Linkage disequilibrium analysis of hexaploid wheat (*Triticum aestivum* L.)

A thesis presented for the degree of Doctor of Philosophy

by

Sherri Anne Kruger

The School of Agriculture, Food and Wine
The University of Adelaide

June 2007

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Acronyms

μg	Micrograms
μL	Microliter
ABI	Applied biosystems incorporated
AFLP	Amplified fragment length polymorphism
Amp	Ampicillin
APS	Ammonium persulfate
ATP	Adenosine triphosphate
BAC	Bacterial artificial chromosome
bp	Base pairs
°C	Degree celsius
CIMMYT	International centre for maize and wheat improvement
CS	Chinese spring
cM	Centimorgan
DaRT	Diversity array technology
DH	Doubled haploid
DNA	Deoxyribonucleic acid
DNTPs	Deoxynucleotide triphosphate
EDTA	Ethylene diamine tetra-acetic acid
EST	Expressed sequence tag
EtBr	Ethidium bromide
EtOH	Ethyl alcohol
g	Gram
GDM	Gatersleben D-genome microsatellite
GSP	Grain softness protein
GWM	Gatersleben wheat microsatellite
HCl	Hydrochloric acid
HMW	High molecular weight
IAA	Isoamyl alcohol
IPTG	Isopropyl β-D-galactopyranoside
ITMI	International triticeae mapping initiative
Kb	Kilobase
L	Litre
LB	Luria bertani broth
LBA	LB with bacto-agar

LD	Linkage disequilibrium
LOD	Log likelihood
MAS	Marker assisted selection
$MgCl_2$	Magnesium chloride
mL	Millilitre
mM	Millimolar
MWM	Molecular weight marker
N_2	Nitrogen
NaCl	Sodium chloride
NaOH	Sodium hydroxide
NCBI	National centre of biotechnology information
ng	Nanogram
NIL	Near isogenic line
nmol	Nanomole
NT	Nullisomic tetrasomic
PCR	Polymerase chain reaction
pH	Potential hydrogen
Pina	Puroindoline a
Pinb	Puroindoline b
PVPP	Polyvinyl polypyrrolidone
QTL	Quantitative trait locus
R40	RNaseA
RAPD	Randomly amplified fragment length polymorphism
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
RNase	Ribonuclease
RO	Reverse osmosis
SDS	Sodium dodecyl sulfate
SNP	Single nucleotide polymorphism
SOC	Salt optimized broth + carbon
SSR	Simple sequence repeats
STS	Sequence tagged site
TAE	Tris-HCl, acetic acid and EDTA
TBE	Tric-HCl, boric acid and EDTA
TE	Tris-HCl and EDTA
Tm	Annealing temperature
U	Units
UV	Ultra violet

V	Volts
VNTI	Vector NTI®
WMC	Wheat microsatellite consortium
WMS	Wheat microsatellite
X-GAL	5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside

Abstract

There has recently been a renewed interest in using a whole-genome approach for identifying regions with relatively small effect on a particular trait of interest. One method that has proven effective in human populations is association mapping or linkage disequilibrium (LD) mapping. With focus on identifying the statistical correlations between marker allele frequency and phenotypes, association mapping, as a result, typically requires a high density marker map and a firm understanding of the extent and patterns of LD in the population.

This study assesses the feasibility of applying LD mapping in hexaploid wheat research for the fine mapping of traits. Adequate marker coverage of the large wheat genome was attained providing a framework enabling the examination of the extent of LD in this species. Results presented in this thesis illustrate how extensive LD is in locally adapted populations of hexaploid wheat, extending up to 100cM in some cases. It is also apparent that statistical associations are not limited only to markers on the same chromosome but include those on different genomes and chromosome groups. One of the main focuses of this study was to evaluate the effect of genetic and evolutionary factors on the levels of statistically significant LD. Type-1 error rate was successfully reduced by accounting for population structure and the presence of rare alleles in the data sets. This research has provided a base from which patterns of LD can begin to be understood in other populations and subsequently assess the applications of association mapping in inbreeding crop species, specifically *Triticum aestivum* L.

Declaration

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being made available in all forms of media, now or hereafter known.

_____ _____

Sherri A. Kruger Date

Acknowledgements

I would like to sincerely thank my supervisors, Professor Peter Langridge, and Dr Jason Able for their guidance, advice and patience during my studies. They both have a passion for science and learning that is truly inspiring. Peter's insight into my project and assistance in allowing me to see the big picture is greatly appreciated. I am indebted to Jason for the seemingly endless editing of chapters and for assisting me with the submission of this manuscript. Both Peter and Jason were always enthusiastic about my project, many many thanks to you both.

Special thanks to Drs Ute Bauman and Andreas Schrieber for their statistical support and genuine interest in my project over the past 4 years. Dr Penny Henschke, and Alison Hay for their assistance in the genotyping and endless trouble shooting of the loveable 3700. I would also like to acknowledge Drs Klaus Oldach, Tim Sutton and Amanda Able for their friendship and encouragement throughout my studies. And of course my "Aussie sisters", Pat Warner and Elise Tucker, I am forever thankful for them lending me their superior lab skills over the years but most of all for their friendships. It was an absolute pleasure working with the rest of the Langridge lab group. Their support, patience and making the lab more than "just a place to work" made the transition from Canada and every year after much more manageable.

I would also like to thank Drs Greg Penner and Peter Jack for their support and for encouraging me to do a PhD in the first place. Without them I may never have gotten myself into this!

I am grateful to the CRC for Molecular Plant Breeding who financially supported this project throughout my studies and who also provided me with a postgraduate research scholarship.

I wish to thank my Mom and Dad, Ma, Walt, Grannie, Darren and my two beautiful sisters Jenni and Chrissy, for absolutely everything. The messaging sessions and never ending phone calls made us seem not so far apart. The visits to Australia gave us a lot to look forward to and the care packages from home got us through many of the difficult times. (And Chrissy you still are the WR champ!).

Most importantly, I am forever grateful for the love and support of my husband, Gwynn, who stood by me every step of the way. He gave up life in Canada to allow me to realise my dreams in Australia. We had many adventures along the way and an experience of a lifetime. Go the transpositron! ☺

Chapter 1: General Introduction

1.1 An introduction to wheat

Wheat is arguably the most important cereal crop grown worldwide. It provides the basis of human nutrition and is a commodity of significant global economic importance. In 2005 more than 624 million tonnes of wheat were produced world wide, which is only slightly less than maize but considerably more than rice where 710 and 421.9 million tonnes respectively, were produced world wide (FAO 2006). As members of the Poaceae family and Triticeae tribe, wheat species are typically characterized by a compound spike, laterally compressed spikelets and two glumes (Lupton 1987). The grains are large, plump and, in most species, separated from the lemma when threshed. Morphologically, bread wheat is distinguished from other members of the *Gramineae* family by hollow stems and glumes which are typically awnless (Peterson 1965).

1.2 A complex genome

Genetically, wheat is divided into three discrete groups; diploid (*Triticum monococcum*), tetraploid (Durum wheat) and hexaploid (bread wheat). *Triticum aestivum*, common or bread wheat, is the most common form of wheat cultivated at present. As an allopolyploid, *T. aestivum* arose from the hybridization of three fully differentiated genomes; A, B, and D. It is believed that this domesticated allohexaploid is the result of at least two ancient hybridization events (Figure 1.1) involving three diploid wheat species, each contributing one of the three genomes of wheat.

Recurrent hybridization is thought to be the most likely mechanism of speciation of bread wheat and intrinsically there is an evolutionary advantage to this method over the theory of a single hybridization event leading to this polyploid

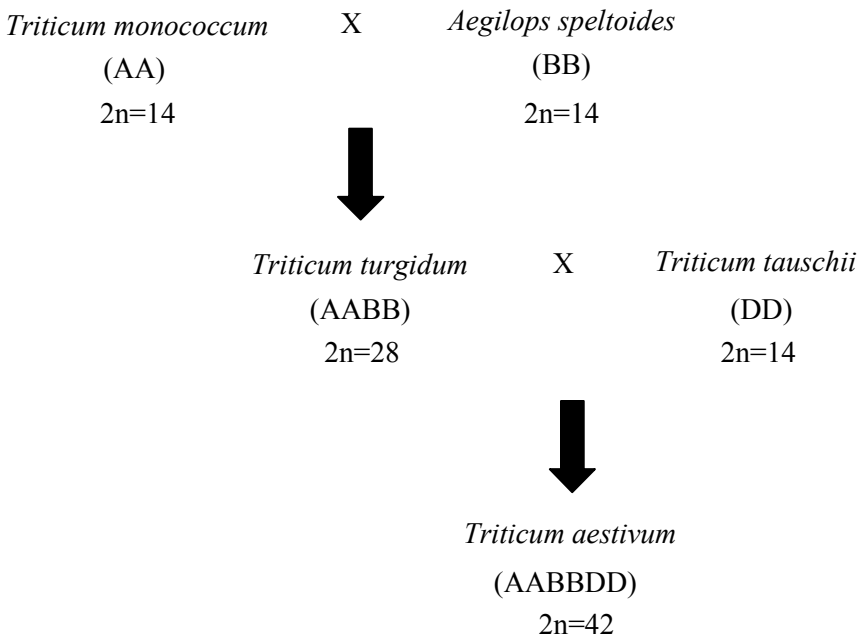


FIGURE 1.1 Flow chart illustrating the origin of hexaploid wheat, *Triticum aestivum*. Hybridisation events between diploid wheat species and diploid/tetraploid wheat species are listed resulting in *T. aestivum*, hexaploid wheat.

species (Soltis and Soltis 1999). In the theory of recurrent hybridization, polyploidization is thought to occur over and over, each time allowing the genetically distinct parents to hybridize into an allopolyploid organism thereby incorporating the genetic diversity from each of the founding populations. This allows an increase in the genetic diversity of the resulting population.

By understanding this genetic diversity in modern wheat varieties as well as ancient varieties, geneticists are able to apply biotechnology and molecular genetic techniques that will continue to enhance varieties. However, there are several aspects of the wheat genome that add to the complexity of breeding and genetic analysis. Firstly wheat has a very large genome size. Estimated to be roughly 16,000 MB, the wheat genome is considerably larger than that of the model crop rice *Oryza sativa* (466 MB) (Yu et al. 2002) and more than 6 times larger than maize (2,500 MB) (Arumuganathan and Earle 1991). Secondly, the polyploid nature of the genome, and the resulting triplication of genetic loci, adds to the complexity of genetic analysis. Analyzing fingerprinting results becomes a challenge when, particularly with restriction fragment length polymorphism (RFLP) analysis, there is quite often the detection of three distinct loci or bands, one from each genome. The benefit of this complexity lies in the fact that three or more loci may be mapped simultaneously to one of the 21 linkage groups of wheat (Nelson et al. 1995). Lastly, there is the repetitive and low polymorphism nature of the genome that, when combined with the aforementioned features, makes genetic analysis of wheat particularly difficult. There are some challenges with respect to the meiotic mapping of wheat due in part to the >80% repetitive DNA (Hayden and Sharp 2001), and low polymorphism levels, particularly in the D genome (Chalmers et al. 2001; Chao et al. 1989; Nelson et al. 1995).

1.3 Genetic tools

Molecular markers, meiosis-derived linkage maps and sophisticated software packages, such as Arlequin (Excoffier et al. 2005), Dnasp (Rozas and Rozas 1999), JoinMap (Stam 1993) as examples, are just a few of the tools used by geneticists and plant breeders to enhance line selection and accelerate variety release times. Table 1.1 lists a recent selection of genetic maps constructed using a number of different molecular marker systems and where appropriate the targeted traits and genes are listed.

TABLE 1.1 A selection of genetic maps developed in *T. aestivum* sp. using a variety of molecular marker systems. Genetic control over several important agronomic traits has also been identified and the traits and/or genes are listed.

Number of loci	Marker Type	Traits	Reference
479	STM		(Hayden et al. 2006)
424	SSR	Fusarium head blight resistance	(Somers et al. 2006)
410	DArT, STM		(Akbari et al. 2006)
167	SSR, HMW glutenin subunits, STS-PCR	24 QTL detected for agronomic traits 26 QTL detected for quality-related traits	(Huang et al. 2006)
464	SSR		(Torada et al. 2006)
567	RFLP, AFLP, SSR, morphological and biochemical	QTL for yield and yield components	(Quarrie et al. 2005)
1, 235	SSR (Consensus)		Somers, 2004
173	SSR	Grain protein content	Prasad et al, 2003
550	RFLP, SSR	Awning	Sourdille et al, 2002

Development of these tools and techniques has significantly contributed to the enhancement of wheat cultivars (Bass et al. 2006; Pestsova et al. 2006; Smith et al. 2007), with the potential to reduce varietal delivery time by up to several years (Hospital et al. 1997).

1.4 Molecular Markers

The results from decades of conventional plant breeding have been substantial and include enhancements such as improved yields, resistance to pests and fungal pathogens, and quality traits. Conventional wheat breeding, however, is a time consuming process taking up to 10 to 15 years to release a new variety.

The development of molecular marker systems over the past few years has contributed to a greater understanding of genome organization and structure within wheat (Akhunov et al. 2003; Boyko et al. 2002; Sandhu and Gill 2002). RFLPs were the first molecular marker system used to generate linkage maps of cereal species (Young et al. 1988). Shortly there after RAPD (Williams et al. 1990) and AFLP (Vos et al. 1995) marker systems were used to add to the framework maps. Currently SSRs, or microsatellites, are the most widely used marker system in plant genetics with SNP development gaining momentum although still in its infancy.

Perhaps the most obvious advantage to implementing molecular markers in plant research is the ability to tag and track favorable alleles through a breeding program (Hayden et al. 2004). Plant material may then be screened for the desirable alleles at an early stage such that time and other resources are not spent on tending to undesirable plants.

The characteristics of the many molecular marker systems available for wheat genetic research are summarized in Table 1.2. These systems have proven effective in a number of studies including those involving; genetic diversity (Balfourier et al. 2007; Fu et al. 2006; Li et al. 2006), marker assisted selection (MAS) (Charmet et al. 2001; Gupta et al. 2005b), gene mapping (Bass et al. 2006; Liu et al. 2005; Sardesai et al. 2005), and association studies (Breseghello and Sorrells 2006; Rhone et al. 2007; Watanabe et al. 2006).

TABLE 1.2 Overview of characteristics of the most commonly used molecular marker systems for wheat.

Marker system	DNA quantity (ug)	DNA (quality)	PCR based	Reproducible	Amenable to HTP	Cost (development /use)	Technical difficulty	Poly-morphisms (per assay)
RFLP	5	High	No	Yes	No	Low/high	Difficult	3
RAPD	.2	High	Yes	Variable	Possible	Low/low	Simple	10
AFLP	.2	High	Yes	Yes	Possible	Med/med	Simple	~50
SSR	.2	Med	Yes	Yes	Yes	High/low	Simple	1-3
SNP	.2	Med	Yes	Yes	Yes	High/low	Simple	1

This section provides an overview of the most common marker systems currently employed in plant genetic research. For other reviews of this area see Langridge et al. (2001) and Gupta et al. (1999).

1.4.1 RFLPs

The RFLP marker system is a Southern hybridization-based technique that is the consequence of either presence or absence of restriction sites, specific sequences in the DNA which are distributed throughout the genome (Chao et al. 1989; Gardiner et

al. 1993; Young et al. 1988). The restriction sites will vary between individuals based on genomic position due to insertions, deletions, inversions, duplications and translocations; and in sequence due of transitions and transversions. The polymorphisms are revealed when total DNA is digested with a particular enzyme, separated by agarose gel electrophoresis and hybridized with a labeled probe. Since this is a co-dominant marker system it is possible to detect heterozygous individuals in a population since the probe generally hybridizes with both alleles.

RFLP probes are largely developed from cDNA clones and hence correspond to expressed genes (Young et al. 1988). It was the initial observation by Chao et al. (1989), of the preserved gene order in the three group 7 chromosomes of wheat, which played a pivotal role in establishing, what is now, the field of comparative genomics. Since there is a high level of genome conservation between related grass species, such as wheat, barley, rice, and maize, (Devos and Gale 2000; Hass et al. 2003) it has been possible to employ probe sets from a number of related species. Since these probes cross-hybridise, wheat RFLP maps consist of a large number of probes from rye, barley and oats (Chao et al. 1989; Paull et al. 1998).

The main reason for the decline in popularity of the RFLP marker system is the technically demanding and labor intensive nature of the assay. It has taken a back seat to PCR-based methods which are typically quicker and far less time consuming.

1.4.2 RAPDs

With the advent of the polymerase chain reaction (PCR) the field of molecular genetics was revolutionized. The first widely adopted marker system that took advantage of the time saving benefits of PCR was the RAPD system. RAPD polymorphisms were initially examined in humans, corn, soybean, and *N. crassa* by

the founders of the technology, Williams et al. (1990). Random segments of genomic DNA are amplified using 10-mer oligonucleotides synthesized with no prior knowledge of the genomic sequence. Polymorphisms are detected between individuals which, can arise due to variations in the genomic sequence at the priming site as well as the presence of large indels that change the size of the amplicon but not preventing amplification (Williams et al. 1990). These polymorphisms are visualised as a presence or absence of an amplicon, or an amplicon shift between individuals, following gel electorphoresis.

Although there are many features that make the RAPD marker system favorable over RFLPs including, low development and implementation costs, easy detection, and amenability to automation, there are several disadvantages that have rendered this method obsolete in most molecular marker programs. RAPD markers have a dominant mode of inheritance and as such are unable to distinguish between homozygous and heterozygous individuals. Furthermore, there is a distinct lack of reproducibility of results between labs due to DNA extraction methods as well as differences in equipment and polymerases (He et al. 1994; Jones et al. 1997).

Despite these drawbacks, once a RAPD marker is shown to be linked to a gene of interest it is relatively easy to convert such a marker to a sequence characterized amplified region (SCAR) (Masojc 2002). These SCAR markers facilitate the rapid detection (or lack) of the allele of interest and also reduces the background problems associated with RAPD visualization. SCAR markers are particularly useful in marker assisted selection applications (Masojc 2002).

1.4.3 AFLPs

AFLP is another PCR based marker system that relies on the selective amplification of restriction fragments generated from a whole genome digest (Vos et al. 1995). This a highly robust marker system that can be used on a variety of DNA samples regardless of their origin and complexity. There are four main steps involved in this technique. First the genomic DNA is digested with restriction enzymes (RE) most commonly a rare 6 bp cutter (*EcoRI*) and a frequent 4 bp cutter (*Mse1*). The second step involves the ligation of double stranded adaptors onto the ends of each restriction fragment thus providing a primer-binding site for subsequent amplifications. Fragments are then amplified by PCR using primers that are complementary to the adaptor sequences, have the RE recognition site and selective nucleotides. In organisms with complex genomes, such as wheat, this PCR amplification step is typically done in two stages in order to maximize the usefulness of the assay (Vos et al. 1995). There is an initial pre-amplification step that uses primers complementary to the adapter plus one additional selective nucleotide at the 3' end followed by a secondary, selective amplification that uses similar primers but with additional nucleotides (2-3) at the 3' end. By modifying the selective nucleotides in the primers the number of amplified fragments and hence detected polymorphisms can be controlled.

AFLP combines the reliability of RFLPs with the ease of PCR providing the additional benefits of reliability (Jones et al. 1997), robustness, and amenability to any type of DNA regardless of origin or complexity (Vos et al. 1995). This marker system has been useful in reducing gaps in framework maps of RFLPs (Lotti et al. 2000), although there is still a reduction in the levels of polymorphism in hexaploid wheat's D genome in comparison to the A and B genomes, which is consistent with observations from other marker systems (Chalmers et al. 2001; Roder et al. 1998).

Due to the large amount of data generated with each assay, applications of this marker system include map construction and saturating map regions harboring loci that contribute to commercially important traits (Parker et al. 1998). Another application lies with targeting specific chromosomal locations associated with genes controlling particular traits of interest via bulked segregant analysis (Michelmore et al. 1991). Bulked segregant analysis (BSA) involves comparing two pooled DNA samples at the molecular level. These pools are typically derived from a segregating F2 population originating from a single cross between two individuals varying for a particular trait of interest, such as a resistant individual and a susceptible individual. Individuals from the resultant segregating F2 population are separated into two groups, each representing the two extremes of the variation. DNA from, typically, 10-14 individuals in each group is pooled and through the use of molecular markers is checked for polymorphisms. The expectation is, any polymorphisms that are detected between the two bulked samples will reflect differences in regions that are tightly linked with the gene controlling the trait of interest since all other loci should be equally distributed throughout the two bulked samples. BSA using AFLP and other molecular markers, such as RFLPs and SSRs, has proven effective in identifying associations with important disease resistance genes in wheat such as leaf rust and stripe rust (William et al. 2006; William et al. 2003; Xu et al. 2005).

