chromosome reference sequence has been achieved with use of a bespoke pipeline and allele frequency algorithm. This pipeline is detailed within this project with attached Perl scripts but is also available within the Discovery Environment of iPlant (The iPlant Collaborative, 2011) (see section 4.2.5) as series of public workflows allowing users with a non-programming background to utilize the methodology and requiring, through an online graphical user interface, only uploads of sequencing data (fastq) files as input. The default algorithm is set for a diploid organism, though with a simple change of parameters a polyploid organism can be processed (as demonstrated in the study here of P2 and P3 hexaploid wheat samples). This allows a specialist mapping-by-sequencing pipeline in a polyploid or diploid plant to be truly accessible to non-specialists for implementation on additional datasets.

The mapping, SNP calling and enrichment techniques detailed here could then be applied to enable a study of methylation patterns in a subset of wheat. The wheat methylation array (6Mbp), or Agilent SureSelect Methyl-Seq Target Enrichment System, was designed based on the pre-validated probe target sequences used for the NimbleGen gene capture array. It was used to enrich samples that were from the wheat variety Chinese Spring and included; plants grown at 12°C and plants grown at 27°C. Bisulfite treatment was also used to allow, after PCR and sequencing, discrimination of methylated cytosine residues. It was found that the methylation array enriched the 3 genomes consistently without bias for methylated and un-methylated regions.

The methylation analysis confirmed our main hypotheses. Genome specific methylation did exist between the A, B and D genomes with the D genome showing preferential methylation at a significant level. The enriched genes that were linked to the D genomes preferential methylation, unlike those linked to the A and B genome, were involved in; chromatin silencing, histone modification, methylation/regulation of gene expression and chromosome organization; adding weight to the hypothesis that the expression of the D genome may be controlled by methylation. Genome independent methylation existed more commonly across the genome than genome specific methylation. Bias for CpG sites was stronger in genome independent sites, while genome specific sites had a consistently higher proportion of CHG/CHH methylation in predicted transcribed and non-transcribed regions, but still a CpG methylation bias (section 5.6).

Genome independent methylation or non-methylation resulted in consistent expression profiles across the 3 genomes i.e. similar levels of expression. Genome specific methylation was seen to effect the expression of the genome in question, across a variety of genes, in relation to the expression profiles of the other 2 genomes, in non-transcribed regions. In transcribed regions genome specific methylation was seen to effect the expression of the D genome only, in relation to the expression profiles of the other 2 genomes. A large decrease in expression was seen when genome D was specifically methylated in transcribed regions in a gene associated with monooxygenase activity. This could be indicative of first exon methylation that is strongly associated wth gene silencing (Brenet *et al.,* 2011). Also a small increase in expression was seen when genome D was specifically methylated in transcribed regions in a gene associated with ATP-dependent DNA helicase activity. This supports the hypothesis that methylation could be directly affecting gene expression in wheat in general across the genome (section 5.8). The observation of preferential D genome methylation in genes that are involved in chromatin silencing, histone modification, methylation/regulation of gene expression and chromosome organization combined with changes in D genome gene expression could be indicative of the use of methylation to control the D genome to maintain stability in the plant after the addition of this, the newest genome to be introduced, to create hexaploid wheat.

Temperature was found to have an effect on methylation; differences between the 12°C and 27°C samples were mainly genome specific but differences were also seen in 2/3 genomes indicating that it is likely to be possible for methylation differences to occur genome independently between samples, however, much more rarely. The adaptation of this study to a larger region of the genome would be likely to confirm this. Some of the temperature dependent differential methylation was found to be likely to be linked to the increased or decreased expression of affected genes with high biological significance (table 5.6); 1 up-regulated gene in the 27°C sample was linked to stress response due to the presence of stress response elements (2 sites correlated with this gene), 1 up-regulated gene was found to be similar to the PRP5 heat shock protein in a BLAST alignment and finally 1 up-regulated gene was related to putative glycine hydroxymethyltransferase. 1 down-regulated gene in the 27°C sample had similarity to the vacuolar ATPase B subunit that has interestingly been linked to tissue specific transcript level decrease under heat stress (Kluge *et al.,* 2003), 1 down-regulated gene was likely to encode a frigida-like protein found typically in winter accessions to prevent flowering until after winter (Risk *et al.,* 2010) and finally a down-regulated gene was related to the F-box protein that has been up-regulated under cold stress (Jain *et al*., 2007) (section 5.8).

This methylation analysis was the first genome wide methylation study in wheat, selecting a subset of wheat genes that were distributed relatively evenly across the wheat genic sequence. Such an analysis should give a clear indication of the trends and methylation patterns that we are likely to see if the study is extended to include the entirety of wheat's genic sequence. This is an obvious next step for this analysis, to increase the region under observation e.g. with adaptation of the ~110Mbp gene capture array target sequence, encompassing the majority of what genes, into a larger scale methylation analysis. This analysis also only allowed the analysis of one strand of DNA and future analyses would benefit from the analysis of both strands. Implementation of the PacBio RS could aid a larger scale approach as the unique sequencing methodology employed by this technology allows discrimination of the methylation status of a cytosine residue without the need for bisulfite treatment. This is due to DNA polymerase synthesizing DNA at slightly different speeds depending on whether it is epigenetically modified or not. Nucleotides emit pulses of fluorescent light as they are added to the DNA and by calculating the lengths of the pulses and distances between them; it is possible to identify methylated sites (Flusberg *et al.,* 2010).