1.4.4 SSRs

Simple sequence repeats (SSRs) or microsatellites, are tandem repeats of a basic two to six nucleotide motif. They are abundant and located randomly throughout the genomes of most eukaryotic organisms (Roder et al. 1998). The variability of the number of these repeat motifs varies between individuals, which results in a highly polymorphic marker system. These sequence variations are identified by PCR amplification, which again is easier to perform than RFLP, and uses primers that have

been designed in the unique DNA sequence flanking the SSR. Allelic variation is thought to result from a process that causes the expansion of short repeated motifs called slipped-strand mispairing (Levinson and Gutman 1987). After this initial expansion further development of these regions is likely to be caused by unequal crossing-over, or unequal sister chromatid exchange, since there is a tendency for these regions to mispair (Levinson and Gutman 1987). The mutation rate has been estimated to be on the order of 10^{-4} in humans (Payseur and Nachman 2000) and 10^{-5} in maize (Vigouroux et al. 2002).

The high mutation rate results in an important source of highly polymorphic markers (Levinson and Gutman 1987). SSR markers are also easy to use and transfer between researchers and this has made the SSR marker system popular (Jones et al. 1997). SSRs are co-dominantly inherited and polymorphisms are easily detected by a shift in mobility using gel based electrophoresis detection methods. For all the aforementioned reasons SSR markers are particularly useful in genetic diversity and evolutionary studies (Lee et al. 1995; Reif et al. 2004), as well as trait tagging and marker assisted selection (MAS) (Hayden et al. 2004). All of the features of this marker system mentioned so far contribute to a reduction in cost and manual labor of each assay when compared to other marker systems. The SSR marker system is also amenable to high-throughput methods and automation, which is important when integrating markers into a breeding program.

However, there are also drawbacks to this system. Most notable are the high cost and the great amount of time required for the development of such markers. In the discovery and validation stages of development, large numbers of potential markers are reduced to, in most cases, less than 30% of the original amount providing amplicons of the expected size (Roder et al. 1998). Squirrell et al. (2003) reviewed

the workload involved in developing SSR markers in plants and summarised the typical attrition steps during this marker discovery. Typical attrition steps, as outlined by Squirrell et al. (2003), include inability to design primers to target clone sequences since the microsatellite may appear to close to the end of an insert thus not having enough flanking sequence to design primers against. There are also failed PCR amplifications, complex multi-banded profiles due to multiple amplicon amplification, and limited genetic polymorphism resulting in monomorphic PCR products.

There are many international groups that are committed to the development and mapping of microsatellite markers in wheat. Within the last decade there has been significant contributions to the abundance of SSR markers in the public domain (Gupta et al. 2002; Pestsova et al. 2000; Roder et al. 1998) as well as through the wheat microsatellite consortium (http://wheat.pw.usda.gov/ggpages/SSR/WMC/ 2004). In 1998, a microsatellite map of wheat was published with 279 SSRs placed on the RFLP framework map of the International Triticeae Mapping Initiative (ITMI) reference population Opata 85 x W7984 (Roder et al. 1998). Since then, hundreds of other SSR markers have been developed and mapped (Eujayl et al. 2002; Pestsova et al. 2000), including a high-density consensus map of hexaploid wheat comprising 1,235 SSR markers and covering 2,569 cM at an average marker density of 1 every 2.2 cM (Somers et al. 2004) and a more recent addition of 347 SSR loci (xbarc) to the ITMI W7984 x Opata population by (Song et al. 2005).

With the continued commitment of these and other international groups, the resolution of the wheat genetic maps is likely to become even greater. This will facilitate the tagging and tracking of favorable alleles in breeding programs and map-

based cloning of agronomically important genes ultimately leading to transgenic plants for improved crops.

1.4.5 SNPs

Single nucleotide polymorphism (SNPs) markers are typically bi-allelic and are the marker system of choice in human genetic research due largely to; the recent completion of the human genome project (Macllwain 2002), the vast amount of available sequence data from a number of individuals, and the amenability to high-throughput detection methods (Kwok and Chen 2003). Approaches to SNP identification in genomes include direct sequencing of PCR products (Ravel et al. 2006) as well as data mining of public sequence databases (Somers et al. 2003). The former approach involves the PCR amplification of a specific region of the genome from several individuals. Once sequences are obtained they can be aligned and the variant loci (SNPs) easily detected. This works well for sequences that are homologous in nature but paralogy (presence of a duplicate copy in the genome) and homoeology (in the case of polyploids, a duplicate copy or multiple copies on the other genomes), may hinder the process of legitimate SNP detection (Ravel et al. 2006). The later method relies on taking advantage of the plethora of sequence data available in public databases, particularly EST databases (Batley et al. 2003; Somers et al. 2003). Sequences amplified from the same target region of the genome are acquired and aligned to identify sequence variations.

Identification and development of SNP markers in the genome of hexaploid wheat is already taking place (Ravel et al. 2006; Somers et al. 2003). A SNP discovery effort by Somers et al. (2003) consisted of more than 90, 000 wheat EST sequences aligned into contiguous sequences (contigs) of which 45 were selected to generate primers to amplify specific alleles. It was found that there was on average 1

SNP in every 540 bp between individual wheat genotypes (Somers et al. 2003); nearly two to three times the frequency observed in humans (1 SNP per 1Kb (Wang et al. 1998); 1 SNP per 1.6Kb (Dunham et al. 2004)); nearly twice as infrequent as that observed in a recent barley SNP discovery project (1 SNP per 200bp (Rostoks et al. 2005)) and nearly four times less frequent than observed in coding regions in a collection of maize elite lines (1 SNP in 124bp (Ching et al. 2002)). In a recent study of bread wheat, Ravel et al. (2006) amplified 21 genes from 26 geographically diverse lines and analysed SNPs based on their direct sequence alignment. The SNP frequency in this study was estimated to be 1 every 267bp in the coding regions which is over double that observed by Somers et al. (2003) EST based SNP estimate. This difference in SNP frequency was attributed to the difference in individuals studied; 26 geographically diverse lines in the Ravel et al. (2006) study as compared to 12 wheat lines chosen to represent those important to the Canadian grain industry by Somers et al. (2003).

Perhaps the most useful application of SNP markers in human genetic research will be in linkage disequilibrium analysis and gene mapping through association (Klein et al. 2005; Maraganore et al. 2005; Weiss and Clark 2002), and, although still in its infancy, the same is likely true for plant systems (Clark et al. 2004; Gonzalez-Martinez et al. 2006; Palaisa et al. 2004). There is also the potential for the generation of high density molecular maps due to SNP abundance within genomes. With the recent completion of the rice genome project (Genome 2005) there is now sufficient sequence data to allow primer design, SNP identification and mapping in this species. This was recently demonstrated in the study by Feltus et al. (2004) in which 408, 898 SNPs/Indels were identified between two rice subspecies, *indica* and *japonica,* with the aim of providing a marker resource useful for breeding and genetics and to provide a framework to carry out high density marker,

evolutionary or functional studies. Synteny between the rice and wheat genome has long been recognized with respect to gene sequences and order (Devos and Gale 1997). Although a recent study of microlinearity between the wheat and rice genomes reveals that researchers should procede with caution when using rice sequence and mapping data to determine the gene order and genome compostion of wheat (La Rota and Sorrells 2004). In this study 5,780 Triticeae EST sequences were compared to the rice genome sequence of 3,280 BAC/PAC clones. La Rota and Sorrells (2004) studied the relationship between the wheat deletion map and rice genomic sequence and observed that there was a breakdown of gene order and content. On average 35% of presumed single copy genes matched rice chromosomes other than the one that was the most similar. Despite this complicating the use of rice as a source of information in cross-species comparisons, La Rota and Sorrells (2004) note that the resolution of this comparative wheat/rice sequenced analysis will contribute to the selection of conserved genome regions for saturation mapping and identifying candidate genes which will aid in the the development of robust molecular markers and in the understanding of the evolution of the Triticeae genome.

1.5 Identifying and localizing traits of interest

Geneticists have been able to use molecular markers to generate detailed genetic linkage maps with the purpose obtaining a better understanding of genome organization and to deliver a method for localizing genes controlling traits of interest. With the escalating resource of polymorphic molecular marker data, genetic linkage maps are becoming more saturated (Somers et al. 2004) allowing for the precise positioning of genes. Recent examples include Somers et al. (2006) who successfully identified 2 QTL that controlled resistance to fusarium head blight (FHB) in a cross of tetraploid durham wheat; a study by Kuchel et al. (2006) in which QTL for dough rheology, dough strength, loaf volume, crumb quality, flour protein content, milling

yield and flour color were identified; and Williams et al. (2006) who identified a novel QTL contributing to the resistance of cereal cyst nematode (CCN) in wheat, which acts additively with QTL previously identified in wheat (Jahier et al. 2001; Williams et al. 2003).

The discovery of a quantitative trait locus (QTL) in inbreeding plant species is done with a mapping population consisting of F_2, BC_1, doubled haploids or recombinant inbred lines (Gupta 2002). In plant species this has been very effective with the identification of a number of genomic regions contributing to quantitative traits (Table 1.1).

Although QTL analysis has had a significant impact on the identification and localization of genetic loci that underpin commercially important traits in the wheat genome, there are still some issues that render it sub-optimal. Populations are generated using two parental genotypes that differ considerably in the trait of interest. The segregating populations are then scored phenotypically and genotyped with molecular markers and subsequently associations are made between the two. Results obtained are specific to this population and may not be transferable to other populations. In addition to population development and maintenance, there is the added drawback of the resolution with which these traits are mapped. Low map resolution is due to the limited number of meioses that have taken place in the generation of the mapping population and typically QTL are mapped only to within a few centimorgans (Ronin et al. 2003). Recently there has been a focus on establishing high resolution mapping populations in an attempt to fine map QTL underlying traits of interest (Cuthbert et al. 2006; Liu et al. 2006; Roder et al. 2007). However, even with these large popuations the resolutions with which these QTL are mapped are still on the order of several cM, in most cases. Cuthbert et al. (2006) successfully fine

mapped a major fusarium head blight (FHB) resistance gene to the short arm of chromosome 3B in two wheat populations, with map intervals of 1.27cM and 6.05cM. In another study Roder et al. (2007) sought to fine map the QTL for grain weight in wheat on chromosome group 7D and were successful in doing so to a map length interval of 7.6cM. Despite some of its short comings, QTL analysis for the detection of genomic regions contributing to economically important traits is not likely to be abandoned anytime soon. There is an increased focus, however on establishing and validating a new technique that is relatively new to crop genetic research, namely, association mapping.

1.6 Association mapping based on linkage disequilibrium

A potentially effective approach to identifying and localizing commercially important genes is association, or linkage disequilibrium (LD) mapping. This relies on drawing a statistical correlation between the molecular genotype and the phenotype of individuals in a collection of wheat accessions. This approach has the potential to localise genes more precisely than with QTL mapping. Association mapping relies on using large, historically-derived populations for which numerous generations of recombination have occurred and allows statistical correlations to be drawn between polymorphic marker data and phenotypic data over greatly reduced distances at the DNA level. Association mapping has proven to be particularly successful in the field of human genetics and has made significant contributions to elucidating disease susceptibility loci (del Bosque-Plata et al. 2004; Horikawa et al. 2000; Parsian et al. 2004; Shifman et al. 2002).

An important contributor to understanding LD in humans and the application of such has been the International HapMap Consortium (Consortium 2003). Established in 2002, the primary aim of the HapMap project is to create a publicly

accessible database of common DNA variants in the human genome using four distinct population samples with ancestries from parts of Africa, Asia and Europe. Phase I of this project contains over one million sequence variants (SNPs) that are polymorphic across the 269 individuals from the four populations mentioned earlier and with phase II underway there is expected to be another 4.6 million SNPs mapped in these populations as well as samples from additional populations (Consortium 2005). It is expected that these results will provide valuable information surrounding the transfereability of LD inferences from these populations to others.

One application of the HapMap data is to make genome-wide association studies possible. There are already several studies in which genome-wide SNP scans are used to draw associations between markers and causal variants (Klein et al. 2005; Maraganore et al. 2005; Ozaki et al. 2002). In one such genome-wide association study Klein et al. (2005) revealed variability within the compliment factor H (*CFH*) gene through the genotyping of just over 116,000 SNPs. This relatively small number of SNPs was used to genotype a small group of 96 cases and 50 controls. Initially strong association was identified between two SNPs, which lie in a LD block that is 41 kb long, and contained entirely within the *CFH* gene. To determine the polymorphism underlying susceptibility to the disease (age-related macular degeneration, AMD) Klein et al. (2005) resequenced all exons in the *CFH* gene, including those outside of the 41 kb LD block, for all 96 cases. It was a polymorphism in an exon located 2 kb upstream from the LD block that proved to be the most strongly associated polymorphism with the disease, increasing the risk of disease by 7 fold in homozygous carriers.

Maraganore et al. (2005) conducted a whole genome association study of Parkinson disease (PD) utilizing a 2 tiered approach. Initially 198,345 SNPs were

genotyped in 443 sibling case-control pairs. They found 1,793 SNP were significantly ($P < 0.01$) associated with PD. This SNP subset was used, in addition to 300 genomic control SNPs, in the subsequent tier 2 study in which Maraganore et al. (2005) 332 unrelated cases-control pairs were genotyped. Eleven SNPs were identified to be associated with PD in both tier 1 and 2 individuals. Four of the 11 SNPs were located in four genes that have compelling potential biological relationships with PD. The remaining 7 SNPs were in non-coding regions of the genome and includes one that is associated with a late-onset PD susceptibility locus.

The studies by Klein et al. (2005) and Maraganore et al. (2005) discussed above demonstrate the usefulness of genome wide association studies and although the field of human association genetics is far more advanced than that of plants, the past few years have shown an increasing interest in using LD and association mapping to localise economically important traits in plant systems (Ching et al. 2002; Kraft et al. 2000; Nordborg et al. 2002; Remington et al. 2001). In the first study of assessing the applications of association mapping in maize, Thornsberry et al. (2001) used association genetics to assess the *Dwarf8* gene and its affect on flowering time and plant height. Through sequence analysis of the *Dwarf8* gene (~3.7 Kb) in 92 inbred maize lines, in the absence of population structure, significant associations were observed between a suite of polymorphisms (9 in total) and differences in flowering time. The results from this study provide optimism with respect to applications of association mapping in plant systems. However, there are several differences between maize and other crop species, such as wheat, that will result in vastly different approaches to utilizing this method. Perhaps the biggest difference will be in the patterns and levels of LD throughout the different genomes.

In an early study by Paull et al. (1998), eight RFLP probes were found to be associated with a disease resistance locus present on an alien chromosome segment on the long arm of chromosome 6A in hexaploid wheat. These markers showed close to 100% association with the trait, despite five opportunities for recombination to break the linkage. This has direct implications for association mapping and LD analysis since this apparent lack of recombination will result in large segments that remain in tact and will be transferred from generation to generation as linkage blocks. This will generate extensive LD and so the most important prerequisite for any association mapping study is the complete understanding of the extent and distribution of LD in the experimental population under examination.

1.7 Extent and patterns of linkage disequilibrium in human genetics

LD in the human genome has been studied extensively and the distribution is highly variable due largely to the choice of experimental population, regions of the genome being examined as well as marker choice. Early estimates based on computer simulations, concluded that levels of LD, to be useful in association mapping efforts, is unlikely to extend beyond 3kb and will require some 500,000 SNP markers to facilitate whole genome studies (Kruglyak 1999). However, empirical data has subsequently shown that this is not entirely the case (Reich et al. 2001; Service et al. 2001).

In an early study Dawson et al. (2002) examined the measurement of LD along the entire human chromosome 22. 218 individuals from 3 populations (77 members from the Centre d'Etude du Polymorphisme Humain (CEPH), 90 unrelated individuals from the UK, and 51 unrelated individuals from Estonia) were genotyped with 1,504 SNP and insertion/deletion markers to estimate the extent of LD along this chromosome. Dawson et al. (2002) observed, in general, that LD decays with

increasing distance between markers but also showed considerable variability along the chromosome. Within the CEPH individuals there were expanses of LD (|D'| >.2), which extended over 400 kb as were there areas of no detectable LD (|D'| <.2) observed, between markers located <5 kb apart.

Another study into the extent and patterns of LD in the human genome involved the examination of 19 genes randomly spaced throughout the genome that exhibited large amounts of variation (Reich et al. 2001). Reich et al. (2001), reported that in comparing these 19 genes in a population of Utah Mormons, half-length of LD (or distance at which |D'| drops below .5) is ~60 kb. Despite this average measure of LD there was great variation in LD across the genomic regions. For example, |D'| remained >.5 for at least 155 kb around the *WASL* gene yet around the *PCI* gene LD remained >.5 for less than 6 kb. Since results presented in this study were obtained from within the same experimental population and from a large number of genes, it has shed some light on the conflicting results obtained when comparing, for example, results from a moderate number (one to three) of genes studied in different populations (Abecasis et al. 2001). Several other studies show similar levels of genomic LD as that observed in the Reich et al. (2001) study (Dunning et al. 2000; Ohashi et al. 2004; Service et al. 2001).

Despite the large amount of variation in genome-wide LD and the apparent randomness of this phenomenon in humans, there have been several reports describing distinct block-like patterns of LD throughout the genome. This block-like pattern was first observed when a 500kb region of the human genome, known to harbor a risk factor for Crohn disease, was analysed (Daly et al. 2001). SNP analysis of this region revealed that it could be divided into 11 discrete haplotype blocks (typically 10-100kb) with relatively little diversity but extensive LD. These regions

were shown to be interspersed with areas of high diversity which exhibited rapidly declining LD, possibly indicative of historical recombination events (Daly et al. 2001). Around this time, a survey of 216 kb in the major histocompatibility complex (MHC), found that it too consisted of blocks of sequence exhibiting high LD interrupted by regions with rapid breakdown in LD (Jeffreys et al. 2001). Subsequently there have been numerous studies reporting a similar trend of block-like LD structure (Gabriel et al. 2002; Johnson et al. 2001) leading to the conclusion that recombination is restricted to specific regions of the human genome known as recombination "hotspots" (Goldstein 2001). As discussed in section 1.5 this restriction of recombination events to particular regions of the genome will affect the patterns and distribution of LD throughout individual genomes.

1.8 Extent of linkage disequilibrium in plants

Plant genomes, like those of humans, have also revealed hotspots, or localization, of recombination (Dooner and Martinez-Ferez 1997; Faris et al. 2000; Tanksley et al. 1992). Although this is known, there is relatively little empirical data on how it relates to the patterns and levels of LD. There have been few studies in which LD has been examined in plants (Table 1.3), but apart from maize and, to a much lesser degree rice, very little research has been done on the evaluation of LD in crop species and in particular the economically important species of wheat.

In maize there have been a number of studies examining the extent and genomic patterns of LD on a local and a whole-genome scale. It has been well documented that recombination within the maize genome is localized to the genic regions resulting in extensive LD over long, intergenic regions followed by a rapid decline in LD within coding regions (Remington et al. 2001; Tenaillon et al. 2001; Thornsberry et al. 2001).

TABLE 1.3 Extent of linkage disequilibrium in plant studies.