Implementation of target enrichment has cost effectively facilitated mapping and SNP calling, mapping-by-sequencing mutant identification, varietal comparisons and methylation analyses in the genic regions of wheat. All of this has been achieved with no finished genome reference sequence and outlines the possibility of such analyses being implemented on other non-model organisms. Discrimination between the 3 genomes has been achieved using homeologous SNPs to allow comparison between them in several of these studies, this is also a limiting factor since the absence of a homeologous SNP hinders genome discrimination. Future work in this field would benefit from longer length sequencing reads allowing more methylation sites to be connected to a homeologous SNP to enable genome discrimination. Correlation of a methylated region with desirable gene expression in wheat could allow methylation status to be used as a marker for improved crop yields. A simple cost effective assay to routinely detect such methylation sites, e.g. the PCR/Sanger sequencing assay implemented within this study, could be used in crop yield development.

The 454 genomic wheat DNA sequence assembly that was generated by Brenchley *et al*. was used throughout this project and consisted of a contig set that is anticipated to represent the complete wheat gene set. As this project ends the IWGSC released the full genome sequence of wheat in July 2014; an improved version of the 454 DNA assembly, containing a set of contigs that have been assigned to an individual wheat genome A, B or D and a wheat chromosomal arm (IWGSC, 2014). Such sequence could be used in future mapping studies similar to those implemented here with enriched data allowing assignment of sequences to individual wheat genomes via mapping rather than using homeologous SNPs. Such analyses would benefit from long, high quality sequencing reads to ensure effective association of a sequencing read with the correct wheat genome given the high degree of similarity between them. Full wheat chromosomal sequences are still under construction by the IWGSC therefore the wheat contigs have still yet to be assigned, in the main, anchored positional information that would allow the construction of an improved set of pseudo-chromosomes with all of the target sequence contigs included and arranged confidently to improve our mutant identification pipeline. The utilization of the IWGSC assembly has been retrospectively applied throughout this project since its release to ascertain if the pseudo-chromosome sequences contain sequence from the correct wheat chromosome (>80% of sequence was correctly assigned to a chromosome and all analyses yielded identical results with un-validated sequence removed). This would not aid validation of the local ordering of the sequence within the pseudo-chromosomes. This recent wheat genome sequence was complemented by 124,201 gene annotations; these would improve the methylation study in chapter 5 allowing differentiation of promoter, first exon, internal exon/intron structures that assist the effort to correlate the positional effect of a methylation site on gene expression; annotations would also assist in the effort to define phenotype inducing SNPs in the P2/P3 datasets in chapter 4, allowing discrimination of synonymous/non-synonymous SNPs.

At the end of this project it is clear that many of the sequencing technologies, mapping/SNP calling tools and mutant identification pipelines that were commonly used at the onset of the project have since been replaced by newer more improved versions or else disregarded altogether. Programs such as SHOREmap, BWA and Bowtie, to name a few, received total overhauls and the newer algorithms and methodologies that have been detailed here are typically much more sophisticated and multi-functional than they were in 2010. Sequencing technologies such as the Roche 454 sequencer are soon to be largely redundant with competitors such as the PacBio RS II entering the market while the SOLiD sequencing methodology has seen an invasion of the market by the increasingly high throughput of the Illumina HiSeq series. In such a fast moving field, with increased accessibility of computing power and huge amounts of sequence data, of increasing read length, that is generated in a single day, bioinformatic analyses are likely to be continually under development.

7. References

Abe, A., Kosugi, S., Yoshida, K., Natsume, S., Takagi, H., Kanzaki, H., Matsumura, H., Yoshida, K., Mitsuoka, C., Tamiru, M., Innan, H., Cano, L., Kamoun, S. & Terauchi, R. (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. *Nature Biotechnology,* **30**, 174-178

Adams, K. L., Cronn, R., Percifield, R. & Wendel, J. F. (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing*, Proc. Natl. Acad. Sci. USA*, **100,** 4649–4654

Agilent Technologies Inc., 2012. *SureSelect Methyl-Seq target Enrichment System for Illumina Multiplexed Sequencing* [online] Available at: http://www.chem.agilent.com/Library/usermanuals/Public/G7530-90002.pdf [Accessed 14 June 2012]

Allen, A. M., Barker, G. L., Berry, S. T., Coghill, J. A., Gwilliam, R., Kirby, S., Robinson, P., Brenchley, R. C., D'Amore, R., McKenzie, N., Waite, D., Hall, A., Bevan, M., Hall, N. & Edwards, K. J. (2011) Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (Triticum aestivum L.). *Plant Biotechnology Journal,* **9**, 1086–1099

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

Applied Biosystems, Life Technologies, 2011. *Overview of SOLiD Sequencing Chemistry* [online] Available at: http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems/solid-sequencing-chemistry.html [Accessed 09 February 2011]

Applied Biosystems, Life Technologies, 2011. *SOLiD Accuracy Enhancement Tool* [online] Available at: http://marketing.appliedbiosystems.com/mk/get/SOLID_ACCURACY_TOOLS;jsessionid=cc30a7c80866$14$16$A?_A=116807&_D=73427&_V=0 [Accessed 06 January 2012]

Applied Biosystems, Life Technologies, 2012. *Applied Biosystems SOLiD 3 System-Library Preparation Guide* [online] Available at: https://banana-slug.soe.ucsc.edu/_media/lab_protocols:solid3_libprep_guide.pdf [Accessed 11 March 2014]

Ashelford, K., Eriksson, M. E., Allen, C. M., D'Amore, L., Johansson, M., Gould, P., Kay, S., Millar, A. J., Hall, N. & Hall, A. (2010) Full genome re-sequencing reveals a novel circadian clock mutation in *Arabidopsis*, *Genome Biology,* **12,** 28

Austin, R., Vidaurre, D., Stamatiou, G., Briet, R., Provart, N., Bonetta, D., Zhang, J., Fung, P., Gong, Y., Wang, P., McCourt, P. & Guttman, D. (2011) Next-generation mapping of Arabidopsis genes. *The plant journal,* **67**, 715-725