Species	Extent of LD	Marker System	Gene/Genome-Wide Studies	Reference
Maize	600 kb	SNP	Gene based	(Palaisa et al. 2004)
	33.1 kb	SNP	Gene based	(Clark et al. 2004)
	2.0 Kb	SNP	Gene based	(Palaisa et al. 2003)
	1.5-7.0 Kb	SSR	Genome wide	(Remington et al. 2001)
Arabidopsis	250 Kb	SNP	Genome wide	(Nordborg et al. 2002)
Rice	100Kb	SSR/SNP	Gene based	Garris 2003
Loblolly pine	800 bp	SNP	Gene based	(Gonzalez-Martinez et al. 2006)
Sugarcane	10cM	RFLP	Genome wide	Jannoo 1999

In a survey of six genes Remington et al. (2001) reported on a general trend of LD decaying with distance. However, these diminishing levels of LD remained highly variable between genes probably due to a large effective population sizes during species evolution and high recombination rates within genes. In this particular study the LD diminished, on average, to an insignificant value within 1.5 Kb. Despite this rapid decline there were two genes in particular that exhibited far greater levels of LD ranging from 2.4 Kb to nearly 12 Kb. It is thought that selection of these genes during domestication is the cause of these elevated LD levels. In addition to these sequenced regions, Remington et al. (2001) studied a set of 47 SSR markers which provided stronger evidence for genome-wide LD than the SNP analysis.

Another experiment, this time involving 21 loci along chromosome 1 in maize, evaluated sequence diversity from an experimental population of diverse germplasm (Tenaillon et al. 2001). LD was shown to decay much more rapidly than the Remington study in which the intragenic LD decayed to a non-significant value in only 100-200 bp. This study was extended to include SSR markers and the extent of LD was calculated using 33 polymorphic SSR markers from the same region of

chromosome 1, presumably to determine whether local LD levels can be indicative of long range measures and vice versa (Tenaillon et al. 2002). The observed LD decay rate was similar. Of the 528 pair-wise comparisons between SSR markers, 33 were significantly associated which was reduced to 5 when the effect of multiple testing was accounted for. Furthermore, one of these significant associations was a consequence of two SSRs being in the same gene.

Based on the results of these studies, the levels of LD within the maize genome are not homogeneous and do not decay at a consistent rate across the genome. There are many possible explanations, including choice of experimental population (that is elite breeding lines versus more diverse germplasm), choice of genes, and genome location.

However, *Arabidopsis* shows a very different story and may prove to be a more accurate reflection of how LD will be structured in wheat. Since *Arabidopsis* is a self pollinating species elevated levels of LD throughout the genome are to be expected, as recombination is less effective at breaking associations between loci throughout the genome when they are largely homozygous, which is the case in self pollinating individuals (Borevitz and Nordborg 2003). There have been several studies in *Arabidopsis* examining nucleotide variation and LD around specific genetic loci (Hagenblad and Nordborg 2002; Jia et al. 2003; Nordborg et al. 2002; Shepard and Purugganan 2003). Each of these studies report similar findings with respect to extensive interlocus and intralocus LD. One study examines a 40 Kb region of chromosome 1 known to contain the *CLV2* gene responsible for shoot meristem development (Shepard and Purugganan 2003). Through SNP analysis, Shepard and Purugganan (2003) found that this 40 Kb region demonstrated significant levels of LD with complete LD extending in some cases over 25 Kb.

Since self pollinating organisms are expected to contain higher levels of LD than out crossing species, sequence diversity of such species will likely be studied on a much larger scale (Hagenblad and Nordborg 2002). Hagenblad and Nordborg (2002) studied 14 short fragments from a 400 Kb region around the flowering time locus (*FRI*), in 20 *Arabidopsis* accessions. Haplotype analysis revealed extensive haplotype structure in this region lending to high levels of LD, up to 250 kb.

Nordborg et al. (2002), also sequenced short segments from around the *FRI* gene and confirmed the rapid decline in LD over 250 kb (~1cM). However, they expanded this study to determine if local levels of LD were indicative of genome-wide levels. Nordborg et al. (2002) evaluated 163 genome-wide SNP markers in their examination of 76 global *Arabidopsis* lines. They observed a weak relationship between LD and distance. One proposed explanation was that SNP markers were adequately spaced such that recombination effectively removed any LD. Furthermore, only a few of the genome wide SNP markers were spaced as far as the most distal SNP markers from the *FRI* region. Interestingly, when local *Arabidopsis* populations were examined using markers around the *RPM1* locus, a gene thought to confer disease resistance, genome-wide LD was very extensive, stretching over 50-100 cM (Nordborg et al. 2002). Local populations are expected to reduce heterzygosity within a genome and thus increase the levels of LD. As such, the authors of this study suggest that it may be possible to 'zoom in' on particular loci by using the appropriate populations.

The varying levels of LD observed in these *Arabidopsis* studies reaffirm the necessity to examine this genetic phenomenon in a number of populations for the organism of interest since the population used will likely vary the levels observed. It is clear from both the maize and *Arabidopsis* studies that local, elite populations will

exhibit elevated levels of LD throughout the genome, making them more amenable to whole genome scans with relatively fewer markers. On the other hand greater diversity amongst individuals and sampling from a global population will result in reduced levels of LD across the genome and will be more conducive to a candidate gene mapping approach requiring a higher marker density to detect the associations. Through the identification of these different populations and understanding the distribution of LD throughout the genome it may be possible to modulate LD levels and facilitate a two tiered approach to association mapping.

1.9 Measuring linkage disequilibrium

Although the premise of LD analysis is fairly straightforward the same cannot be said for the statistics used in measuring associations. To help understand these statistical methods, assume that there are two segregating loci A and B for which there are four alleles A, a, B, and b. The four possible gametic types are therefore: AB, Ab, aB, and ab for which the frequencies are f_{AB}, f_{Ab}, etc. The two most common measures of LD are D' and r^2 each of which are derivatives of the D statistic. The D statistic is used to measure the difference between observed and expected allele frequencies at two segregating loci (Lewontin 1960), and is defined as;

$$D_{ab} = (f_{AB}-f_A f_B),$$

where f_A, f_a, f_B, and f_b are the allele frequencies at the two loci and f_{AB} is the frequency with which the AB haplotype occurs in the population. The expected value of f_{AB} in the linkage equilibrium state would be equal to the product of the individual allele frequencies, thus equalling zero (Jorde 2000). Where the AB alleles are observed to co-occur more frequently in a population compared to the product of their individual frequencies, D will be greater than zero and hence there is evidence of LD. This

disequilibrium will gradually be eroded through the acts of recombination but is dependant on the frequency and hence number of generations as well as other evolutionary factors such as selection (Lewontin 1960).

One derivative of this LD statistic is r^2. This statistic is defined as the squared correlation coefficient between two marker loci (Lewontin 1988) and is defined as;

$$r = D/ \sqrt{f_A f_B f_a f_b},$$

which can also be written as;

$$r^2 = (f_{AB}-f_A f_B)^2/ (f_A f_B f_a f_b)$$ (Pritchard and Przeworski 2001)

The r^2 values will range between 0 and 1 indicating complete linkage equilibrium and complete linkage disequilibrium, respectively. However, it is strongly affected by allele frequencies at the loci being examined (Jorde 2000).

D' is another derivative statistic from the first equation above and ranges between -1 and +1, representing equilibrium and disequilibrium, respectively. D' is defined as (Lewontin 1988);

$$D' = D / D_{max,}$$

where;

$$D_{max} = \min (p_1 q_2, q_1 p_2) \text{ when } D>0,$$

when;

$$D_{min} = \min (p_1p_2, q_1q_2) \text{ when } D<0.$$

This D' statistic is indicative of recombination and is also dependant on allele frequency (Delvin and Risch 1995).

In a study by Delvin and Risch (1995), in which they evaluated the properties of 5 disequilibrium measures for gene mapping (which included D' and r^2) the D' statistic was superior to r^2 in terms of fine scale mapping and although each measurement had some sensitivity to allele frequencies, the r^2 statistic was decidedly more so. This led Delvin and Risch (1995) to the conclusion that r^2 would be most useful in simple LD mapping where the allele frequencies didn't vary greatly from locus to locus but that the best estimate of LD would be that obtained by D' estimates.

The presence of multiple alleles at two loci, as is the case with SSR markers, adds further complexity to the LD calculations. In the software package *Powermarker*, used predominantly in this study of LD in hexaploid wheat, the LD calculation for multiple alleles is as follows (Liu and Muse 2004);

$$LD = \sum_u \sum_v p_u p_v |LD_{uv}$$

where LD can be either r^2 or D' and is the weighted mean of the absolute allele pair LD.

There are several statistical tests that can be used in determining marker correlations in studies of LD. The two most commonly used measures are described above and each are affected by factors such as recombination, and population history in very different ways. Ultimately, it is up to the researcher to decide which method is

most appropriate depending on the questions asked and the nature of the data being studied, particularly with respect to the allele frequencies in the data set and marker type for example.

1.10 Factors affecting linkage disequilibrium

It is clear from the number of studies discussed so far that there exist very distinctive patterns of LD both locally and on a genome-wide scale. LD has been shown to extend from a few base pairs (Tenaillon et al. 2001), to tens of centimorgans (Nordborg et al. 2002). Variations in the extent and distribution of LD within genomes are dependant on a number of factors including both physical processes and effects from the population dynamics of the experimental population.

1.11 Physical factors affecting LD

1.11.1 Mutations and recombinations

Perhaps the most obvious generator of LD is a new mutation. The emergence of a new mutation in a particular generation leaves all remaining alleles in that region, in that individual, completely associated with it. From generation to generation this extensive LD is gradually broken down by recombination until the allele frequencies are ultimately returned to equilibrium within the population.

Recombination is the most important physical factor that works to vary genomic levels of LD within an organism by breaking down associations. With meiotic crossovers essentially being restricted to “hot spots” within genomes there is a general pattern consisting of long stretches of conserved sequence with few haplotypes but extensive LD (Daly et al. 2001; Gabriel et al. 2002; Johnson et al. 2001). Within plant systems the location of these recombination events are

heterogeneous with repressed crossing over near the centromere. The probability of recombination increases while moving distally from the centromere towards the telomeres (Tenaillon et al. 2001; Weil 2002). In maize it is well understood that recombination occurs more frequently in genes than in intergenic regions. The *bronze (bz)* locus, for example, a gene that conditions anthocyanin pigmentation in the aleurone layer of seeds, has been shown to be a recombination hotspot demonstrating more than 100 fold higher recombination over the genome average (Dooner and Martinez-Ferez 1997). The cross over events across the *bz* locus occur uniformly. However, the presence of insertions or deletions affect these frequencies through reduction in cross-over rates (Dooner and Martinez-Ferez 1997). In studying the flanking sequence of the *bz* locus it was noted that the distal gene rich side of the locus had a much higher frequency of recombination relative to the proximal side of the locus which exhibited reduced recombination due to the presence of retrotransposons. This further confirms the hypothesis that recombination in maize is largely restricted to genes (Fu and Dooner 2002).

The same is true for wheat, in that recombination is not uniform throughout the genome (Akhunov et al. 2003). Several studies have shown that there is a tendency for meiotic crossing over to occur at an increased frequency distal to the centromere (Dvorak et al. 1998; Gill et al. 1996a; Gill et al. 1996b). These studies and those discussed above lead to the question: what impact will recombination have on the expected levels of LD in wheat? In maize it is clear that in gene rich regions, such as those distal to the centromere, there will be low levels of LD as compared to the regions devoid of genes that will demonstrate more extensive LD. Interestingly in wheat, unlike most other plants, the recombination frequency is dependant more on the location along the centromere/telomere axis as opposed to the local gene density

(Akhunov et al. 2003). As a result it is expected that LD in wheat will be high around the centromeric regions and gradually decay as one moves distally.

1.12 Population genetic factors affecting LD

Physical factors within the genome such as mutations and recombination help to create and break down levels of LD, respectively. There are additional factors resulting from population genetics that also contribute to varying levels of LD in the genome. Factors such as selection, population structure, mating systems and gene conversion need to be considered when studying LD in an organism.

1.12.1 Selection

Selection at a locus, whether natural or artificial will affect the pattern of LD in a profound way. Selection for an allele or region of a genome will reduce genetic diversity around a locus resulting in the levels of LD to increase, at least temporarily, in the surrounding area.

The effects of selection have been studied extensively in maize with respect to domestication. In a recent study, Palaisa et al. (2004) examined how far reaching the effects of selection were from *Y1*, a maize gene responsible for endosperm color. In addition to genotype-phenotype associations and genetic diversity, they also examined LD at increasing distance from the *Y1* gene, 1.2 Mb in the 5' direction and ~700 kb in the 3' direction. LD was calculated between pairs of informative SNPs as |D'| and r^2. There was significant LD between SNPs in genes located 1kb and 600 kb downstream from the *Y1* gene, with the thresholds for LD being 0.1 for r^2 and 0.5 for D'. Differences in yellow and white diversity ratios were calculated and were supportive of a selective sweep, or reduction of sequence diversity in the region surrounding an allele that has been the target of a recent selection, in this case caused

by the selection at *Y1* (Palaisa et al. 2004). Upstream the diversity ratio was higher than that observed downstream, where the diversity ratio was very low (0.1) and extended 600 kb. LD analysis of this region confirmed the patterns of sequence diversity in this region. Upstream, the 5' region of *Y1* had significant LD scores of |D'| 0.7 and r^2 0.09 with a region located 200 kb away, reducing to |D'| 0.51 and r^2 0.05 for a region that is 550 kb away. Downstream however, there is significant LD with the 5' region of *Y1* and all tested regions (|D'| 0.5 to 0.8 at distances of 13 kb and 575 kb respectively, with a region 700 kb away demonstrating |D'| 0.7; and r^2 0.1 from 1 kb to 700 kb), with the exception of one, *sbe1*. Palaisa et al. (2004), demonstrate through diversity analysis that there has been a recent selective sweep around the *Y1* gene. They confirm this finding through LD analysis in the region which revealed perhaps the longest reported stretch of LD in maize to date (Clark et al. 2004; Palaisa et al. 2003; Remington et al. 2001).

In an earlier study, Remington et al. (2001) sought to evaluate the patterns and extent of LD in 102 diverse maize inbred lines, through the evaluation of six candidate genes for important agronomic traits. Five of the six candidate genes studied showed r^2 values decreasing to less than 0.1 in ~2 kb. Only the *su1* candidate gene locus exhibited a greater range of LD extending ~7 kb. In a subsequent study, Whitt et al. (2002) examined selection at six loci through sequence diversity analysis in maize and *Z. mays* ssp. *parvigmulis*, one of these six loci being *su1*. The *su1* gene exhibited average diversity levels in *Z. mays* ssp. *parvigmulis*, but low levels of diversity in maize which, the authors point out is consistent with selection during domestication or breeding. The reduced sequence variation observed at this locus due to selection, is a probable explaination for the greater than ususal LD observed at the same locus (*su1*) in the Remington et al. (2001) study.

1.12.2 Population dynamics

Since patterns of LD are indicative of ancient recombination events, populations differing in size, origin, and growth will likely display different patterns of LD. As noted previously there has been a reduction in the levels of sequence diversity in plant species, particularly maize during domestication (Clark et al. 2004; Palaisa et al. 2004; Remington et al. 2001). This reduced sequence diversity is due in part to the small initial populations in crops relative to the population size of wild relatives and also in part to selection for agronomically important traits (Clark et al. 2004; Zeder et al. 2006). The former effect is more commonly referred to as a population bottleneck and is a result of only a subset of the genetic variation found in the wild progenitors being retained in the new founder population (Zeder et al. 2006). The extent of this reduced variation is dependant on the duration of the bottleneck and the number of individuals in the founding population (Zeder et al. 2006).

A comparative study of genetic diversity in the cultivated rice specie *Oryza sativa* and the wild progenitors (*Oryza rufipogon* and *Oryza nivara*) revealed greater nucleotide variation in the wild progenitors and a reduction of such in the cultivated ssp. *indica* and *japonica* (Zhu et al. 2007). In fact there is retention of 20 and 10% of the diversity present in the wild relatives in *O. sativa* ssp. *indica* and *O. sativa* ssp. *japonica*, respectively. Simulations of bottlenecks with respect to duration and population size were carried out to better understand the process of domestication. Results from these simulations detected a bottleneck and sequence diversity currently present in the rice genome is likely explained by a founding population of 1,500 individuals. LD was calculated as r^2 and the lowest levels were present in *O. rufipogon* where r^2 dropped below 0.1 within 400bp and the highest was present in

the japonica where LD extended (r^2 =0.7) to the entire sequenced region that was ~900 bp. (Zhu et al. 2007) conclude that this small effective population size is most likely the cause of higher LD in the cultivated rice species since only 10-20% of diversity present in the wild relatives is present in these species.

Another population dynamic that affects LD, particularly in association studies is population admixture. Perhaps the most widely referenced study demonstrating the implications of population admixture on association studies is one in which type 2 diabetes was examined in two Native American tribes in southern Arizona (Knowler et al. 1988). This study showed a correlation between one particular haplotype at the human leukocyte antigen (*HLA*) locus and the reduced incidents of diabetes. However, upon further analysis it became clear that this was a spurious association as a result of population admixture. In this case the haplotype associated with reduced diabetes was more common in Europeans that Native Americans, and the diabetics in this study indeed had a lower proportion of European ancestry relative to the controls. This was elucidated only when analysis was repeated within the sub-population of individuals with reduced European background and the resulting association disappeared.

An early study in maize used association mapping to evaluate *Dwarf8* sequence variants in 92 inbred lines (Thornsberry et al. 2001). When accounting for population structure, the type 1 error was reduced by up to 4.7 fold. By incorporating this statistical correction to determine sub-structure, Thornsberry et al. (2001) were successful in identifying a number of *Dwarf8* polymorphisms that were associated with differences in flowering time. The authors also conclude that obstacles encountered as a result of this population stratification in crop plants can be overcome

by sampling unlinked markers and statistically accounting for population structure, which can improve the identification of alleles contributing to quantitative traits.

Being aware of the fact that population structure exists is an important point to consider at the outset of any LD study. There are numerous statistics available but perhaps the most popular method of accounting for population structure in experimental populations lies in a clustering software program called *Structure* (Falush et al. 2003; Pritchard et al. 2000). This software infers the population sub-structure based on multi locus genotype data. To test the effectiveness of this software Rosenberg et al. (2001) conducted a study in which 600 individual chickens were genotyped with 27 SSR markers which was then used to infer the population structure. The chickens represented 20 discrete breeds for which population structure was already known. The clustering algorithm was successful in correctly clustering individuals belonging to the same sub-population ~98% of the time. In addition, they examined the clustering success rate as a function of marker number and numbers of individuals used. They found that, for similar studies the optimal number of highly divergent markers would be 12-15 when studied in a sample size of 15-20 individuals per hypothesised cluster, to obtain a success rate of 90% or higher.

This clustering method has been successfully applied to a number of populations to account for population structure prior to commencing association analysis (Liu et al. 2003; Rosenberg et al. 2002; Thornsberry et al. 2001). The use of such statistical methods will likely improve the reliability and confidence of results generated in association studies. This is important, as LD generated through population structure is likely to be prevalent in plant systems due to their breeding histories.

1.12.3 Mating systems

The breeding system of organisms will also have a positive impact on LD. As discussed earlier, recombination works to reduce LD in a genome by breaking associations. However, if a large proportion of the genome is homozygous, as is the case with self-pollinating species, the effective recombination is greatly reduced. The effect this has on LD is a positive one, causing it to extend over greater distances.

Arabidopsis, like wheat, is largely self-pollinating (99%) (Nordborg et al. 2002), and has been shown to exhibit 250-fold higher LD to that observed in maize, which is a predominantly an out-crossing species. The differences observed between mating systems will generate different approaches to association mapping. Where LD is extensive, as in predominantly inbreeding species, the marker density required to identify a significant association will be several times less (for example 1 SNP every 50 Kb in *Arabidopsis* (Nordborg et al. 2002)) than that required for out-crossing species such as maize (for example 1 SNP every 100-200 bp as discussed by Tenaillon et al. (2001)). These and other implications of variations in LD are illustrated in Figure 1.2.

There are several factors affecting the levels and patterns of LD throughout the genomes of various species. In addition to those discussed above, other aspects such as genomic location (Palaisa et al. 2003), choice of experimental population (Eaves et al. 2000; Nordborg et al. 2002; Remington et al. 2001; Tenaillon et al. 2002; Tenesa et al. 2003) and population bottlenecks and admixture (Jannoo et al. 1999; Knowler et al. 1988; Lonjou et al. 2003) will all contribute to the diversity of this genetic phenomenon.

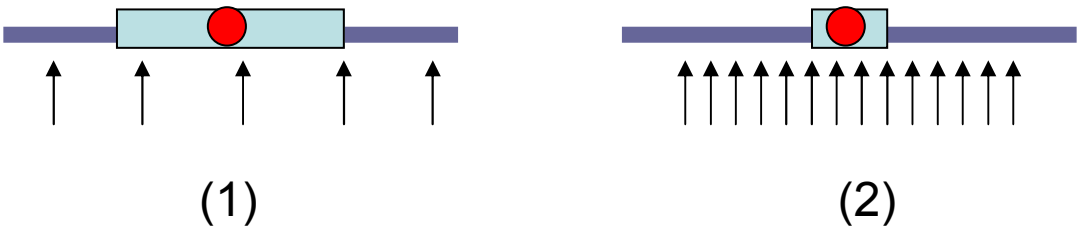


FIGURE 1.2 Schematic representations of the effects of LD on association mapping strategies from (Rafalski 2002). In diagram (1) LD is persisting over a greater distance (light green rectangle) around the gene affecting the trait of interest (red circle). The consequence of this is a lower map resolution with respect to the gene; however fewer markers (arrows) are needed to identify, by association, the gene of interest making this scenario amenable to whole genome scans. On the other hand, diagram (2) represents a scenario in which LD declines rapidly around the gene of interest and requires a very dense marker map to localize the region. In this situation, however the gene may be mapped with greater precision, providing the marker density is high enough, and is more amenable to a candidate gene approach to association mapping.

1.13 Summary

Association mapping is proving to be a useful tool in human genetics and several studies in maize provide optimism for other crop species. However, it is important as a first step, to identify local and genome-wide levels of LD in a number of different populations. In doing so, populations could be identified that harbour extensive LD as well as those with far less LD. The former scenario would be ideal for whole genome scans, requiring few polymorphic markers and identifying broad regions of the genome containing genes of interest. The latter scenario would be more amenable to a candidate gene approach requiring a much higher marker density but would result in a much higher map resolution of the gene. There are additional factors that complicate LD analysis, in particular those discussed in section 1.12. However, by understanding the organism and specifically the population under investigation, it is possible to circumvent these issues, particularly spurious associations or false positives.

With the increase in available marker and sequence data from complex higher plants the association mapping system that has proven so successful in humans is likely to be realized in economically important crop species such as wheat.

Chapter 2: Sequence based analysis of linkage disequilibrium in wheat

2.1 Introduction

Precise positioning of genes contributing to agronomically important traits in wheat (Kuchel et al. 2006; Williams et al. 2006), through QTL analysis, is improving with the availability of dense genetic linkage maps (Somers et al. 2004). Relying mainly on inbred populations genotyped with numerous genomic markers, this traditional method of trait localisation identifies broad regions with the gene controlling the trait located within regions spanning several centimorgans in most cases (Cuthbert et al. 2006; Roder et al. 2007). Correlations between genotype and phenotype in a group of individuals can also be detected through association mapping based on the linkage disequilibrium present in the genome (Aranzana et al. 2005; Yu and Buckler 2006). There are two main advantages to association mapping as compared to traditional linkage mapping methods; firstly, it allows greater precision in QTL location over family-based linkage anaylsis and secondly, association mapping can be applied to a range of natural populations and does not require family or pedigree information as is the case in family-based linkage methods (Mackay and Powell 2007). The success of association mapping strategies, for fine mapping genes controlling traits of interest, is dependant on the persistence of LD within the genome being examined as this will determine the number and density of markers required for this type of analysis (Flint-Garcia et al. 2003). As such, understanding the levels and patterns of LD throughout the genome will be important. In populations where LD persists over large distances, scanning the the whole genome with molecular markers to identify marker/trait associations will be feasible with the added bonus that the number of genetic markers required to do will be a lot less, compared to those populations in which LD persists over relatively short distances and require a higher marker density to detect marker/trait associations (Rafalski 2002).

While LD in plant species has not been thoroughly examined, it is a phenomenon that can now be intensively investigated with the complete genomic sequences for *Arabidopsis thaliana* (The Arabidopsis Initiative, 2000) and rice (Yu et al. 2002) available. The levels and distribution of genomic LD vary dramatically between species and particularly between populations, such as locally adapted and global populations. In a recent study of global and local *Arabidopsis* lines, LD was estimated to extend up to 250 Kb and beyond 100cM, respectively (Nordborg et al. 2002). In maize, it is estimated that LD is very high in intergenic regions of the genome and rapidly breaks down within genes over approximately 1500 bp (Remington et al. 2001). Wheat, as well as being polyploid, harbours a large number of repetitive elements throughout the genome making sequence based LD analysis challenging in this economically important species.

This experiment explores for the first time the genetic concept of LD at a sequence level in a collection of *T. aestivum* lines, by examining two single copy genes on chromosome 5D known to be the main determinants of grain texture. Grain texture is perhaps the most important quality trait in wheat as is ultimately determines the end us of the grain (Morris 2002). There are three genes at the "Hardness" (*Ha*) locus on the short arm of chromosome 5D that were found to be tightly linked; *puroindoline a* (*Pina-D1*), *puroindoline b* (*Pinb-D1*), and *Grain Softness Protein* (*Gsp-1*) (Chantret et al. 2004). A study by Gautier et al. (2000) revealed that the Pina and Pinb genes were present in the diploid ancestor wheats as well as the D genome of hexaploid wheat but were absent in the tetraploid (AABB), *T. turgidum*. Gsp-1 is the only gene present in the *Ha* locus that is also present in the tetraploid T. turgidum and the A and B genomes of hexaploid wheat (Chantret et al. 2005). It is still not clear why the puroindoline genes are absent from the A and B genomes of allopolyploid wheats. In examining the evolution of this region in diploid and

polyploidy wheat species Chantret et al. (2005) conclude that the loss of the Pina and Pinb genes from the A and B genomes of polypolid wheat species is due to one or more large genomic deletions containing these and at least 4 other genes at the time of polyploidisation. Despite the locus controlling "hardness" it the soft grain phenotype that is dominant. When both genes, Pina and Pinb are present and in their native forms the resulting grain texture will be soft whereas the absence of or mutation in one of these genes will result in a hard endosperm texture (Giroux and Morris 1997). There are 7 allelic combinations that will result in the hard endosperm phenotype (*Pina-D1b*, and *Pinb-D1b* through *Pinb-D1g*) (Giroux and Morris 1997; Lillemo et al. 2002; Morris et al. 2001). The Pina gene appears to have two forms the native or "wild-type" designated *Pina-D1a* and an apparent null mutation designated *Pina-D1b* (Giroux and Morris 1998). When Giroux and Morris (1998) examined sequence variations at the Pina and Pinb genes they found that wheats with the hard endosperm characteristic either had an amino acid change within the Pinb gene or had the Pina null mutation.

In this chapter the single copy *Pina-D1* and *Pinb-D1* genes were amplified in 69 and 85 hexaploid wheat lines, respectively. Of those *Pina-D1* and *Pinb-D1* amplified products 44 lines and 62 lines, respectively, provided good reliable sequence results, which allowed for an initial examination of the LD in this region at the sequence level. As the results will show there was a certain lack of diversity in these genes in the hexaploid wheat samples and as such an examination of the Pina and Pinb gene sequences, as well as Gsp-1 gene sequences, were conducted in the more diverse D genome progenitor, *Aegilops tauschii*. As was observed through the examination of 50 *Aegilops tauschii* sequences, this highly diverse set of germplasm had LD present beyond 1300 bp in this region of chromosome 5D.

2.2 Materials and methods

2.2.1 Plant material

Ninety-six Australian hexaploid wheat lines were selected as a representation of the current germplasm in Australian wheat breeding programs, as well as historical varieties that contribute to their individual pedigree (Appendix A). Seed for 52 of the 96 lines used in this study were sourced locally and kindly provided by Drs S Jefferies, Australian Grain Technologies and G Holamby, Roseworthy Agricultural College. The remaining 44 lines were sourced through the Australian Winter Cereals Collection in Tamworth, NSW.

Seed from these 96 wheat lines were sown in temperature-regulated greenhouses. At the 4 to 5 leaf stage 2 grams of leaf tissue was collected from a single plant for each variety, snap frozen in liquid nitrogen and stored at -80°C in preparation for DNA extraction.

A subset of 87 lines from the 96 Australian hexaploid wheat varieties mentioned above, were selected for use in this chapter. Reliable sequence data was obtained from *Pina-D1* and *Pinb-D1* genes in 44 and 62 lines, respectively and are outlined in Table 2.1. The low number of successful sequences obtained was attributed to failed sequencing reactions and not PCR amplification of the genes. The Pina gene, as discussed earlier is known to have 2 alleles *Pina-D1a* wild type and *Pina-D1b* null mutation and as such not all wheat lines amplified a product for this gene. Seven of the 87 lines were included in duplicate to ensure sequence integrity in the *T. aestivum* samples.

TABLE 2.1 List of the Australian hexaploid wheat accessions used in evaluating sequence level LD in the Pina and Pinb genes. Source, origin and year of release are included for a majority of the lines as well as the puroindoline gene from each line that was sequenced and analysed for LD in this study.

Germplasm	Source	Origin	Year	Pina Sequence	Pinb Sequence
Andes	Tamworth	MEX	1969	X	
Anza	Tamworth	MEX,USA:Calf	1971	X	
Aus10894	Tamworth	AUS	1975	X	
Banks	Waite	AUS:NSW	1979	X	X
Beulah	Tamworth	AUS:Vic	1993	X	X
Bindawarra	Waite	AUS:SA	1980	X	
Bluebird	Tamworth	MEX	1969		X
Camm	Tamworth			X	X
Cascades	Waite	AUS:WA	1994	X	X
CD87	Waite			X	
Ciano67	Tamworth	MEX	1967		X
Condor	Waite	AUS:NSW	1973		X
Cook	Waite	AUS:Qld	1977	X	X
Cranbrook	Waite	AUS:WA	1985	X	
Chinese Spring	Waite	CHN:Sichuan		X	X
Cunningham	Waite	AUS:Qld	1990	X	
Dagger	Waite	AUS:SA	1983	X	X
Diamondbird	Roseworthy	AUS:NSW	1997		X
Dirk48	Tamworth	AUS:SA	1951		X
Excalibur	Waite	AUS:SA	1990		X
Federation	Waite	AUS:NSW	1901	X	
Festiguay	Waite	AUS:NSW	1963	X	
Frontana	Tamworth	BRA:Rio	1930	X	X
Fultz	Tamworth	USA:Penn	1871	X	
Gamenya	Waite	AUS:NSW	1958		X
Ghurka	Waite	AUS:Vic	1924	X	X
H45	Tamworth			X	X
Halberd	Waite	AUS:SA	1969		X
Hartog	Waite	AUS:Qld	1982		X
Heron	Waite	AUS:NSW	1958		X
Insignia	Waite	AUS:Vic	1946	X	X
Janz	Waite	AUS:Qld	1989		X
Kalyansona	Tamworth	IND	1967		X
Katepwa	Waite	CAN:MB	1981	X	
KenyaSupremo	Tamworth			X	X
Kloka	Tamworth	DEU	1965	X	X
Krichauff	Roseworthy	AUS:SA	1998	X	X
Kukrirac820	Waite				X
LermaRojo	Tamworth	MEX	1955		X
Machete	Waite	AUS:SA	1985		X
Meering	Waite	AUS:Vic	1984	X	X
Norin10Brevor	Waite	USA:Wash		X	X
OpataM85	Waite	MEX	1985		X

Table 2.1 Continued

Germplasm	Source	Origin	Year	Pina Sequence	Pinb Sequence
Orfed	Tamworth	USA:Wash	1943		X
Oxley	Waite	AUS:Qld	1974	X	X
PavonS	Tamworth	MEX	1977		X
Pitic62	Waite	MEX	1962	X	X
Rac177	Waite	AUS:SA	1977		X
Ranee	Waite	AUS:Vic	1924	X	
Raven	Waite	AUS:NSW	1963		X
Rosella	Waite	AUS:NSW	1985	X	X
Schombergk	Roseworthy	AUS:SA	1986		X
Scimitar	Tamworth	AUS:SA	1930		X
Silverstar	Waite	AUS:Vic	1998	X	X
Spear	Waite	AUS:SA	1983	X	X
Stiletto	Waite	AUS:SA	1993		X
Sunco	Waite	AUS:NSW	1986	X	X
Sunstate	Tamworth	AUS:NSW	1993		X
Sunvale	Tamworth	AUS:NSW	1994	X	X
Supremo	Tamworth	USA:Tx	1948		X
Synthetic	Waite	BGR		X	
Tatiara	Waite	AUS:SA	1988		X
Timgalen	Tamworth	AUS:NSW	1967		X
Tincurrin	Waite	AUS:WA	1977	X	
Trident	Waite	AUS:SA	1993	X	X
Trintecino	Tamworth	BRA	1936	X	X
UNICULM492	Tamworth	ISR			X
Veranapolis	Waite			X	X
VPM1	Tamworth	FRA	1981		X
Warigal	Waite	AUS:SA	1978		X
WW15	Waite	AUS:NSW	1969	X	X

2.2.2 Nucleic acid isolation

Collected leaf tissue was ground into a fine powder using a mortar and pestle with liquid nitrogen. After allowing the tissue to thaw slightly, 4.5 mL of DNA extraction buffer (1% sarkosyl, 100 mM Tris-HCl, 100 mM NaCl, 10 mM ETDA and 2% PVPP) was added. This mixture was then poured into a 10 mL plastic tube to which 4.5 mL of phenol/chloroform:IAA (25:24:1) was added and left to mix by inversion on an orbital rotor for 15 mins.

The samples were centrifuged at 5000 rpm for 10 mins after which the supernatant was poured into a silica matrix tube, with a further 4.5 mL of phenol/chloroform:IAA added. The samples were centrifuged for a further 10 mins at 5000 rpm and the supernatant transferred to a fresh 10 mL plastic tube where 400 µL of 3M sodium acetate (pH 4.8) and 4 mL of isopropanol were added. The DNA was precipitated by gently inverting the tubes several times. After spooling the DNA from the tube with a glass pasteur pipette, it was carefully transferred to a 2 mL microfuge tube containing 1 mL of 70% ethanol. The DNA was centrifuged for 2 mins at 5000 rpm after which the ethanol was decanted and any excess removed with a cotton bud. The DNA pellets were allowed to air dry for approximately 30 mins and were then re-suspended in 350 µL of R40 (40 mg mL^{-1}) at room temperature for 2 hours on an orbital rotor.

2.2.3 PCR amplification of Pina and Pinb genes

The Pina and Pinb genes were amplified from genomic DNA isolated from the lines described above using the DNA extraction method described in section 2.2.2. Primer sequences specific to the *Pina-D1* and *Pinb-D1* genes were obtained from Gautier et al. (1994) and are presented in Table 2.2.

TABLE 2.2 Gene specific primers used to amplify the full length Pina and Pinb genes in the hexaploid wheat varieties outlined in Table 2.1. These primer sequences were obtained by (Gautier et al. 1994) and the PCR amplicon sizes for each gene are also listed here.

Puroindoline Gene	Forward Primer Sequence (5' to 3')	Reverse Primer Sequence (5' to 3')	PCR Amplicon Size (bp)
Pina	5'-ATGAAGGCCCTCTTCCTCA-3'	5'-TCACCAGTAATAGCCAATAGTG-3'	447
Pinb	5'-ATGAAGACCTTATTCCTCCTA-3'	5'-TCACCAGTAATAGCCACTAGGGAA-3'	447

Each PCR was conducted in 96-well microtitre plates (ABgene, UK) with a final reaction volume of 50 µL. Each reaction contained 1x PCR Buffer, 1.5 mM $MgCl_2$, 0.2 mM each dNTP (0.8 mM), 1 µM of the forward and reverse primer, 1U Taq DNA polymerase (Invitrogen, USA) and 50 ng total genomic DNA. Since the Pina gene has two alleles, wild type and a null mutation, PCRs were initially performed with an internal control (glutenin) to distinguish true null alleles from failed amplifications.

The PCR program consisted of an initial denaturation step of 95°C for 7 mins followed by 35 cycles of 94°C for 10 seconds, 58°C for 30 seconds, and 68°C for 45 seconds. This was followed by a final extension of 72°C for 10 mins.

Each PCR product was run on a 1% TAE agarose gel for visualisation and purification. Each product was excised from the gel using a scalpel and long wave UV transilluminator and placed into fresh 1.5 mL Eppendorf tubes for subsequent gel purification.

2.2.4 Fragment purification and sequencing of Pina and Pinb genes

Each PCR product isolated in agarose gel fragments was purified using the Qiagen gel purification kits as per the manufacturers' instructions. Once purified, 2 µL of the PCR product was loaded on a 1% agarose gel for quantification.

Sequencing PCRs were set up with 50 ng of purified PCR product, 3.2 pmol primer, 1 µL of Big Dye Terminator (Version 3.1) (Perkin Elmer, Applied Biosystems Division), 3 µL of the Big Dye dilution buffer, and water to a final volume of 12 µL. The PCR program consisted of an initial denaturation step of 96°C for 30 seconds followed by 25 cycles of 96°C for 10 seconds, 50°C for 5 seconds, and

an extension of 60°C for 4 mins. Upon completion, samples were allowed to equilibrate to room temperature following which each sample was cleaned as follows. Seventy-five microlitres of 0.2 mM $MgSO_4$ was added to each sample, the plate was sealed and vortexed to ensure thorough mixing. The samples were then allowed to precipitate at room temperature for 15 mins followed by centrifugation at 4000 rpm also for 15 mins.

The plate was then removed from the centrifuge and the supernatant removed by gently inverting the plate over paper towels. To ensure that the supernatant was completely removed the plate was centrifuged for an additional minute upside down at 1000 rpm followed by an additional 15 mins where the samples were allowed to air dry. The plate was then submitted to the Australian Genome Research Facility (AGRF) for sequencing.

2.2.5 Data analysis

Sequences from the two puroindoline genes, Pina and Pinb, in the *T. aestivum* samples were edited and aligned using Vector NTI suite 7 (InforMax Inc., USA). The *A. tauschii* gene sequences for Pina, Pinb and GSP were obtained from the publicly available GenBank database at the National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/, 2005) and are denoted by the accession numbers AY251946-AY252095. All SNP and LD analysis was carried out using the computer software DNasp (Version 3.51) (Rozas and Rozas 1999). In each case, *T. aestivum* and *A. tauschii*, LD was estimated for each gene independently as well as for the concatenated sequences of each, Pina and Pinb and Pina, Pinb and GSP-1, respectively. The LD between polymorphic sites and the significance of the data was calculated using the r^2 estimate and the Fisher's exact test and Bonferroni corrections (a statistical correction aiming to avoid the spurious rejection of the null hypothesis in

multiple tests, assuming all tests are independant) as performed in the software DNasp (Version 3.51) (Rozas and Rozas 1999).

2.3 Results

2.3.1 Sequence evaluation of Pina and Pinb genes from hexaploid wheat

Gene specific PCR primers successfully amplified the two single locus puroindoline genes from two of the three homoeologous group 5 nullisomic-tetrasomic (NT) lines. Figure 2.1 is a gel image showing the absence of *Pina-D1* and *Pinb-D1* PCR products in the wheat line, which lacks chromosome 5D, and the presence of products in the other two lines, one that lacks chromosome 5A only and the other, which lacks only the 5B chromosome. These results indicate that the gene specific primers described in Table 2.2 are in fact targeting the 447 bp gene on the short arm of chromosome 5D in the hexaploid wheat lines used in this study.

The Pina gene, having a null allele, was amplified from the wheat genomic DNA together with a glutenin gene as an internal control, to determine whether the lack of PCR product was a legitimate null allele or merely a failed PCR reaction. Figure 2.2 is the gel image of the aforementioned PCR experiment and shows the 19 samples that are true null alleles.