Baralle, D. & Baralle, M. (2005) Splicing in action: assessing disease causing sequence changes, *Journal of Medical Genetics,* **42**, 737-748

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

Bennetzen, J. L. & Ma, J. (2003) The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis, *Curr. Opin. Plant Biol.*, **6**, 128-133

Bent, E., Johnson, S. & Bancroft, I. (1998) BAC representation of two low-copy regions of the genome of *Arabidopsis thaliana, Plant J.,* **13** (6), 849-55

Berglund, E. C., Kiialainen, A. & Syvänen, A. (2011) Next-generation sequencing technologies and applications for human genetic history and forensics, *Investigative Genetics* **2**, 23

Bevan, M. & Walsh, S. (2005) The *Arabidopsis* genome: A foundation for plant research, *Genome Research,* **15,** 1632-1642

Bottley, A., Xia, G. M. & Koebner, R. M. D. (2006) Homoeologous gene silencing in hexaploid wheat, *Plant Journal*, **47**, 897–906

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G., D'Amore, R., Allen, A., McKenzie, N., Kramer, M., Kerhornou, A., Bolser, D., Kay, S., Waite, D., Trick, M., Bancroft, I., Gu, Y., Huo, N., Luo, M., Sehgal, S., Gill, B., Kianian, S., Anderson, O., Kersey, P., Dvorak, J., McCombie, W., Hall, A., Mayer, K., Edwards, K., Bevan, M. & Hall, N. (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature,* **491,** 705–710

Brenet, F., Moh, M., Funk, P., Feierstein, E., Viale, A., Socci, N. & Scandura, J. (2011) DNA Methylation of the first exon is tightly linked to transcriptional silencing, *PLOS one,* **6** (1), 14524

Burnett, C. J., Lorenz, K. J. & Carver, B.F. (1995) Effects of the 1B/1R translocation in wheat on composition and properties of grain and flour, *Euphytica,* **86,** 159-166

Campoli, C., Pankin, A., Drosse, B., Casao, C. M., Davis, S. J. & Korff, M. (2013) HvLUX1 is a candidate gene underlying the early maturity 10 locus in barley: phylogeny, diversity, and interactions with the circadian clock and photoperiodic pathways. *New Phytol*. **199**, 1045-1059

Canadian Food Inspection Agency. (2013) *Plant Breeders' Rights, Crop Reports, CDC Utmost* [online] Available at: http://www.inspection.gc.ca/english/plaveg/pbrpov/cropreport/whe/app00007656e.shtml [Accessed 14 October 2013]

Cantu, D., Vanzetti, L., Sumner, A., Dubcovsky, M. & Matvienko, M. (2010) Small RNAs, DNA methylation and transposable elements in wheat, *BMC Genomics,* **11**, 408

Cao1, X., Zhou1, J., Gong, X., Zhao, G., Jia, J. & Qi1, X. (2012) Identification and validation of a major QTL for slow-rusting resistance to stripe rust in wheat, *Journal of Intergrative Plant Biology,* **5,** 330-44

Cavanagh, C. R., Chao, S., Wang, S., Huang, B. E., Stephen, S., Kiani, S., Forrest, K., Saintenac, C., Brown-Guedira, G., Akhunova, A., See, D., Bai, G., Pumphrey, M., Tomar, L., Wong, D., Kong, S., Reynolds, M., Lopez de Silva, M., Bockelman, H., Talbert, L., Anderson, J. A., Dreisigacker, S., Baenziger, S., Carter, A., Korzun, V., Morrell, P. L., Dubcovsky, J., Morell, M., Sorrells, M. E., Hayden, M. J. & Akhunov, E. (2013) Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars, *PNAS*, doi/10.1073/pnas.1217133110

Chaudhuri, S. & Messing, J. (1994) Allele-specific parental imprinting of dzr1, a posttranscriptional regulator of zein accumulation, *Proc Natl Acad Sci USA*, **91**, 4867–4871

Chen, X. M., Moore, M., Milus, E. A., Lon, D. L., Line, R. F., Marshall, D. & Jackson, L. (2002) Wheat stripe rust epidemics and races of Puccinia striiformis f. sp. tritici in the United States in 2000, *Plant Dis.*, ***86***, 39–46

Choulet, F., Wicker, T., Rustenholz, C., Paux, E., Salse, J., Leroy, P., Schlub, S., Le Paslier, M. C., Magdelenat, G., Gonthier, C., Couloux, A., Budak, H., Breen, J., Pumphrey, M., Liu, S., Kong, X., Jia, J., Gut, M., Brunel, D., Anderson, J. A., Gill, B. S., Appels, R., Keller, B. & Feuillet, C. (2010) Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces, *Plant Cell,* **22**, 1686–1701

Clarke, J., Wu, H., Jayasinghe, L., Patel, A., Reid, S. & Bayley, H. (2009) Continuous base identification for single molecule nanopore DNA sequencing, *Nature Nanotechnology,* **4,** 265-270

Clark, M. J., Chen, R., Lam, H., Karczewski, K., Chen, R., Euskirchen, G., Butte, A. & Snyder, M. (2011) Performance comparison of exome DNA sequencing technologies, *Nature Biotechnology,* **29,** 908-914

Columbia University Press, 2005. *The Columbia Electronic Encyclopedia, 6th edition, Wheat* [online] Available at: http://www.infoplease.com/encyclopedia/science/wheat-wheat-varieties-their-uses.html [Accessed 18 October 2013]

Covaris, The sample prep advantage, 2011. *Adaptive Focused Acoustics-DNA shearing with AFA* [online]. Available at: http://www.covarisinc.com/dna-shearing.html [Accessed 13 March 2012]