With respect to SNP analysis only informative polymorphic loci, or those with a base variant in more than 1 variety, were included in the calculations of LD in the Dnasp software (Rozas and Rozas 1999). Base changes in one line offer little information and are most likely attributed to errors in sequencing. Good reliable Pina sequence was obtained from 44 hexaploid wheat lines (Table 2.1), among which there were no

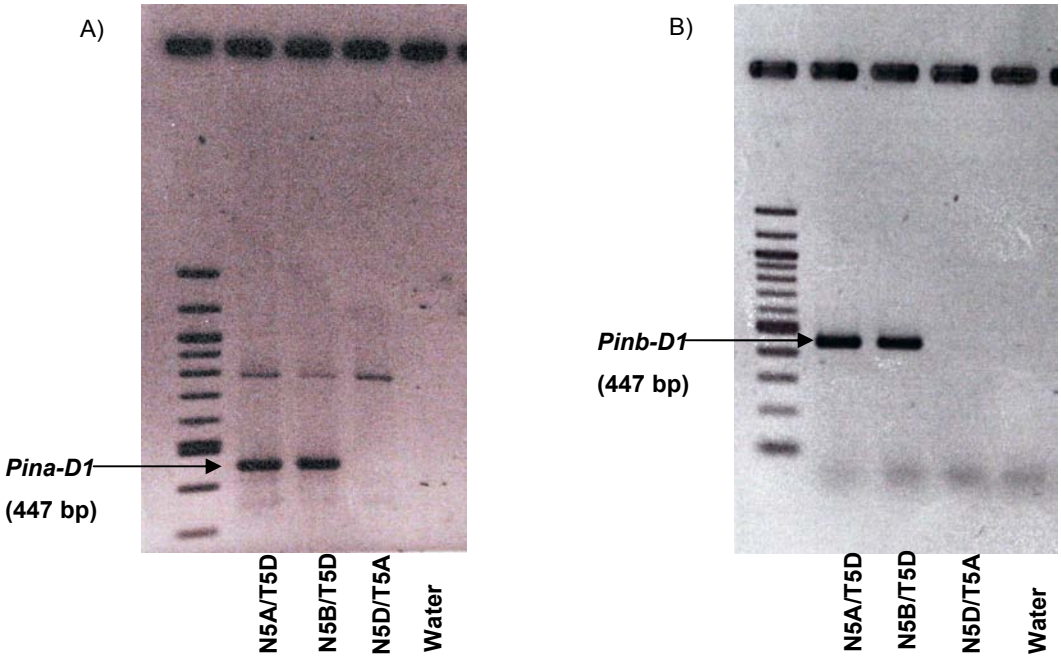


FIGURE 2.1 Gel images of three group 5 nullisomic tetrasomic (Nulli-Tetra) Chinese Spring wheat lines and a water control, amplified with the gene specific primers for *Pina-D1* (A) and *Pinb-D1* (B). Illustrated (in both (A) and (B)) is the absence of a PCR product in the final DNA sample which is lacking chromosome 5D (N5D/T5A) indicating that the gene specific primers are infact amplifying a single locus in the hexaploid wheat genome.

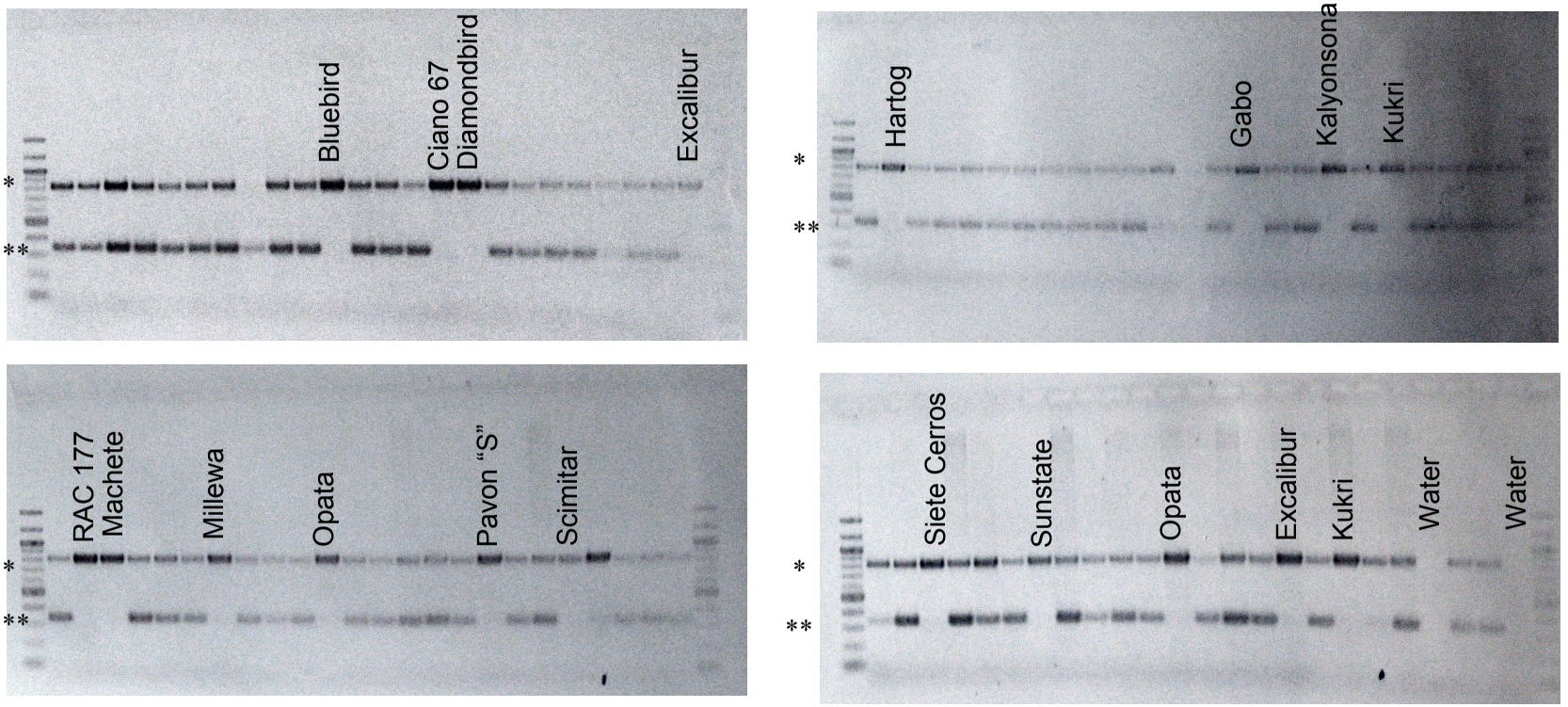


FIGURE 2.2 Gel images of 87 Australian hexaploid wheat lines amplified with the Pina gene specific primers in addition to a glutenin gene used as an internal control in order to detect null *Pina-D1b* alleles (*, Glutenin amplicon; **, Pina amplicon). Seven samples were duplicated to ensure consistency of results and two water controls were included in this assay three of which were amplified in duplicate meaning that 69 unique lines amplified with the Pina primer set.

sequence variations detected. Figure 2.3 shows the consensus DNA sequence of the Pina gene for 44 wheat lines and the 100% conservation that exists within these lines.

Sequence variation amongst the Pinb genes was greatly restricted. Only 1 SNP was detected amongst the 62 wheat lines for which sequence data was obtained. This guanine to adenine base change occurs at nucleotide position 223, which corresponds to amino acid 46, a previously described mutation *Pinb-D1b*, resulting in a hard endosperm (Giroux and Morris 1997).

Figure 2.4 illustrates the 2 haplotypes that are present in the 62 wheat lines sequenced for the Pinb gene. Of these 62 wheat lines 32 have the soft endosperm *Pinb-D1a* allele and the remaining 30 have the glycine to serine change at position 46 (*Pinb-D1b*) resulting in hard grain texture. Within the Pinb gene sequence comparison there were 2 instances of other SNPs in 1 line although each only occurred once and therefore was not used in the LD calculations. RAC 177 has base changes at nucleotide positions 4 and 18. The first substitution at base 4 results in a lysine to a stop codon amino acid change whereas the base change at nucleotide 18 is a synonymous change thereby not altering the leucine residue at that position to any other amino acid.

With respect to the Pinb gene it is reasonable to conclude that there is very little variation between all the lines and this entire region can be separated into 2 distinct haplotypes (Figure 2.4). Because of the notable lack of variation in this region, nucleotide diversity and LD was assessed in a much more diverse germplasm set, *Ae. tauschii,* the progenitor of the D genome to hexaploid wheat.

***Pina-D1a* 1**	ATGAAGGCCC	TCTTCCTCAT	AGGACTGCTT	GCTCTGGTAG	CGAGCACCGC
***Pina-D1a* 51**	CTTTGCGCAA	TATAGCGAAG	TTGTTGGCAG	TTACGATGTT	GCTGGCGGGG
***Pina-D1a* 101**	GTGGTGCTCA	ACAATGCCCT	GTAGAGACAA	AGCTAAATTC	ATGCAGGAAT
***Pina-D1a* 151**	TACCTGCTAG	ATCGATGCTC	AACGATGAAG	GATTTCCCGG	TCACCTGGCG
***Pina-D1a* 201**	TTGGTGGAAA	TGGTGGAAGG	GAGGTTGTCA	AGAGCTCCTT	GGGGAGTGTT
***Pina-D1a* 251**	GCAGTCGGCT	CGGCCAAATG	CCACCGCAAT	GCCGCTGCAA	CATCATCCAG
***Pina-D1a* 301**	GGGTCAATCC	AAGGCGATCT	CGGTGGCATC	TTCGGATTTC	AGCGTGATCG
***Pina-D1a* 351**	GGCAAGCAAA	GTGATACAAG	AAGCCAAGAA	CCTGCCGCCC	AGGTGCAACC
***Pina-D1a* 401**	AGGGCCCTCC	CTGCAACATC	CCCGGCACTA	TTGGCTATTA	CTGGTGA

FIGURE 2.3 Consensus sequence of the Pina gene as determined from sequence analysis of 44 Australian hexaploid wheat lines. There were no SNPs identified across all 44 accessions and this haplotype sequence is consistent with the previously denoted allele *Pina-D1a* (Morris 2002).

Pinb-D1a **1**	ATGAAGACCT	TATTCCTCCT	AGCTCTCCTT	GCTCTTGTAG	CGAGCACAAC
Pinb-D1b **1**	……	……..	……...	……..	……
Pinb-D1a **1**	CTTCGCGCAA	TACTCAGAAG	TTGGCGGCTG	GTACAATGAA	GTTGGCGGAG
Pinb-D1b **51**	……	……	……	……...	……
Pinb-D1a **101**	GAGGTGGTTC	TCAACAATGT	CCGCAGGAGC	GGCCGAAGCT	AAGCTCTTGC
Pinb-D1b **101**	……	……	……	……	……
Pinb-D1a **151**	AAGGATTACG	TGATGGAGCG	ATGTTTCACA	ATGAAGGATT	TTCCAGTCAC
Pinb-D1b **151**	……	……	……	……	……
Pinb-D1a **201**	CTGGCCCACA	AAATGGTGGA	AG**G**GCGGCTG	TGAGCATGAG	GTTCGGGAGA
Pinb-D1b **201**	……	…….....	…..**A**……...	……	……
Pinb-D1a **251**	AGTGCTGCAA	GCAGCTGAGC	CAGATAGCAC	CACAATGTCG	CTGTGATTCT
Pinb-D1b **251**	……	……	……	……	……...
Pinb-D1a **301**	ATCCGGCGAG	TGATCCAAGG	CAGGCTCGGT	GGCTTCTTGG	GCATTTGGCG
Pinb-D1b **301**	……	……	……	……	……...
Pinb-D1a **351**	AGGTGAGGTA	TTCAAACAAC	TTCAGAGGGC	CCAGAGCCTC	CCCTCAAAGT
Pinb-D1b **351**	……	……	……	……	……
Pinb-D1a **401**	GCAACATGGG	CGCCGACTGC	AAGTTCCCTA	GTGGCTATTA	CTGGTGA
Pinb-D1b **401**	……	……	……	……..	……...

FIGURE 2.4 Pinb gene consensus sequences as determined from sequence analysis of 62 Australian hexaploid wheat lines. One SNP at nucleotide position 223 resulted in two distinct haplotypes in the germplasm studied indicated by the allele names *Pinb-D1a* and *Pinb-D1b* following the naming convention by Giroux et al. (1997). The dotted lines in the *Pinb-D1b* sequence indicate no difference in sequence from the *Pinb-D1a* allele sequence directly above it.

2.3.2 LD across the hardness locus of *Ae. tauschii*

At the time that this research was conducted, there was no published information on LD in this species using the public domain sequence data. However, Massa et al. (2004) have recently described the extent of LD in *Ae. tauschii* with results similar to those described in this thesis.

Following the evaluation of gene sequences of the 50 *Ae. tauschii* lines it became clear that there existed far more variation in this set of germplasm relative to the hexaploid wheat lines described earlier. The Pina gene has the least number of SNPs of all 3 genes at the *Ha* locus (Pina, Pinb and Gsp-1) with a total of 5 polymorphic sites and only 3 of which have base variants in more than one of the 50 lines examined. Based on these SNPs the Pina gene can be divide into 6 distinct haplotypes (Figure 2.5) denoted by allele names *Pina-D1a*, *Pina-D1c*, *Pina-D1d*, *Pina-D1e*, *Pina-D1f*, and *Pina-D1g* following the *Pina-D1* naming convention in *T. aestivum* (Morris 2002).

The 3 informative nucleotide polymorphisms are located at bases 57, 66, and 257 with the latter being the only non-synonymous SNP resulting in an amino acid residue change from arginine to glutamine. The remaining two SNPs at positions 156 and 321 are synonymous changes and do not affect the resulting protein sequence.

Gsp-1 has a similar SNP count to that of the Pina gene with 7 polymorphic sites. Three of these 7 SNPs are synonymous base substitutions resulting in transitions and are at nucleotide postitions 66, 315, and 459.

Pina-D1a	1	ATGAAGGCCC	TCTTCCTCAT	AGGACTGCTT	GCTCTGGTAG	CGAGCACCGC
Pina-D1c	1	……	……	……	……	……
Pina-D1d	1	……	……	……	……	……
Pina-D1e	1	……	……	……	……	……
Pina-D1f	1	……	……	……	……	……
Pina-D1g	1	……	……	……	……	……
Pina-D1a	51	CTTTGC**G**CAA	TATAG**C**GAAG	TTGTTGGCAG	TTACGATGTT	GCTGGCGGGG
Pina-D1c	51	……**G**……	……**C**……	……	……	……
Pina-D1d	51	……**A**……	……**C**……	……	……	……
Pina-D1e	51	……**G**……	……**T**……	……	……	……
Pina-D1f	51	……**A**……	……**C**……	……	……	……
Pina-D1g	51	……**G**……	……**C**……	……	……	……
Pina-D1a	101	GTGGTGCTCA	ACAATGCCCT	GTAGAGACAA	AGCTAAATTC	ATGCAGGAAT
Pina-D1c	101	……	……	……	……	……
Pina-D1d	101	……	……	……	……	……
Pina-D1e	101	……	……	……	……	……
Pina-D1f	101	……	……	……	……	……
Pina-D1g	101	……	……	……	……	……
Pina-D1a	151	TACCTGCTAG	ATCGATGCTC	AACGATGAAG	GATTTCCCGG	TCACCTGGCG
Pina-D1c	151	……G……..	……	……	……	……
Pina-D1d	151	……G……..	……	……	……	……
Pina-D1e	151	……G……..	……	……	……	……
Pina-D1f	151	……A……..	……	……	……	……
Pina-D1g	151	……G……..	……	……	……	……
Pina-D1a	201	TTGGTGGAAA	TGGTGGAAGG	GAGGTTGTCA	AGAGCTCCTT	GGGGAGTGTT
Pina-D1c	201	……	……	……	……	……
Pina-D1d	201	……	……	……	……	……
Pina-D1e	201	……	……	……	……	……
Pina-D1f	201	……	……	……	……	……
Pina-D1g	201	……	……	……	……	……
Pina-D1a	251	GCAGTC**G**GCT	CGGCCAAATG	CCACCGCAAT	GCCGCTGCAA	CATCATCCAG
Pina-D1c	251	……**A**……	……	……	……	……
Pina-D1d	251	……**A**……	……	……	……	……
Pina-D1e	251	……**A**……	……	……	……	……
Pina-D1f	251	……**A**……	……	……	……	……
Pina-D1g	251	……**G**……	……	……	……	……
Pina-D1a	301	GGGTCAATCC	AAGGCGATCT	CGGTGGCATC	TTCGGATTTC	AGCGTGATCG
Pina-D1c	301	……	……	C……	……	……
Pina-D1d	301	……	……	C……	……	……
Pina-D1e	301	……	……	C……	……	……
Pina-D1f	301	……	……	C……	……	……
Pina-D1g	301	……	……	T……	……	……
Pina-D1a	351	GGCAAGCAAA	GTGATACAAG	AAGCCAAGAA	CCTGCCGCCC	AGGTGCAACC
Pina-D1c	351	……	……	……	……	……
Pina-D1d	351	……	……	……	……	……
Pina-D1e	351	……	……	……	……	……
Pina-D1f	351	……	……	……	……	……
Pina-D1g	351	……	……	……	……	……
Pina-D1a	401	AGGGCCCTCC	CTGCAACATC	CCCGGCACTA	TTGGCTATTA	CTGGTGA
Pina-D1c	401	……	……	……	……	……
Pina-D1d	401	……	……	……	……	……
Pina-D1e	401	……	……	……	……	……
Pina-D1f	401	……	……	……	……	……
Pina-D1g	401	……	……	……	……	……

FIGURE 2.5 Consensus sequence of 6 haplotypes of the Pina gene as determined from sequence analysis of 50 *Ae. tauschii* wheat lines. There were 5 SNPs identified resulting in the 6 haplotypes. Three of the 5 SNPs were informative, occurring in >1 of the 50 lines examined, which are indicated in bold. The alleles are named *Pina-D1a* and *Pina-D1c* through *D1g* based on the naming convention applied to the same alleles in hexaploid wheat (McIntosh *et al.* 1998).

The remaining 4 SNPs are non-synonymous base substitutions, which result in the alteration of the resultant protein sequence. Nucleotide base postitions 37, 101, 278, and 445 give rise to the following amino acid substitutions; valine to leucine, alanine to valine, glutamine to glycine, and isoleucine to leucine, respectively. Each of these 7 SNPs resulted in 7 distinct haplotypes for this gene which were designated *GSP-D1b*, *GSP-D1c*, *GSP-D1d*, *GSP-D1e*, *GSP-D1f*, *GSP-D1g*, and *GSP-D1h* (Morris 2002).

Of the three genes located at the hardness locus in *A. tauschii*, (Pina, Pinb and Gsp-1) Pinb proved to be the most diverse with a total of 33 polymorphic sites within the 447 bp making up this gene. This equates to, on average, one nucleotide polymorphism every 13.5 bases. These 33 SNPs can be separated into synonymous and non-synonymous base changes and in fact there are 18 synonymous sites and 15 non-synonymous base substitutions (Table 2.3).

TABLE 2.3 Summary of Pinb nucleotide variation and their impact on the resultant protein as observed in the 50 *Ae. tauschii* lines.

Nucleotide Position	SNP	Non-Synonymous (NS) / Synonymous (S)	Amino Acid
48	G to A	S	THR
57	G to A	S	ALA
96	T to C	S	GLY
98	C to G	NS	ALA/GLY
99	A to G	S	ALA
106	A to G	NS	SER/GLY
120	C to T	S	CYS
125	T to A	NS	LEU/GLN
144	C to A	NS	SER/ARG
150	T to C	S	CYS
159	T to C	S	TYR
167	G to A	NS	GLY/GLU
169	T to C	NS	TRP/ARG
171	G to A	S	TRP
186	G to A	S	LYS
196	T to G	NS	PHE/VAL

Table 2.3 Continued

Nucleotide Position	SNP	Non-Synonymous (NS) / Synonymous (S)	Amino Acid
201	T to C	S	THR
210	G to A	S	THR
228	T to C	S	GLY
252	C to G	NS	ASN/LYS
285	G to A	S	GLN
294	C to T	S	CYS
306	A to G	S	ARG
307	G to C	NS	GLY/ARG
310	A to G	NS	MET/VAL
323	A to G	NS	LYS/ARG
339	T to G	NS	PHE/LEU
342	A to C	S	GLY
357	T to G	NS	ASP/GLU
367	A to C	NS	LYS/GLN
370	A to C	NS	ILE/LEU
411	A to C	S	GLY
423	A to G	S	LYS

Based on these 33 base pair substitutions the 50 *A. tauschii* lines examined may be divided into 4 distinct haplotypes (Figure 2.6), the least number of haplotypes amongst the three genes at this locus. One of these 4 haplotypes was identical to a previously identified allele and hence assigned *Pinb-D1a* (Morris 2002), however the remaining 3 haplotypes were unique and have subsequently been assigned allele names *Pinb-D1h*, *Pinb-D1i,* and *Pinb-D1j* (Massa et al. 2004).