Copenhaver, G., Nickel, K., Kuromori, T., Benito, M., Kaul, S., Lin, X., Bevan, M., Murphy, G., Harris, B., Parnell, L., McCombie, W. R. Martienssen, R., Marra, M. & Preuss, D. (1999) Genetic Definition and Sequence Analysis of *Arabidopsis* Centromeres, *Science,* **286** (5449), 2468-2474

Cullum, R., Alder, O. & Hoodless, P. A. (2010) The next generation: using new sequencing technologies to analyze gene regulation, *Respirology,* **2,** 210-22

Darst, R. P., Pardo, C. E., Ai, L., Brown, K. D. & Kladde, M. P. (2010) Bisulfite sequencing of DNA, Curr. Protoc. Mol. Biol., **7,** 1-17

Davey, J. W. & Blaxter M. L. (2010) RADseq: next-generation population genetics, *Briefings in Functional Genomics,* **9** (5-6), 416-423

Deng, Z., Zhang, X., Wang, X., Jing, J. & Wang, D. (2004) Identification and Molecular Mapping of a Stripe Rust Resistance Gene from a Common Wheat Line Qz180, *Acta Botanica Sinica,* **46,** 236-241

DePristo, M., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl, C., Philippakis, A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T., Kernytsky, A., Sivachenko, A., Cibulskis, K., Gabriel, S., Altshuler, D. & Daly, M. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data, *Nature Genetics*, **43**(5), 491-498

Doitsidou, M., Poole, R. J., Sarin, S., Bigelow, H. & Hobert, O. (2010) C. elegans mutant identification with a one-step whole-genome-sequencing and SNP mapping strategy. *PLoS One,* 5, e15435. doi:10.1371/journal.pone.0015435

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

Dvorak, J., Akhunov, E. D., Akhunov, A. R., Deal, K. R. & Luo, M. C. (2006) Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat, *Mol. Biol. Evol.,* **23,** 1386–1396

Earley, E. J. & Jones, C. D. (2011) Next-generation mapping of complex traits with phenotype-based selection and introgression, *Genetics,* **4,** 1203-9

Edgar, R. C., Asimenos, G., Batzoglou, S. & Sidow, A., 2012. *Evolver: a whole-genome sequence evolution simulator* [online] Available at: http://www.drive5.com/evolver/ [Accessed 12 March 2014]

Finnegan, E. J., Genger, R. K., Peacock, W. J. & Dennis, E. S. (1998) DNA methylation in plants, *Annual Review of Plant Physiology and Plant Molecular Biolog*y, **49**, 223-247

Flusberg, B., Webster, D, Lee, J., Travers, K., Olivares, E., Clark, T., Korlach J. & Turner, S. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing, *Nature Methods,* **7**, 461-465

Foresight, The Government Office for Science, 2011. *The Future of Food and Farming, Final Project Report* [online] Available: http://www.bis.gov.uk/assets/foresight/docs/food-and-farming/11-546-future-of-food-and-farming-report.pdf [Accessed 12 March 2014]

Frisch, M. & Melchinger, A. E. (2005) Selection theory for marker-assisted backcrossing, *Genetics,* **170** (2), 909-17

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

Gawroński, P., Ariyadasa, R., Himmelbach, A., Poursarebani, N., Kilian, B., Stein, N., Steuernagel, B., Hensel, G., Kumlehn, J., Sehgal, S. K., Gill, B. S., Gould, P., Hall, A. & Schnurbusch, T. (2014) A distorted circadian clock causes early flowering and temperature-dependent variation in spike development in the Eps-3A mutant of einkorn wheat, *Genetics,* **196** (4), 1253-61

Glaus, P., Honkela, A. & Rattray, M. (2012) Identifying differentially expressed transcripts from RNA-seq data with biological variation**,** *Bioinformatics,* **28,** 1721-1728

Glazier, A. M. Nadeau, J. H. & Aitman, T. J. (2002) Finding Genes That Underlie Complex Traits, *Science,* **298 (**5602), 2345-2349

Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B. M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W. L., Chen, L., Cooper, B., Park, S., Wood, T. C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R. M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A. & Briggs, S. (2002) A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. japonica), *Science,* **296**(5565), 92-100

Gu, Y. Q., Coleman-Derr, D., Kong, X. & Anderson, O. D. (2004) Rapid Genome Evolution Revealed by Comparative Sequence Analysis of Orthologous Regions from Four Triticeae Genomes, *Plant Physiology,* **135,** 459-470

Gu, Y. Q., Salse, J., Coleman-Derr, D., Dupin, A., Crossman, C., Lazo, G. R.. Huo, N., Belcram, H., Ravel, C., Charmet, G., Charles, M., Anderson, O. D. & Chalhoub, B. (2006) Types and Rates of Sequence Evolution at the High-Molecular-Weight Glutenin Locus in Hexaploid Wheat and Its Ancestral Genomes, *Genetics*, **174,** 1493-1504

Harmer, S. L., Hogenesch, J. B., Straume, M., Chang, H.S., Han, B., Zhu, T., Wang, X., Kreps, J.A. & Kay, S. A. (2000) Orchestrated transcription of key pathways in Arabidopsis by the circadian clock, *Science*, **290**, 2110-2113

Harris, R. A., Wang, T., Coarfa, C., Nagarajan, R., Hong, C., Downey, S., Johnson, B., Fouse, S., Delaney, A., Zhao, Y., Olshen, A., Ballinger, T., Zhou, X., Forsberg, K., Gu, J., Echipare, L., O'Geen, H., Lister, R., Pelizzola, M., Xi, Y., Epstein, C., Bernstein, B., Hawkins, D., Ren, B., Chung, W., Gu, H., Bock, C., Gnirke, A., Zhang, M., Haussler, D., Ecker, J., Li, W., Farnham, P., Waterland, R., Meissner, R., Marra, M., Hirst, M., Milosavljevic, A. & Costello, J. (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications, *Nature Biotechnology,* **28,** 1097-1105