In order to estimate the LD in this important region of the wheat progenitor genome, pairwise comparisons between informative SNPs, or those SNPs occurring more than once, was carried out for each gene individually as well as the whole region. Of the 447bp in the Pina gene there are 3 informative sites that can be used to calculate the LD across this gene. Of the 3 pairwise comparisons only one proves to be significant at the $P < 0.001$ level after the Bonferroni correction. There are 7

Pinb-D1a	1	ATGAAGACCT	TATTCCTCCT	AGCTCTCCTT	GCTCTTGTAG	CGAGCACAAC
Pinb-D1h	1	……	……..	……	……	……G….
Pinb-D1i	1	……	……..	……	……	……G….
Pinb-D1j	1	……	……..	……	……	……A….
Pinb-D1a	51	CTTCGCGCAA	TACTCAGAAG	TTGGCGGCTG	GTACAATGAA	GTTGGCGGAG
Pinb-D1h	51	……G …..	……	……	……	……..T...CA..
Pinb-D1i	51	……G …..	……	……	……	……..T...CG..
Pinb-D1j	51	……A …..	……	……	……	……..C...GA..
Pinb-D1a	101	GAGGTGGTTC	TCAACAATGT	CCGCAGGAGC	GGCCGAAGCT	AAGCTCTTGC
Pinb-D1h	101	……..A……..	……C	……..T……..	……	…….C…..…..T
Pinb-D1i	101	……..A……..	……C	……..T……..	……	…….C.. ……..T
Pinb-D1j	101	……..G……..	……T	……..A……..	……	…….A.. ……..C
Pinb-D1a	151	AAGGATTACG	TGATGGAGCG	ATGTTTCACA	ATGAAGGATT	TTCCAGTCAC
Pinb-D1h	151	……..T..	……G...T..	G……	……..G……..	……..T……..
Pinb-D1i	151	……..T..	……G...T..	G……	……..G……..	……..T……..
Pinb-D1j	151	……..C..	……A...T..	G……	……..A……..	……..G……..
Pinb-D1a	201	CTGGCCCACA	AAATGGTGGA	AGGGCGGCTG	TGAGCATGAG	GTTCGGGAGA
Pinb-D1h	201	T……..G	……	……T….	……	……
Pinb-D1i	201	T……..G	……	……T….	……	……
Pinb-D1j	201	C……..G	……	……C….	……	……
Pinb-D1a	251	AGTGCTGCAA	GCAGCTGAGC	CAGATAGCAC	CACAATGTCG	CTGTGATTCT
Pinb-D1h	251	...C……	……	……	……G…….	……C……
Pinb-D1i	251	...C……	……	……	……G…….	……C……
Pinb-D1j	251	...G……	……	……	……G…….	……C……
Pinb-D1a	301	ATCCGGCGAG	TGATCCAAGG	CAGGCTCGGT	GGCTTCTTGG	GCATTTGGCG
Pinb-D1h	301	……..AG….A	……	…..A……	……T…	...A……
Pinb-D1i	301	……..AG….A	……	…..A……	……T…	...A……
Pinb-D1j	301	……..AG….A	……	…..A……	……T…	...A……
Pinb-D1a	351	AGGTGAGGTA	TTCAAACAAC	TTCAGAGGGC	CCAGAGCCTC	CCCTCAAAGT
Pinb-D1h	351	……T……	……...A…..A	……	……	……
Pinb-D1i	351	……T……	……...A…..A	……	……	……
Pinb-D1j	351	……T……	……...A…..A	……	……	……
Pinb-D1a	401	GCAACATGGG	CGCCGACTGC	AAGTTCCCTA	GTGGCTATTA	CTGGTGA
Pinb-D1h	401	……	A……	….A……...	……	…….
Pinb-D1i	401	……	A……	….A……...	……	…….
Pinb-D1j	401	……	A……	….A……...	……	…….

FIGURE 2.6 Consensus sequence of 4 haplotypes of the Pinb gene as determined from sequence analysis of 50 *Ae. tauschii* wheat lines. A total of 33 SNPs were identified, resulting in 4 distinct haplotypes listed below. All 33 SNPs were informative markers, or markers that occurred in more than one of the 50 lines examined, and were all used in the analysis of LD across this gene. The alleles are named *Pinb-D1a* and *Pinb-D1h* through *D1j,* names that were subsequently given to these alleles by Massa *et al.* (2005).

informative polymorphisms that may be used in the estimation of LD across the Gsp-1 gene resulting in 21 pairwise comparisons.Of these 21 comparisons, 6 (29%) proved to be significant using Fisher's exact test yet only one of these remained significant at the $P < 0.001$ level after the Bonferroni correction was applied to the data. Finally the Pinb gene, which has 33 informative polymorphic sites, was evaluated for levels of LD. The 33 informative sites generated 528 pairwise comparisons of which 477 were significant using Fisher's exact test and was subsequently reduced to 422 when the Bonferroni correction was applied. When the r^2 values are plotted against physical distance it is clear that there is gradual decay in LD over distance (Figure 2.7). The logarithmic trend line is superimposed on the graph to visualise this trend and to show that across this gene the level of LD remains above the arbitrarily chosen threshold of $r^2 = 0.2$.

For the entire *Ha* region that is, across all 3 genes (Pina, Pinb and Gsp-1), an estimate of r^2 was determined. For each of the 50 *Ae. tauschii* lines, the gene sequences were concatenated in the order that they appear on chromosome 5D beginning distally with Gsp-1, followed by Pina, and finally the proximally located Pinb. The resulting length of this sequence is 1389 bp and contains 43 informative sites. Of the 903 possible pairwise comparisons 667 (or 74%), are significant according to the Fisher's exact test and is reduced to 54% (486) which are significant at the $P < 0.001$ level following the Bonferroni correction. Figure 2.8 illustrates the LD decay over distance across the *Ha* locus on chromosome 5DS of the 50 *Ae. tauschii* lines. The logarithmic trend line superimposed on Figure 2.8 intercepts the $r^2 = 0.2$ threshold at nearly the full length of this locus indicating that LD does in fact extend across this region and potentially even beyond this region.

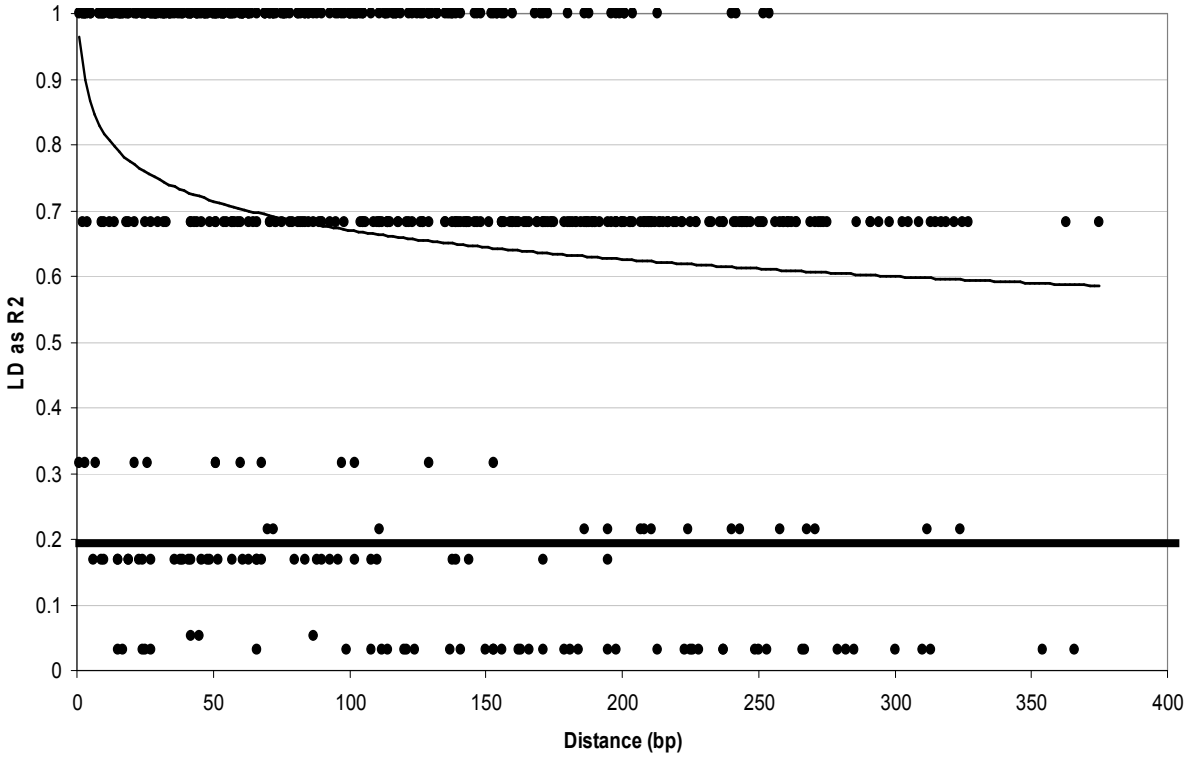


FIGURE 2.7 Scatter plot of LD (r^2) as a function of distance in base pairs across the 447 bases in the Pinb gene of 50 *Ae. tauschii* lines. The trend line is superimposed to highlight the decay of LD with distance, yet this remains above the LD threshold of 0.2 (Rafalski *et al.* 2004).

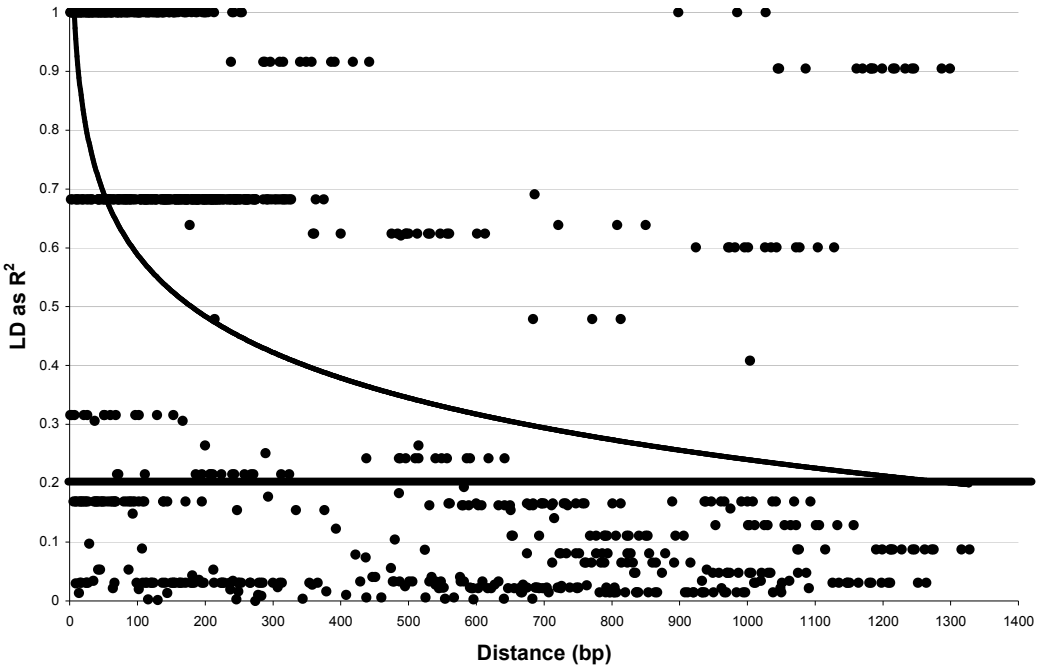


FIGURE 2.8 Scatter plot of LD (r^2) as a function of distance in base pairs across the hardness locus of 50 *Ae. tauschii* lines. Gene sequences from each of the Gsp-1, Pina, and Pinb were concatenated and LD measured across the resulting 1389 bp of sequence. The trend line is superimposed over approximately 1300 bp to highlight the decay of LD to the threshold of $r^2 = 0.2$ (Rafalski *et al.* 2004).

2.4 Discussion

In this chapter, LD was estimated at the sequence level using the Pina and Pinb genes in 44 and 62 wheat lines, respectively. The sequence results obtained demonstrate very little sequence diversity at this region with only 1 SNP being identified in the Pinb gene and none identified in the Pina gene. Based on the sequence of these two genes, the hardness locus can be separated into two distinct haplotypes. The lack of nucleotide diversity and haplotype structure observed across this number of wheat lines didn't allow for the estimation of LD in this region in *T. aestivum*. As such, a comparison was made with a more diverse set of lines (*Ae. tauschii)* since genetically diverse lines will be derived from a more complex recombinational history, it is anticipated that LD will be less extensive than for more genetically related populations (Yu and Buckler 2006). By examining gene sequences from 3 genes at the hardness locus in, 50 *Ae. tauschii* lines, it was evident that the diversity at these genes in this collection was far greater than that of the hexaploid wheat lines examined. There are several reasons that explain why this is the case. Firstly, *Ae.tauschii* is the progenitor of the D genome in common wheat. However; during the hybridisation events leading to hexaploid wheat it is conceivable that only a small number of *Ae. tauschii* lines were involved in the formation of the amphidiploid (Tanksley and McCouch 1997). As a result the diversity observed in the wild species is under represented in the common wheat form. Secondly, selection has a role in shaping the genetic diversity of these genes. Since grain texture is an economically important trait there is higher selection pressure placed on the alleles that govern its control. During domestication and subsequent line development particular alleles were selected that have resulted in the desired phenotype. Lines have therefore gone through a bottleneck resulting in a negative correlation with diversity (Buckler IV and

Thornsberry 2002) and the *A. tauschii* lines, not having undergone this selection, thus remain diverse.

The level of LD within the *Ha* locus in the *Ae. tauschii* lines is elevated yet a decay in LD with physical distance is evident (Figure 2.8). In particular the Pinb gene, the most genetically diverse of the three genes at this locus, has an average LD estimate of r2 = 0.68. The trend line which is superimposed on this figure illustrates that marker associations decay with distance from $r^2 = 1.0$ to $r^2 = 0.58$ over a physical distance of ~375 bp and remain well above the threshold of $r^2 = 0.2$. The most dramatic decline in r^2 values occurs between markers that are separated by 50 bp with an observed decrease in LD from $r^2 = 1.0$ to approximately $r^2 = 0.72$. While there is a continued decline in LD as the physical distance between markers increases, it appears to lessen for markers separated by more than 50 bp. The decline in LD across all three genes (Gsp-1, Pina, and Pinb) at the *Ha* locus, as illustrated in figure 2.8, is more prominent than in the Pinb gene alone with estimated levels of LD declining from $r^2 = 1.0$ to $r^2 = 0.2$ over ~1300 bp. Although there are only 1389 bp of sequence analysed here, the physical distance over which these three genes are located on the short arm of chromosome 5D equates to roughly 70 Kb, based on BAC sequencing in *T. monococcum* (Chantret et al. 2004).With the extended elevated levels of LD in this wheat progenitor sample it is possible to speculate that, at least in this region of the genome, LD it would indeed extend even further in the less diverse species of hexaploid wheat.

How does this compare with other sequence studies based on LD in autogamous plant species? *Arabidopsis thaliana* is 99% self-pollinating and has been shown to have extensive LD up to 250 kb, equating to ~1 cM (Nordborg et al. 2002). In soybean (*Glycine max*), significant LD has been reported to extend beyond 50 kb

and declined to a much lower value of r^2 (0.10) at distances greater than 2.5 cM (Zhu et al. 2003). The distance that LD is appearing to decay in these self-pollinating plant species remains a function of numerous genetic and evolutionary factors, as discussed previously. The most significant difference between these populations and the hexaploid wheat data set is that they consist of a global selection of lines. In general global populations will demonstrate LD that decays at an accelerated rate, relatively speaking, when compared to those, which are locally adapted (Nordborg 1997). Since these latter populations typically have a reduction in heterozygosity at each locus, recombination is less effective at reducing associations and thus there is an increase in the level of LD observed.

Although the LD observed in *Arabidopsis* and soybean is declining rapidly with distance, within certain genes, it remains to be determined if in fact this is a reflection of the whole genome. It has been found in maize that LD is highly variable depending on the genes studied and reports have demonstrated a range of diminished r^2 values between 0.1 and 0.04 over 1.5 kb and 7 kb, respectively (Remington et al. 2001).

In a recent study of the closely related species Hordeum, Caldwell (2004) examined sequence level LD at the hardness locus in 123 lines representing cultivated, landraces, and *H. spontaneum*. Caldwell (2004) observed that the cultivated lines exhibit higher levels of LD in the hardness region of barley with gradual decay observed in the landraces followed by *H. spontaneum*, which demonstrated rapid decays of LD with physical distance. However, unlike the hexaploid wheat results presented in this Chapter, there was greater genetic diversity amongst the cultivated barley lines. The presence of 4 and 8 distinct haplotypes were observed in the cultivated barley varieties across the *Hina* and *Hinb* genes

respectively. This is higher than the complete uniformity observed across the Pina and Pinb genes in the hexaploid wheat samples. Grain texture in barley has been associated with malting quality (Brennan et al. 1996), and as discussed by Caldwell (2004), has only recently become an important trait selected for in breeding programs. This is in contrast to the *Ha* locus in wheat which, because it is such an important quality trait, has been selected by breeders for well over a hundred years (CSIRO 2005). Therefore, despite the similarities in mating type between these two crops there has been a greater period of time for the fixation of the *Ha* locus in wheat in comparison to barley. As such there is a reduction in diversity at this locus in wheat and it is likely that LD will extend over greater distances in this species.

A recently published article on the extent of LD in a polyploid wheat species will undoubtedly provide the most appropriate comparison for the results obtained in this hexaploid wheat study. In the tetraploid wheat species durum, widespread LD has been observed throughout the genome through the use of SSR markers (Maccaferri et al. 2005). This includes markers, which are tightly linked (<10 cM) as well as markers that are independent, or located on different chromosomes (Maccaferri et al. 2005). The highest number of significant LD pairs was observed between markers separated less than 10 cM (average D' = 0.67) and nearly all of them significant at the $P < 0.01$ level. The magnitude of the LD value decreased slightly (D' of approximately 0.3) and the significant pair-wise comparisons was reduced by nearly half when independent markers were assessed (54 % significant at the P <0.01 level). It would not be surprising then to predict that similar levels of LD in hexaploid wheat will be observed.

A rapid breakdown in LD would be ideal for the fine mapping of QTL on a candidate gene basis. However, the results presented here suggest that this would not

be feasible in hexaploid wheat. There is a need to account for evolutionary and genetic factors such as population structure and rare alleles as these are known to temporarily increase the observed levels of non-random associations (Thornsberry et al. 2001). It remains unclear as to whether or not association mapping of traits based on the premise of LD will be feasible in hexaploid wheat and further research is required, within this population in particular, to determine the point at which LD decays to insignificant levels.

Chapter 3: Population Structure within experimental hexaploid wheat populations

3.1 Introduction

Through the use of molecular markers, association mapping has proved a useful tool in the identification and localisation of disease genes in humans. Association mapping or linkage disequilibrium (LD) mapping relies on drawing marker/trait associations between individuals in large natural populations. By genetically surveying a large group of individuals it is possible to draw correlations between phenotype and a causative genotype or allele. This becomes problematic when there is a population sub-division or sub-structure in the large group of individuals being examined. Consider for example, a large group of individuals which is segregating for a specific phenotype. Within that group there exists distinct sub-groups, to one of which the phenotype is specific. By genetically surveying the population as a whole however, alleles which are present in a relatively high frequency in the remaining sub-groups could appear to be linked to the phenotype.

This presents an elevated type 1 error rate and the consequences of not accounting for population structure are that alleles that are not genuinely associated with the target phenotypic variation may appear to be so. There is a good example of the effect of population structure in maize when estimating the association of the *Dwarf8* gene with variations in flowering time. Thornsberry et al. (2001) tested association of the *Dwarf8* gene with flowering time in a population of 92 inbred maize lines that knowingly could be divided into three major groups. By comparing results from analysis of the population as a whole and within sub-populations, the type 1 error rate was reduced up to 4.7-fold.