Hartwig, B., James, G. V., Konrad, K., Schneeberger, K. & Turck, F. (2012) Fast isogenic Mapping-by-Sequencing of Ethyl Methanesulfonate-Induced Mutant Bulks, *Plant Physiology,* **160,** 591-600

Hashida, S. N., Uchiyama, T., Martin, C., Kishima, Y., Sano, Y. & Mikami, T. (2006) The Temperature-Dependent Change in Methylation of the *Antirrhinum* Transposon Tam3 Is Controlled by the Activity of Its Transposase, *The Plant Cell*, **18,** 104-118

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

Heng, Li, Broad Institute, 2010. *Aligning new-sequencing reads by BWA* [online] Available at: http://www.broadinstitute.org/files/shared/mpg/nextgen2010/nextgen_li.pdf [Accessed 23 November 2011]

Ho, S. (2008) The Molecular Clock and Estimating Species Divergence, *Nature Education,* **1** (1), 168

Hoad, S. P., Davies, D. H. K. & Topp, C. F. E. (2006) How to select varieties for organic farming: science and practice in *Aspects of Applied Biology 79, What will organic farming deliver?* (eds. Atkinson, C., Ball, B., Davies, D. H. K., Rees, R., Russell, G., Stockdale, E. A., Watson, C. A., Walker, R. & Younie, D.) 117-120 (*COR 2006*, Association of Applied Biologists)

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

Illumina, 2011. *Sequencing Technology* [online] Available at: http://www.Illumina.com/technology/sequencing_technology.ilmn [Accessed 09 November 2011]

Illumina, 2014. *Sequencing systems | Sequencer comparison table* [online] Available at: http://www.Illumina.com/systems/sequencing.ilmn [Accessed 05 March 2014]

Illumina, 2014 (b). *HiSeq X Ten | 1000 dollar genome sequencing* [online] Available at: http://www.Illumina.com/systems/hiseq-x-sequencing-system.ilmn

International Wheat Genome Sequencing Consortium (IWGSC) (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome, *Science*, **345** (6194), DOI: 10.1126/science.1251788

Iqbal, Z., Caccamo, M., Turner, I., Filcek, P. & McVean, G. (2012) De novo assembly and genotyping of variant using coloured de Bruijn graphs, *Nature Genetics,* **44,** 226-232

Jain. M., Nijhawan, A., Arora, R. Agarwal, P., Ray, S., Sharma, P., Kapoor, S., Tyagi, A. & Khurana, J. (2007) F-Box Proteins in Rice. Genome-Wide Analysis, Classification, Temporal and Spatial Gene Expression during Panicle and Seed Development, and Regulation by Light and Abiotic Stress, *Plant physiol.,* **143,** 1467-1483

James, G. V., Patel, V., Nordström, K. J., Klasen, J. R., Salomé, P. A., Weigel, D. & Schneeberger, K. (2013) User guide for mapping-by-sequencing in *Arabidopsis, Genome Biology,* **14,** 61

Jiang, K. & Goertzen, L. (2011) Spliceosomal intron size expansion in domesticated grapevine (*Vitis vinifera*), *BMC Research Notes,* **4,** 52

Kawaura, K., Mochida, K., Enju, A., Totoki, Y., Toyoda, A., Sakaki, Y., Kai, C., Kawai, J., Hayashizaki, Y., Seki, M., Shinozaki K. & Ogihara Y. (2009) Assessment of adaptive evolution between wheat and rice as deduced from full-length common wheat cDNA sequence data and expression patterns, *BMC Genomics*, **10,** 271

Kermicle, J. L. (1978) Imprinting of gene action in maize endosperm in *Maize Breeding and Genetics* (ed Wladen, D. B.) 357–371 (John Wiley and Sons, New York, 1978)

Kersey, P., Allen, J., Christensen, M., Davis, P., Falin, L., Grabmueller, C., Hughes, D. S., Humphrey, J., Kerhornou, A., Khobova, J., Langridge, N., McDowall, M. D., Maheswari, U., Maslen, G., Nuhn, M., Ong, C. K., Paulini, M., Pedro, H., Toneva, I., Tuli, M. A., Walts, B., Williams, G., Wilson, D., Youens-Clark, K., Monaco, M. K., Stein, J., Wei, X., Ware, D., Bolser, D. M., Howe, K. L., Kulesha, E., Lawson, D. & Staines, D. M. (2013) Ensembl Genomes 2013: scaling up access to genome-wide data, *Nucleic acids research*, epub ahead of print

Kim, Y., Schumaker, K. & Zhu, J. (2006) EMS Mutagenesis of Arabidopsis in Methods in Molecular Biology, vol. 323:Arabidopsis Protocols, Second Edition (eds. Salinas, J. & Sanchez-Serrano, J. J.) (Humana Press Inc., Totowa, NJ)

Kluge, C., Lamkemeyer, P., Tavakoli, N., Gollgack, D., Kandlbinder, A. & Dietz, K. J. (2003) cDNA cloning of 12 subunits of the V-type ATPase from *Mesembryanthemum crystallinum* and their expression under stress, *Molecular Membrane Biology,* **20,** 171-183

Koboldt, D., Chen, K., Wylie, T., Larson, D., McLellan, Mardis, E., Weinstock, G., Wilson, R. & Ding, L. (2009), VarScan: variant detection in massively parallel sequencing of individual and pooled samples, *Bioinfomatics,* **17,** 2283-5

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

Kovarik, A., Matyasek, R. & Fojtova, M. (2001) Cytosine methylation of plastid genome in higher plants. Fact or artifact? *Plant Sci.,* **160** (4), 585-593

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M. & Turner, D. J. (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes, *Nature Methods,* **6**, 291-295

Krueger, F. & Andrews, S. R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications, *Bioinformatics,* **27 (**11), 1571-1572

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J. & Marra, M. A. (2009) Circos: an Information Aesthetic for Comparative Genomics*, Genome Res.*, **19**, 1639-1645

Lamoureux, D., Peterson, D. G., Li, W. & Fellers, J. P. (2005) The efficacy of Cot-based gene enrichment in wheat (*Triticum aestivum L.*), *Genome,* **48** (6), 1120-1126

Langevin, S. & Kelsey, K. (2013) The Fate Is Not Always Written in the Genes: Epigenomics in Epidemiologic Studies, *Environmental and Molecular Mutagenesis,* **7,** 533-41

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biology*, **10:**R25

Lei, L., Zhu, X., Wang, S., Zhu, M., Carver, B. & Yan, L. (2013) TaMFT-A1 Is Associated with Seed Germination Sensitive to Temperature in Winter Wheat, *PLoS ONE,* **8**(9), e73330

Li, H. & Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform, *Bioinformatics*, **25**, 1754-6

Li, H. & Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **5,** 589-95

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools, *Bioinformatics*, **25**, 2078-9

Li, Y., Niu, Y. & Chen, X (2009) Mapping a stripe rust resistance gene YrC591 in wheat variety C591 with SSR and AFLP markers, *Theor. Appl. Genet.*, **118**, 339–346

Life technologies, 2014. *Ion Proton System specifications* [online] Available at:

http://www.lifetechnologies.com/uk/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-proton-system-for-next-generation-sequencing/ion-proton-system-specifications.html [Accessed 05 March 2014]

Liu, S. X. & Anderson, J. A. (2003) Targeted molecular mapping of a major wheat QTL for fusarium head blight resistance using wheat ESTs and synteny with rice, *Genome,* **46**, 817-823

Lister, R. O., O'Malley, R. C., Tonti-Filippinin, J., Gregory, B. D., Berry, C. C., Millar, A. H. & Ecker, J. R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis, *Cell,* **133,** 523-36

Locke, J.C., Kozma-Bognar, L., Gould, P.D., Feher, B., Kevei, E., Nagy, F., Turner, M. S., Hall, A. & Millar, A. J. (2006) Experimental validation of a predicted feedback loop in the multi-oscillator clock of Arabidopsis thaliana, *Mol. Syst. Biol.*, **2**(59)

Lukens, L. & Zhan, S. (2007) The plant genome's methylation status and response to stress: implications for plant improvement, *Current Opinion in Plant Biology*, **10**, 317–322

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

Mane, S. P., Modise, T. & Bruno, W. (2011) Sobral Analysis of High-Throughput Sequencing Data in *Plant Reverse Genetics: Methods and Protocols* (ed. Andy Pereira), Methods in Molecular Biology, vol. 678, 1-11 (Springer Science+Business Media)

Maniatis, N., Collins, A., Gibson, J., Zhang, W., Tapper, W. and Morton, N. (2004) Positional cloning by linkage disequilibrium, *AJHG,* **74** (5), 846-855

Mardis, E. R. (2008) Next-generation DNA sequencing methods, *Annu. Rev. Genomics Hum. Genet.,* **9,** 387-402

Marra, M., Kucaba, T., Sekhon, M., Hillier, L., Martienssen, R., Chinwalla, A., Crockett, J., Fedele, J., Grover, H., Gund, C., McCombie, W. R., McDonald, K., McPherson, J., Mudd, N., Parnell, L., Schein, J., Seim, R., Shelby, P., Waterson, R. & Wilson, R. (1999) A map for sequence analysis of the *Arabidopsis thaliana* genome, *Nature Genetics,* **22,** 265-270

McHale, L., Tan, X., Koehl, P. & Michelmore, R. (2006) Plant NBS-LRR proteins: adaptable guards, *Genome Biology,* **7,** 212

McIntosh, R. A.,Hart, G. E., Devos, K. M., Gale, M. D. & Rogers, W. J. (1998) *Catalogue of gene symbols for wheat*, Slinkard A E. Proceedings of the 9th International Wheat Genetics Symposium, Saskatoon: University Extension Press, 1–235

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M. A. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Genome Res.*, **20**(9), 1297-303

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

Middleton, C., Senerchia, N., Stein, N., Akhunov, E., Keller, B., Wicker, T. & Kilian, B. (2014) Sequencing of Chloroplast Genomes from Wheat, Barley, Rye and Their Relatives Provides a Detailed Insight into the Evolution of the Triticeae Tribe, *PLOS one,* DOI: 10.1371/journal.pone.0085761

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

Mizuno, N., Nitta, M., Sato, K. and Nasuda, S. (2012) A wheat homologue of PHYTOCLOCK 1 is a candidate gene conferring the early heading phenotype to einkorn wheat. *Genes Genet. Syst*., **87,** 357-367

Mooney, E. H & McGraw, J. B. (2007) Effects of self-pollination and outcrossing with cultivated plants in small natural populations of American ginseng, *Panax quinquefolius* (Araliaceae), *American Journal of Botany,* **94**, 1677-1687

Mozo, T., Dewar, K., Dunn, P., Ecker, J. R., Fischer, S., Kloska, S., Lehrach, H., Marra, M., Martienssen, R., Meier-Ewert, S. & Altmann, T. (1999) A complete BAC-based physical map of the *Arabidopsis thaliana* genome, *Nature Genetics,* **22,** 271-275

Nadeau, J. H. (2001) Modifier genes in mice and humans, *Nature Reviews Genetics,* **2**(3), 165-174

New England Biolabs, 2009. *Studying DNA methylation* [online] Available at: http://www.neb.com/nebecomm/tech_reference/epigenetics/epigenetics_technology.asp [Accessed 13 june 2012]