There are many further reasons in addition to association genetics that render the understanding of sub-populations important. Understanding the history of a population, classifying sub-species, identifying species origin and determining 'within species differentiation' are all important reasons to study structure within populations. In plants particularly, the main method of classifying individuals is based on phenotype, origin and pedigree. Although effective for some methods of differentiation, these are all subjective and, in the case of pedigree, can be the source of some debate. By using genotype data some of the uncertainty is removed and this is possibly the most accurate method of determining the true structure of a population. Pritchard et al. (2000) have developed a software package called *Structure,* which uses multi-locus genotype data to classify individuals into genetic clusters based on the amount of genetic background that was contributed by the group.

Pritchard et al. (2000) successfully applied this algorithm to two different data sets, for which prior sub-structure was known, in order to test its ability to separate individuals into their respective clusters. The first data set consisted of 155 birds sampled from 4 different geographic locations. Using 7 SSR markers they were able to assign nearly all of the birds to a cluster that was indicative of the geographic region from which they were sampled. In the second data set 162 humans were sampled from two distinct ancestries. Using 30 bi-allelic restriction site polymorphisms, they were able to correctly assign the individuals to one of the distinct clusters. This algorithm is capable of distinguishing relatively small populations into a small number of sub-groups with a small number of markers. To determine if the same algorithm could be used on much larger sample sizes with a greater number of clusters, Rosenberg et al. (2001) sought to classify 600 chickens into 20 different clusters based on previously known breed types. Using a sample of

27 SSR markers, they were able to identify clusters representing the 20 chicken breeds at a success rate of approximately 98% each time.

As it is clear that population structure will directly impact the results from association studies, it is important to account for it in some way. The *Structure* software provides a reliable method to make a distinction between very similar individuals in a particular population. *Structure* was therefore used to infer population sub-structure in the wheat data set used in this study.

A total of 96 and 225 wheat lines specific to South Australia and the United Kingdom, respectively, were examined for population structure with 25 multi-locus markers. This marker data set is similar in size to that of the chicken study and includes wheat varieties from two very different parts of the world. Historical pedigrees from each of the lines were compared to determine if the clusters generated by *Structure* were consistent with the pedigree information.

3.2 Materials and Methods

3.2.1 Plant material

The 96 Australian wheat lines used in this experiment are outlined in Section 2.2.1.

Two hundred and twenty five hexaploid wheat varieties from the United Kingdom were also included in this study (Appendix B). These lines represent key wheat varieties that are currently used in the UK wheat breeding programs as well as varieties that have contributed to the individual pedigrees. Permission to include the genotype data from these lines in this study was kindly provided by Pauline Stephenson, John Innes Centre, Norwich, UK.

3.2.2 Microsatellite markers

A total of 150 wheat microsatellite markers were used in the genotyping of the 96 Australian wheat varieties. Twenty-three microsatellite primer sequences were obtained from the Agrogene wheat microsatellite consortium (WMC) (Gupta et al. 2002). The remaining 127 microsatellite primer sequences were developed at the Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben (GWM and GDM) (Pestsova et al. 2000; Roder et al. 1998). The markers were selected based on their map location, in order to provide basic coverage of the wheat genome (Table 3.1).

TABLE 3.1. Genomic distribution of 150 microsatellite markers used in genotyping 96 Australian wheat varieties.

Genome	A	B	D	Total from each group
Group				
1	4	6	6	16
2	10	6	8	24
3	5	7	6	18
4	5	5	2	12
5	6	6	8	20
6	3	4	4	11
7	14	20	15	49
Total from each genome	47	54	49	Total: 150

Marker locations were based on consensus molecular marker maps generated by (Chalmers et al. 2001; Somers et al. 2004), and from the KOMUGI Integrated

Wheat Science Database,
(http://www.shigen.nig.ac.jp/wheat/komugi/maps/markerMap.jsp).

Within the Australian germplasm, the group 7 chromosomes were selected for high marker density coverage to obtain information on optimal marker density for subsequent LD and association mapping studies in wheat. A total of 49 SSRs specific to the group 7 chromosomes of wheat were included in this study. Each SSR primer pair consisted of a forward primer, which was fluorescently labelled at the 5' end with one of three tags: FAM (Invitrogen, USA), NED (Applied Biosystems, USA) or HEX (Invitrogen, USA) while the reverse primer (Invitrogen, USA) remained unlabelled.

The microsatellite genotyping of UK wheat varieties was conducted at the John Innes Centre as part of the Genome Analysis of Agriculturally Important Traits (GAIT) program. Genotype data for 97 SSR loci were available from the JIC database (http://jic-bioinfo.bbsrc.ac.uk/cereals/). Of these 97 SSR loci, 25 were in common with those used in the Australian germplasm study (Table 3.2).

As a result, these 25, unlinked, SSR markers were chosen to evaluate the population structure of these two larger datasets.

3.2.3 PCR amplification microsatellite markers in hexaploid wheat

DNA isolations are outlined in sections 2.2.1 and 2.2.2. Each PCR was set up in 384-well microtitre plates (ABgene, UK) with a final reaction volume of 12.5 µl. In each case the reaction mix contained 1x PCR Buffer, 1.5 mM $MgCl_2$, 0.2 mM each dNTP (0.8 mM), 1 µmole of each primer (labelled forward and unlabelled reverse), 1U Taq DNA polymerase (Invitrogen, USA) and 50 ng total genomic DNA.

TABLE 3.2. Twenty-five, unlinked, SSR markers used in estimating population structure of the Australian and UK hexaploid wheat data sets.

SSR locus	Left Primer	Right Primer
Xgwm357-1A	TAT GGT CAA AGT TGG ACC TCG	AGG CTG CAG CTC TTC TTC AG
Xgwm99-1A	AAG ATG GAC GTA TGC ATC ACA	GCC ATA TTT GAT GAC GCA TA
Xgwm11-1B	GGA TAG TCA GAC AAT TCT TGT G	GTG AAT TGT GTC TTG TAT GCT TCC
Xgwm337-1D	CCT CTT CCT CCC TCA CTT AGC	TGC TAA CTG GCC TTT GCC
Xgwm458-1D	AAT GGC AAT TGG AAG ACA TAG C	TTC GCA ATG TTG ATT TGG C
Xgwm558-2A	GGG ATT GCA TAT GAG ACA ACG	TGC CAT GGT TGT AGT AGC CA
Xgwm311-2A	TCA CGT GGA AGA CGC TCC	CTA CGT GCA CCA CCA TTT TG
Xgwm257-2B	AGA GTG CAT GGT GGG ACG	CCA AGA CGA TGC TGA AGT CA
Xgwm157-2D	GTC GTC GCG GTA AGC TTG	GAG TGA ACA CAC GAG GCT TG
Xgwm539-2D	CTG CTC TAA GAT TCA TGC AAC C	GAG GCT TGT GCC CTC TGT AG
Xgwm369-3A	CTG CAG GCC ATG ATG ATG	ACC GTG GGT GTT GTG AGC
Xgwm674-3A	TCG AGC GAT TTT TCC TGC	TGA CCG AGT TGA CCA AAA CA
Xgwm155-3A	CAA TCA TTT CCC CCT CCC	AAT CAT TGG AAA TCC ATA TGC C
Xgwm389-3B	ATC ATG TCG ATC TCC TTG ACG	TGC CAT GCA CAT TAG CAG AT
Xgwm160-4A	TTC AAT TCA GTC TTG GCT TGG	CTG CAG GAA AAA AAG TAC ACC C
Xgwm304-5A	AGG AAA CAG AAA TAT CGC GG	AGG ACT GTG GGG AAT GAA TG
Xgwm186-5A	GCA GAG CCT GGT TCA AAA AG	CGC CTC TAG CGA GAG CTA TG
Xgwm291-5A	CAT CCC TAC GCC ACT CTG C	AAT GGT ATC TAT TCC GAC CCG
Xgwm126-5A	CAC ACG CTC CAC CAT GAC	GTT GAG TTG ATG CGG GAG G
Xgwm540-5B	TCT CGC TGT GAA ATC CTA TTT C	AGG CAT GGA TAG AGG GGC
Xgwm190-5D	GTG CTT GCT GAG CTA TGA GTC	GTG CCA CGT GGT ACC TTT G
Xgwm292-5D	TCA CCG TGG TCA CCG AC	CCA CCG AGC CGA TAA TGT AC
Xgwm325-6D	TTT CTT CTG TCG TTC TCT TCC C	TTT TTA CGC GTC AAC GAC G
Xgwm46- 7B	GCA CGT GAA TGG ATT GGA C	TGA CCC AAT AGT GGT GGT CA
Xgwm295-7D	GTG AAG CAG ACC CAC AAC AC	GAC GGC TGC GAC GTA GAG

The PCR program was a standard touchdown PCR with an initial 20 cycles of 94°C for 30 seconds, 60°C for 30 seconds (decreasing by 0.5°C with each cycle) and 72°C for 30 seconds. There was a further 29 cycles of 94°C for 30 seconds, 50°C for 30 seconds and 72°C for 30 seconds followed by final extensions of 72°C for 5 minutes and 25°C for 5 minutes.

3.2.4 ABI Prism 3700 DNA analyzer

All microsatellite PCR products were analysed using the ABI Prism 3700 DNA analyser (Applied Biosystems, CA), a multi-capillary electrophoresis system capable of analysing multiple runs of 96 and 384 samples. One microliter of each PCR product was diluted in 50 µL of water in order to be at an acceptable concentration for analysis. From this dilution, 1.5µL of FAM labelled products, 4.5 µL of HEX labelled products and 5 µL of NED labelled products were pooled with a maximum of 3 primers per labelled dye being multiplexed together. These pooled samples were brought to a final volume of 28 µL with water and centrifuged briefly. Three microliters from each pooled sample were aliquoted into a clean plate and 5 µL of ROX:Formamide (1:100) were added to each sample. The plate was then centrifuged and the samples were placed on the DNA analyser.

Sample plates were placed into "carrier plates" and secured on the platform. The "Genescan" option was selected for the SSR marker analysis and program parameters were set following the manufacturers recommendations.

3.2.5 Microsatellite analysis

Results from the ABI 3700 were first analysed with Genescan Version 3.6 (Applied Biosystems, USA), which standardises fragment sizes based on the ROX internal size standard. Further analysis was done using Genotyper Version 3.6 (Applied

Biosystems, USA). The Genotyper software was used to automate the fluorescent microsatellite allele calling and allowed for the analysis of large data sets at one time. Automated allele calling was largely accurate, however since SSRs typically generate stutter bands, and with the ploidy level of hexaploid wheat, manual verification of each marker in each variety was carried out. The genotyper software was also used to set up categories, sort peak data (fragment sizes) and correlate peak and marker information into tables that were exported into MS Excel for further analysis.

3.2.6 Cluster analysis

Using the *Structure* software by Pritchard et al. (2000), population structure of the Australian and UK hexaploid wheat lines was examined using the admixture model. This model allows for the assumption that each individual in the study has inherited some proportion of its genome from, potentially, each population and incorporates the possibilities that each allele at a particular locus may have arisen from a different effective population.

Structure also allows for the arbitrary assignment of a number of different parameters including; the estimated number of sub-populations (K values), the burn-in period (the number of runs prior to data collection to reduce the effect of the starting configuration), and subsequent number of runs (appropriate number of runs in order to obtain accurate estimates of the data). It is important when estimating population sub-division from this cluster analysis that the results between runs are consistent, therefore a number of iterations were carried out and each set of results plotted and compared. Since the appropriate parameter settings need to be determined, analysis, including 90 of the 96 original South Australian wheat lines, was initially carried out. The total number of lines included in this analysis was reduced to 90, since 4 had missing data values >25% and a further 2 lines were very

closely related to lines already included in the study. Subsequently the population structure was estimated in the UK data set, which also contained fewer lines than the original population. The number of UK wheat lines was reduced in this study from 225 to 184 lines since 40 lines had missing data values >25% and one line was duplicated in the data set.

Using the 25, unlinked, SSR markers common to each the Australia and UK data set, an initial run was set up with a burn-in period and run length of 5 000 and 50 000 iterations, respectively and K values set from 1 to 23. A total of five different run configurations were performed with the burn-in period ranging from 5 000 to 150 000 iterations, subsequent run lengths ranging from 3 000 to 100 000 but keeping the potential K values to a range of between 1-23 (Table 3.3).

TABLE 3.3. Five experiments set up in Structure in order to determine the most appropriate parameter settings for the program. Results from these experiments were used in estimating the underlying population structure in the Australian and UK hexaploid wheat data sets.

Experiment Number	Total run length	Burn-in period	Subsequent Runs	K values	Iterations
1	8 000	5 000	3 000	1-23	5
2	55 000	5 000	50 000	1-23	5
3	80 000	20 000	60 000	1-23	5
4	100 000	80 000	20 000	1-23	5
5	250 000	150 000	100 000	1-23	5

Five iterations of each experiment were performed. The *Structure* software is time consuming to run requiring 14-18 days for a couple of these experiments. Running five iterations over large potential K values provided an idea of the

distribution of lines in different populations. Upon deciding that a burn-in period of 80 000 runs together with a run length of 20 000 runs was sufficient, the 180 UK varieties were then evaluated for their level of population subdivision using the same 25 SSR markers.

The results from both analyses were examined and only those lines that consistently grouped together and had a portion of their genome of 80% or higher originating from one of the estimated number of clusters were selected and used in the subsequent LD analysis (see Chapter 4). Once the most likely number of sub-populations was established, data from the other K values were not included in the rest of the analysis.

3.3 Results

3.3.1 Establishing program parameters

Initial analyses were performed to identify the most appropriate run parameters for consistently separating lines within the complete set of Australian wheat lines. As mentioned above, Table 3.3 specifies the types of tests that were performed and the run lengths of each. Results from each of these runs were indicative of some population structure, since the log probability of the structure (lnPr(X|K)) increased with each run, starting with K = 1 population.

The posterior probabilities of each iteration from Experiments 1 to 5 were plotted in MS Excel to determine which of the experiments provided the most consistent results. Experiment 1 started with an lnPr(X|K) value of approximately -6 800 for K = 1 population. This number increased steadily to approximately -4 440 for K = 10 populations and resulted in an average probability of nearly -5 300 at the

maximum number of populations tested. Despite the increase to a relatively high probability value, the plot of these values against the number of proposed sub-populations shows that the values are inconsistent between runs. Figure 3.1 illustrates the increase in the log probability of structure for each of the 5 iterations using the parameters outlined in Table 3.3 for Experiment 1. From the data in this graph, the results from each iteration, up to K = 10, is relatively consistent. At K = 11 however, the consistency of the results between each run breaks down.

In Experiment 2 the burn-in period was left at the same length as in Experiment 1 but the number of subsequent runs was increased nearly 17-fold to determine if this parameter had an influence on the consistency of the results beyond K = 10. Figure 3.2 again represents the lnPr(X|K) plotted for each proposed value of K from 1-23. The results from Experiment 2 are very similar to those of Experiment 1 with the log probabilities starting at approximately -6 570 for K = 1 and increasing to averages of -4 200 and -5 820 for K = 10 and 23, respectively. From these results and those obtained in Experiment 1, the increase in run length alone is not enough to gain consistency amongst results and hence allow for the confident selection of the correct K value.

With this in mind the number of runs in Experiment 3 were increased by 25 000, in addition to increasing the burn-in period to 20 000 reps followed by 60 000 subsequent runs. LnPr(X|K) values ranged between approximately -6 580 for K = 1 and approximately -4 200 for K = 10 after which the results from each of the 5 iterations remained very sporadic (Figure 3.3). In Experiment 4 the burn-in period was increased 4-fold to a total of 80 000 runs followed by a run length of 20 000. The results from this experiment were more consistent between runs and followed a steady increase in the probability of the data from K = 1 to K = 23. LnPr(X|K) values

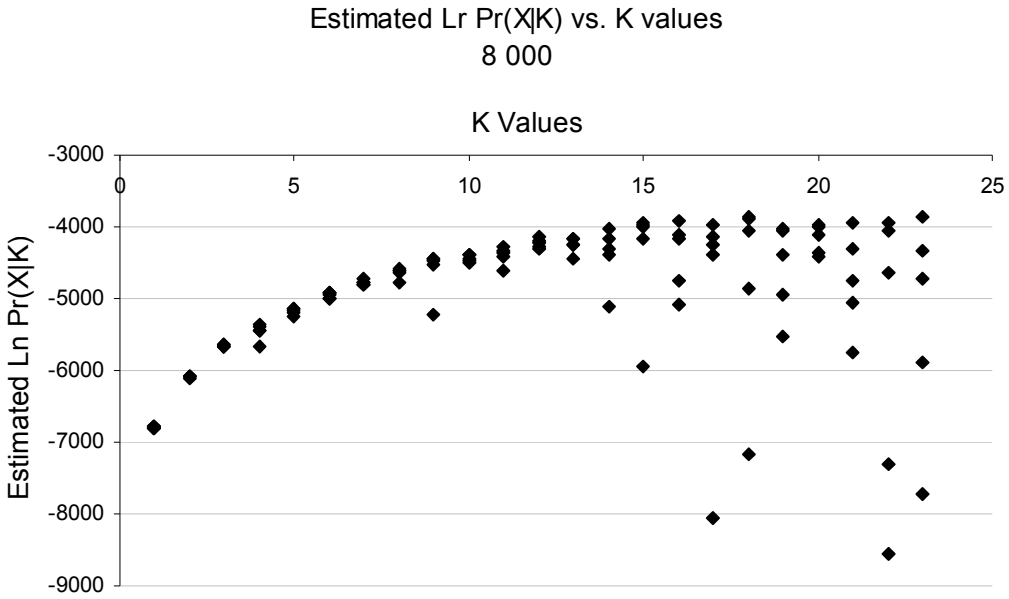


FIGURE 3.1 Experiment 1. Scatter plot of the log probability of the data versus the estimated K values (1 to 23) using 90 Australian wheat lines. These results represent the five iterations carried out in Structure when determining the most appropriate parameter settings. The parameters used in this experiment were a burn-in period of 5,000 followed by 3,000 subsequent runs for a total run length of 8,000.

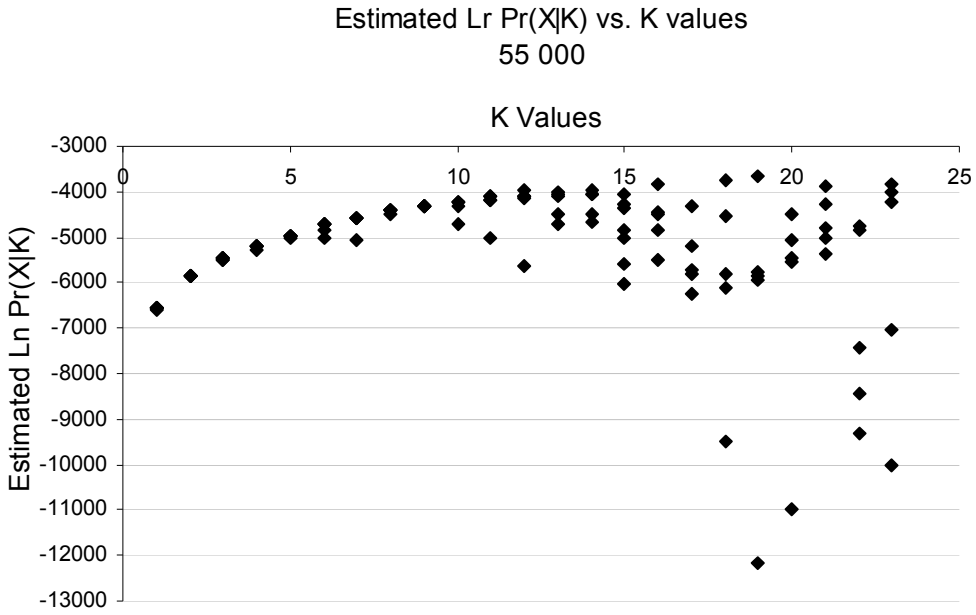


FIGURE 3.2 Experiment 2. Scatter plot of the log probability of the data versus the estimated K values (1 to 23) using 90 Australian wheat lines. These results represent the five iterations carried out in *Structure* when determining the most appropriate parameter settings. The parameters used in this experiment were a burn-in period of 5,000 followed by 50,000 subsequent runs for a total run length of 55,000.