Nickerson Ltd. (2007) *Xi19 winter wheat* [online] Available at: http://www.cerealsdb.uk.net/CerealsDB/Documents/PDFs/Xi19.pdf [Accessed 14 October 2013]

NimbleGen, Roche, 2011. *NimbleGen Sequence capture:Discover more sequence less* [online] Available at: http://www.nimblegen.com/products/lit/05227887001_SeqCapBroch_Oct2011.pdf [Accessed 09 November 2011]

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

Pacific Biosciences, 2014. *SMRT Technology* [online] Available at: http://www.pacificbiosciences.com/products/smrt-technology/ [Accessed 05 March 2014]

Pacific Biosciences, 2014 (b). *Understanding Accuracy in SMRT Sequencing* [online] Available at: http://www.pacificbiosciences.com/pdf/Perspective_UnderstandingAccuracySMRTSequencing.pdf [Accessed 03 June 2014]

Pollard, D. A. (2012) Design and construction of Recombinant Inbred Lines, in *Quantitative Trait Loci (QTL): Methods and Protocols, Methods in Molecular Biology* (ed. Rifkin, A.) (New York: Humana Press)

Quarrie, S. A., Lazic-Jancic, V., Kovacevic, D., Steed, A. & Pekic, S. (1999) Bulk segregant analysis with molecular markers and its use for improving drought resistance in maize, *Journal of Experimental Botany,* **50** (337), 1299-1306

Quince, C., Lanzen, A., Curtis, T. P., Davenport, R. J., Hall, N., Head, I. M., Read, L. F. & Sloan, W. T. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods,* **6**, 639–641

Rabinowicz, P. D., Citek, R., Budiman, M. A., Nunberg, A., Bedell, J. A., Lakey, N., O'Shaughnessy, A. L., Nascimento, L. U., McCombie, W. R. & Martienssen, R. A. (2005) Differential methylation of genes and repeats in land plants, *Genome Res*., **15**, 1431-1440

Ratnaparkhe, M. B., Wang, X., Li, J., Compton, R. O., Rainville, L. K., Lemke, C., Kim, C., Tang, H. & Paterson, A. H. (2011) Comparative analysis of peanut NBS-LRR gene clusters suggests evolutionary innovation among duplicated domains and erosion of gene microsynteny, *New Phytologist,* **192,** 164-178

Risk, J. M., Laurie, R. E., Macknight, R. C. & Day, C. L. (2010) FRIGIDA and related proteins have a conserved central domain and family specific N- and C- terminal regions that are functionally important, *Plant Mol. Biol.,* **73,** 493-505

Roche, 454 Life Sciences, 2006. *Genome Sequencer System-Whole Genome Assembly using paired end reads in E. coli, B. licheniformis and S. cerevisiae* [online] Available at: http://www.genoseq.ucla.edu/images/2/29/Paired_end.pdf [Accessed 13 March 2012]

Roche, 454 Life Sciences, 2007. *How is genome sequencing done?* [online] Available at: http://www.454.com/downloads/news-events/how-genome-sequencing-is-done_FINAL.pdf [Accessed 13 March 2012]

Rothberg, J. M. & Leamon, J. H. (2008) The development and impact of 454 sequencing, *Nature Biotechnology,* **26**, 1117-1124

Round, E. K., Flowers, S. K. & Richards, E. J. (1997) Arabidopsis thaliana Centromere Regions: Genetic map positions and repetitive DNA structure, *Genome Res.,* **7**, 1045-1053

Ruffalo, M., LaFramboise, T. & Koyutu, M. (2011) Comparative analysis of algorithms for next-generation sequencing read alignment, *Bioinformatics,* **27** (20), 2790-2796

Sanger, F., Nicklen, S. & Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors, *Proc. Natl. Acad. Sci. USA,* **74**, 5463–5467

Saintenac, C., Jiang, D. & Akhunov, E. D. (2011) Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome, *Genome Biology*, **12**, 1-17

Sarin, S., Prabhu, S,. O'Meara, MM., Pe'er, I. & Hobert, O. (2008) Caenorhabditis elegans mutant allele identification by whole-genome sequencing, *Nat. Methods*, **5**, pp865-867

Sato, S., Kotani, H., Nakamura, Y., Kaneko, T., Asamizu, E., Fukami, M., Miyajima, N. & Tabata, S. (1997) Structural Analysis of *Arabidopsis thaliana* Chromosome 5. I. Sequence Features of the 1.6 Mb Regions Covered by Twenty Physically Assigned PI Clones, *DNA Research,* **4,** 215-230

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A. H., Nielsen, K. L., Jørgensen, J., Weigel, D. & Andersen, S. U. (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing, *Nature Methods*, **6** (8), 500-501

Schuster, S. C. (2008) Next-generation sequencing transforms today's biology, *Nature methods,* **5** (1), 16-18

Shewry, P. R. (2009) Wheat, *Journal of Experimental Botany*, **60**(6), 1537–1553

Smith, D. R., Quinlan, AR., Peckham, H. E., Makowsky, K., Tao, W., Woolf, B., Shen, L., Donahue, W. F., Tusneem, N., Stromberg, M. P., Stewart, D. A., Zhang, L., Ranade, S. S., Warner, J. B., Lee, C. C., Coleman, B. E., Zhang, Z., McLaughlin, S. F., Malek, J. A., Sorenson, J. M., Blanchard, A. P., Chapman, J., Hillman, D., Chen, F., Rokhsar, D. S., McKernan, K. J., Jeffries, T. W., Marth, G. T. & Richardson, P. M. (2008) Rapid whole-genome mutational profiling using next-generation sequencing technologies, *Genome Res.*, **18**, 1638-1642

Smith, D. & Flavell, R. (1975). Characterization of the wheat genome by renaturation kinetics, *Chromosoma,* **50**, 223–242

Sourceforge.net, 2009. *Picard* [online] Available at: http://picard.sourceforge.net [Accessed 07 March 2014]

Spannagla, M., Mayera, K., Durnerb, J., Haberera, G. & Fröhlichb, A. (2011) Exploring the genomes: From Arabidopsis to crops, *Journal of Plant Physiology,* **168***,* 3-8

Sulonen, A. M., Ellonen, P., Almusa, H., Lepistö, M., Eldfors, S., Hannula, S., Miettinen, T., Tyynismaa, H., Salo, P., Heckman, C., Joensuu, H., Raivio, T., Suomalainen, A. & Saarela, J. (2011) Comparison of solution-based exome capture methods for next generation sequencing, *Genome Biol*., **12,** 94

Tanksley, S. D., Ganal, M. W. & Martin, G. B. (1995) Chromosome landing: a paradigm for map-based gene cloning in plants with large genomes, *Trends in genetics,* **11 (**2), 63-68

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana, Nature,* **408,** 796-815

The Arabidopsis Information Resource (TAIR), 2011. [online] Available at: http://www.arabidopsis.org/index.jsp [Accessed 23 November 2011]

The International Barley Genome Sequencing Consortium (2012) A physical, genetic and functional sequence assembly of the barley genome, *Nature,* **491,** 711-716

The International Barley Genome Sequencing Consortium (2012b). *Sequencing data download FTP site* [online] Available at: ftp://ftpmips.helmholtz-muenchen.de/plants/barley/public_data/ [Accessed 16th June 2014]

The International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon, Nature,* **463,** 763-768

The iPlant Collaborative, 2011. [online] The iPlant Collaborative: Cyberinfrastructure for Plant Biology. Available at: https://www.iplantcollaborative.org [Accessed 14 April 2014]

The National Science Foundation, 2009. *Arabidopsis: The model plant*. [online] Available at: http://www.nsf.gov/pubs/2002/bio0202/model.htm [Accessed 20 November 2010]

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

United States Department of Agriculture, 2012. *World Agricultural Supply and Demand Estimates. Report No. WASDE-511* [online] Available at: http://usda01.library.cornell.edu/usda/current/wasde/wasde-10-11-2012.pdf

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J. A., Costa, G., McKernan, K., Sidow, A., Fire, A. & Johnson, S. M. (2008) A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res*, **18,** 1051-1063

Wang, Z., Gerstein, M. & Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics, *Nature Reviews Genetics,* **10,** 57-63

Wang, W., Zhao, X., Pan, Y., Zhu, L., Fu, B. & Li, Z. (2011a) DNA methylation changes detected by methylation-sensitive amplified polymorphism in two contrasting rice genotypes under salt stress. *Journal of Genetics and Genomics* , **38,** 419-424

Wang, J., Liu, D., Guo, X., Yang, W., Wang, X., Zhan, K. & Zhang, A. (2011b) Variability of Gene Expression After Polyhaploidization in Wheat (*Triticum aestivum* L.), *G3 Genes Genomes Genetics*, **1,** 27-33

Watanabe, S., Mizoguchi, T., Aoki, K., Kubo, Y., Mori, H., Imanishi, S., Yamazaki, Y., Shibata, D. & Ezura, H. (2007) Ethylmethanesulfonate (EMS) mutagenesis of Solanum lycopersicum cv. Micro-Tom for large-scale mutant screens, *Plant Biotechnology,* **24,** 33-38

Wei, Q., Claus, R., Heilscher, T., Mertens, D., Raval, A., Oakes, C., Tanner, S. M., Chapelle, A., Byrd, J., Stilgenbauer, S. & Plass, C. (2013) Germline Allele-Specific Expression of DAPK1 in Chronic Lymphocytic Leukemia, *PLOS one*, **8**, 1

Weigel, D. (2012) Natural Variation in Arabidopsis: From Molecular Genetics to Ecological Genomics, *Plant Physiology,* **158,** 2-22

Wellings, C. R. & McIntosh, R. A. (1990) Puccinia striiformis f. sp. tritici in Australia: pathogenic changes during the first 10 years, *Plant Pathol*, **39**, 316–325

Wicker, T., Yahiaoui, N., Guyot, R., Schlagenhauf, E., Liu, Z D., Dubcovsky, J. & Keller, B. (2003) Rapid Genome Divergence at Orthologous Low Molecular Weight Glutenin Loci of the A and Am Genomes of Wheat, *The Plant Cell*, **15**, 1186–1197

Widman, N. Jacobsen, S. E. & Pellegrini, M. (2009) Determining the conservation of DNA methylation in Arabidopsis, *Epigenetics,* **4 (**2), 119-124

Winfield, M., University of Bristol, 2011. Wheat Genomics:*Wheat BP...the big picture* [online] Available at: http://www.cerealsdb [Accessed 12 March 2014]

Winfield, M. O., Wilkinson, P. A., Allen, A. M., Barker, G. L., Coghill, J. A., Burridge, A., Hall, A., Brenchley, R. C., D'Amore, R., Hall, N., Bevan, M. W., Richmond, T., Gerhardt, D. J., Jeddeloh, J. A. & Edwards, K. J. (2012) Targeted re-sequencing of the allohexaploid wheat exome. *Plant Biotechnology Journal*, **10**, 733–742

Wisconsin Crop Improvement Association. (2011) *Winter wheat data for 2011 planting* [online] Available at: http://wcia.wisc.edu/Winter_Wheat_2011.pdf [Accessed 11 October 2013]

Xu, Y. (2010) *Molecular plant breeding*. CABI

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W., Chen, H., Henderson, I. R., Shinn, P., Pellegrini, M., Jacobsen, S. E. & Ecker, J. R. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis, *Cell*, **126**, 1189-1201