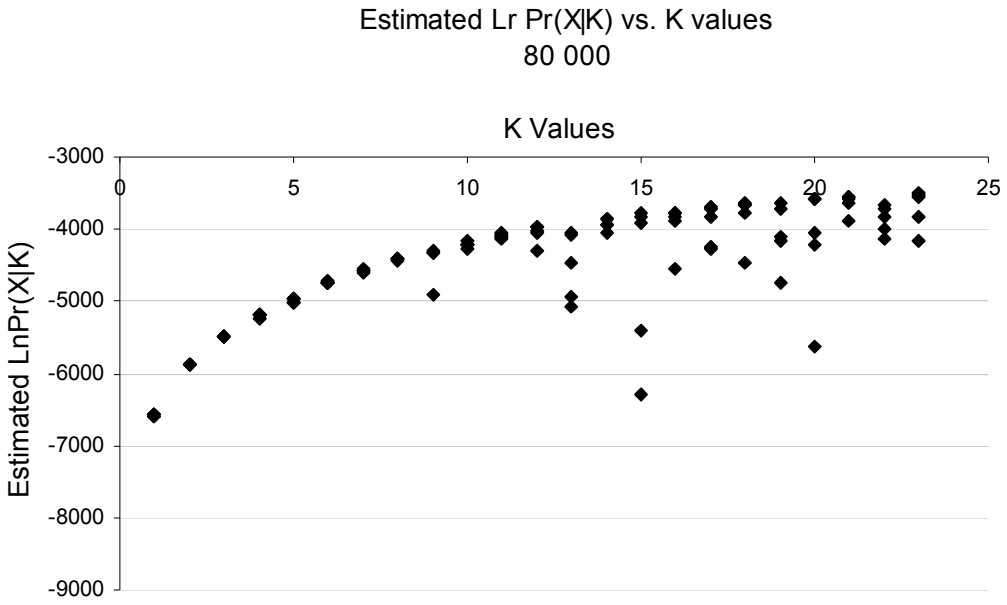


FIGURE 3.3 Experiment 3. Scatter plot of the log probability of the data versus the estimated K values (1 to 23) using 90 Australian wheat lines. These results represent the five iterations carried out in *Structure* when determining the most appropriate parameter settings. The parameters used in this experiment were a burn-in period of 20,000 followed by 60,000 subsequent runs for a total run length of 80,000.

again varied from roughly –6 570 at K = 1 to roughly -3 560 for K = 23. Through the substantial increase in the burn-in period, the consistency of the results between the 5 iterations conducted in Experiment 4 was greatly improved over the other parameter settings (Figure 3.4).

To determine whether a further increase in the length of both the burn-in and subsequent runs would generate substantially different results, Experiment 5 was conducted. As illustrated in Figure 3.5, the consistency of results between each run is comparable to that observed in Experiment 4. The LnPr(X|K) values ranged from -6 530 for population K = 1 to approximately -3 020 for 21 estimated populations. It should be noted that in this last experiment, the number of calculations that were required was considerably larger than the previous 4 experiments. As such the memory capacity of the computer used for such simulations reached its maximum and thus data is only available for estimations in population numbers from 1 to 21 in Experiment 5 and not 1 through 23, as in the first four experiments.

Since there were no substantial differences between the results obtained in Experiments 4 and 5 the parameters used in Experiment 4 were used in the subsequent estimation of sub-populations in the larger UK data set. The UK data set was examined by using the same burn-in and subsequent run lengths of 80 000 and 20 000, respectively. Figure 3.6 illustrates the steady increase in the log probability of the data as well as consistency between the 5 runs. The LnPr(X|K) values were significantly lower than those observed in the Australian data set starting with -10 200 for K = 1 populations, -7 040 for K = 10, and – 5 830 for the maximum number of estimated sub-populations (23).

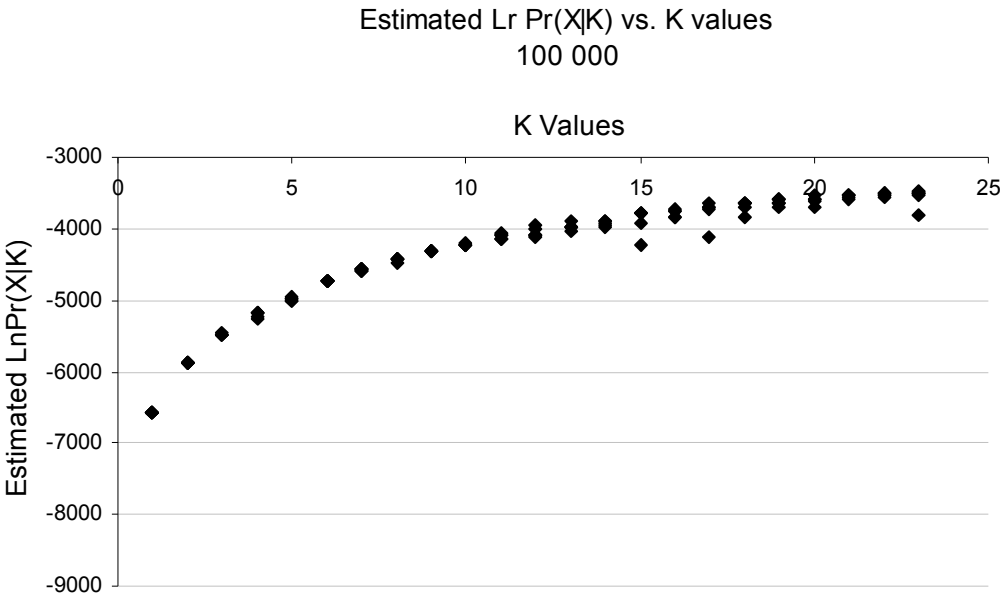


FIGURE 3.4 Experiment 4. Scatter plot of the log probability of the data versus the estimated K values (1 to 23) using 90 Australian wheat lines. These results represent the five iterations carried out in *Structure* when determining the most appropriate parameter settings. The parameters used in this experiment were a burn-in period of 80,000 followed by 20,000 subsequent runs for a total run length of 100,000.

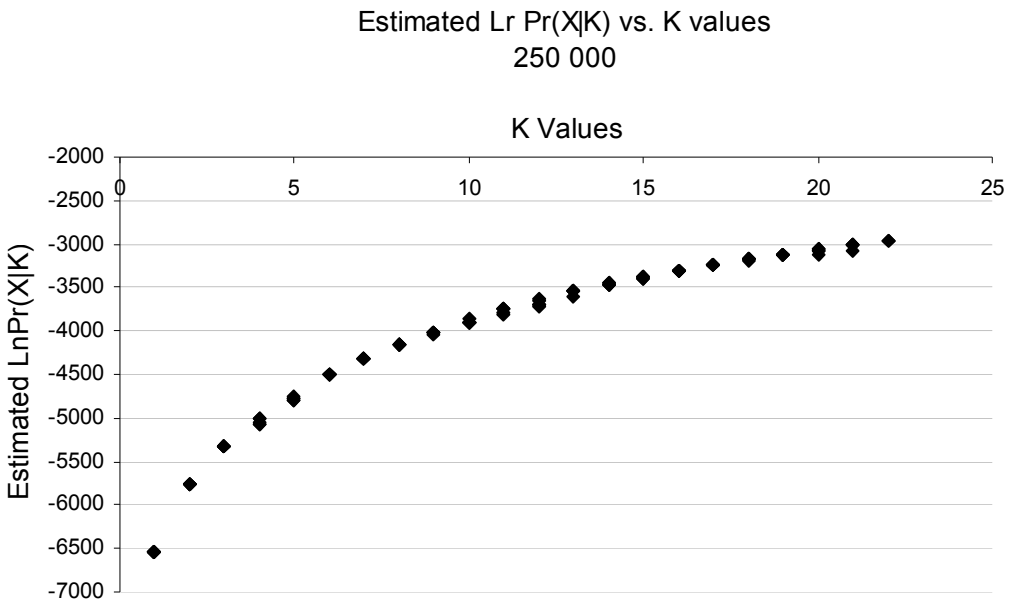


FIGURE 3.5 Experiment 5. Scatter plot of the log probability of the data versus the estimated K values (1 to 23) using 90 Australian wheat lines. These results represent the five iterations carried out in Structure when determining the most appropriate parameter settings. The parameters used in this experiment were a burn-in period of 150,000 followed by 100,000 subsequent runs for a total run length of 250,000.

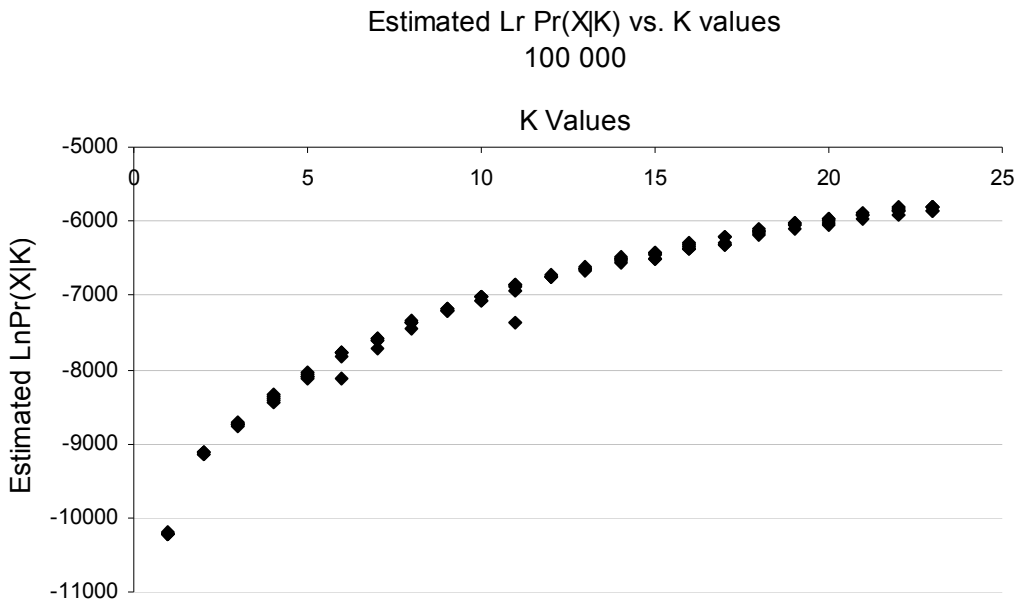


FIGURE 3.6 UK data set. Scatter plot of the log probability of the data versus the estimated K values (1 to 23) using 184 UK wheat lines. These results represent the five iterations carried out in Structure when determining the most appropriate parameter settings. The parameters used in this experiment were a burn-in period of 80,000 followed by 20,000 subsequent runs for a total run length of 100,000.

3.3.2 Deciding on the most appropriate value of K

With the parameters set to those established in Experiment 4, the most appropriate value of K needed to be determined. Estimating the number of sub-populations from a larger data set is not straightforward. (Pritchard et al. 2000) points out that probability values for data in sub-populations less than the ideal number are very small and will eventually increase and plateau when a higher value of K is attained. It is where the likelihood of the K values plotted on the graph is similar that the true number of sub-populations exists and is typically the smaller number of populations. In both the Australian and UK germplasm after 5 iterations of the algorithm, the results were very consistent and the probability figures became slightly less variable around K = 16. Beyond this point there remained slight variation between the likelihood values for each estimation of K.

Factors such as the model used and the level of inbreeding within the population may all work to over estimate the probability of having higher sub-populations than there should be (Pritchard et al. 2000). What is clearly lacking in the results from both data sets in this study is an indication of an appropriate estimate of the number of sub-populations in the form of the probability values leveling off. Since there is no clear indication as to the most appropriate number of sub-populations in these larger data sets, *Structure* is perhaps not the most suitable method for use in determination of sub-population identity. This point will be discussed later.

3.3.3 Identifying sub-populations in Australian and UK data sets

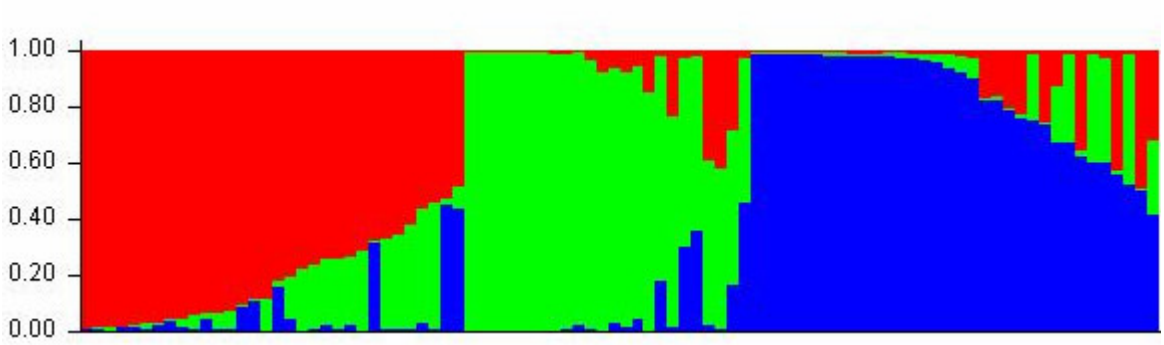
Using the admixture model accounts for the possibility of 'X' percent of the genome originating in one of the K estimated populations. To obtain a population with

minimal structure, lines that consistently grouped together and had >80 percent of their genome originating from one population were evaluated.

Since there was no clear indication as to which K value was more 'correct', the lines that were sub-divided in the estimated K = 3 populations were chosen for further analysis such that at least some of the population structure was accounted for in the subsequent LD analysis. An example of the output of the structure analysis for K = 3 populations is depicted in Figure 3.7 for both the Australian and UK data set. Each individual in the larger data set is represented by a thin vertical coloured bar which indicates the proportion of the genome originating from a particular population. Each of the five iterations generated nearly identical partitioning of lines into sub-groups in both data sets. The Australian sub-group with the largest number of lines having >80% of their genome originating from one sub-population was the third section of the plot, indicated by the blue section in Figure 3.7A. Lines from this sub-population are listed in Table 3.4.

Pedigree information provided by breeders is another method by which the preliminary partitioning of lines into smaller groups may be confirmed. Table 3.4 also contains pedigree data for the 22 Australian wheat lines that were initially grouped together in *Structure* based purely on their genetic make-up. Perhaps the most notable feature of the 22 lines in this table is that they all originated in Australia (except for Bowie); furthermore, half of them are from South Australia. It is also clear from looking at the pedigree data that the lines are very closely related, either sharing common ancestors, like Bindawarra and Warimek, or being ancestral lines themselves, such as Ranee and Ghurka.

(A)



(B)

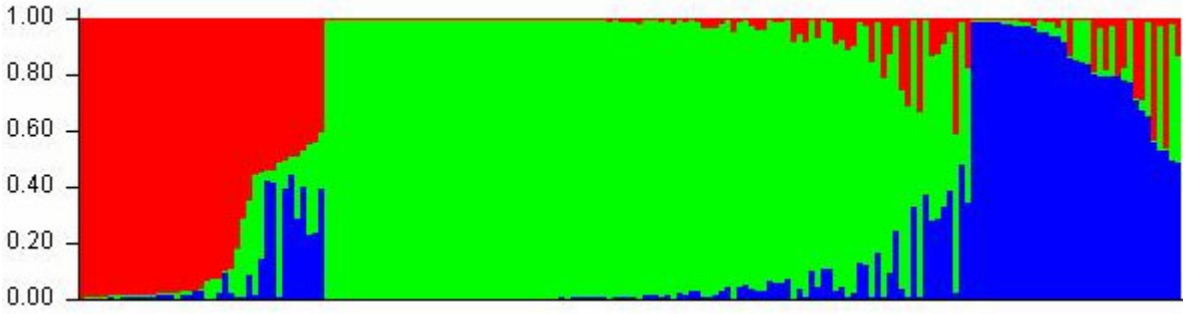


FIGURE 3.7 Estimation of population structure for the Australian (A) and UK (B) data sets for population estimations K = 3. Percent of genome originating from each sub-group is depicted by the three different colours in each plot. The Australian data set of 90 lines (A) when divided into K = 3 populations results in 22 lines consistently grouping together, which is represented by the blue section below. The UK data set of 180 lines (B) when separated into K = 3 populations results in 93 lines consistently grouping together (depicted by the green section below).

TABLE 3.4 Pedigree information for 22 Australian wheat lines, making up the largest sub-population based on the genetic clustering algorithm Structure.

GermplasmName	Source	Pedigree	Origin	Year
Bindawarra	Waite	Mexico-120/Koda//Raven,	AUS:SA	1980
Bowie	Tamworth	Renacimiento/Kenya-C-10862, Renacimiento//Kenya/Gular	USA:Tx	1953
Camm	Tamworth			
Dagger	Waite	Rac-111/Insignia,	AUS:SA	1983
Dirk48	Tamworth	Gabo/4*Dirk;	AUS:SA	1951
Frame	Tamworth	Molineux/3*Dagger,	AUS:SA	1997
Gabo-Aus	Waite	Bobin(S)/(Tr.Dr)Gaza//(S)Bobin;Gular/(Tr.Dr)Gaza//Gular;Bobin*2/Gaza;Bobin-W-39//Bobin-W-39/Gaza;	AUS:NSW	1951
Gamenya	Waite	Kenya-117-A/2*Gabo//Mentana/6*Gabo, Gabo/3/Gabo*5/Mentana//Gabo*2/Kenya-117-A, Gabo/4/Gabo*5/Mentana//2*Gabo/3/Kenya-117-A,	AUS:NSW	1958
Ghurka	Waite	Gallipoli/3/Currawa//Indian-4-E/Federation, Indian-H/Federation//Currawa, Zaff/Yandilla-King,	AUS:Vic	1924
Halberd	Waite	Scimitar/Kenya-C-6042//Bobin/3/Insignia-49,	AUS:SA	1969
Heron	Waite	Ranee/Doubbi//Ranee(R.D.R.)/3/4*Insignia-49,	AUS:NSW	1958
Insignia	Waite	Ghurka/Ranee,	AUS:Vic	1946
Machete	Waite	Mec-3/2*Gabo(Rac-177)//Madden,	AUS:SA	1985
Mengavi	Tamworth	Eureka,Aus/Ci-12632//2*Gabo/3/Mentana/6*Gabo,	AUS:NSW	1958
Olympic	Waite	Baldmin/Quadrat;	AUS:Vic	1956
Rac177	Waite	Mec-3/2*Gabo;	AUS:SA	1977
Ranee	Waite	Indian-F/Federation,	AUS:Vic	1924
Spear	Waite	Rac-111/Insignia,	AUS:SA	1983
Tatiara	Waite	Mexico-120/Koda//Raven/3/Mengavi/Siete-Cerros-66,	AUS:SA	1988
Trident	Waite	Vpm-1/5*Cook//4*Spear, Vpm-1/4*Spear, Spear*4/Vpm-1,	AUS:SA	1993
Uraguay1064	Tamworth			
Warimek	Tamworth	Mexico-120/Koda,	AUS:SA	1971

With respect to the UK data set, similar results were observed. Of the initial 180 lines examined for population structure, the largest sub-group of lines at the K = 3 value consisted of 93 individuals (Figure 3.7B). Table 3.5 contains the pedigree data for this sub-group of lines that consistently group together in *Structure*. Again as with the analysis of the Australian data sets, all lines, except Daphne and Samson, originate from Europe with 69% of the lines originating from Great Britain. The pedigree data also reveals the relatedness of the lines again though common ancestors, such as Adroit and Aristocrat and many of the "Maris" lines, or being present in the pedigree itself, such as Cappelle Deprez and Hobbit.

3.4 Discussion

In this chapter, 25 microsatellite markers distributed across the wheat genome, were used to detect underlying sub-structure of 90 Australian wheat lines and 184 wheat lines from the UK. An estimated number of sub-populations (K = 1-23) provided little insight into the population structure of these larger data sets and *Structure* alone was unable to divide these lines, based on statistical probability, into distinct groups.

From the graphs in Figures 3.4 and 3.6 there is no clear, distinct plateau observed in the data which is described by (Pritchard et al. 2000) as the identifying point at which the estimated number of sub-groups is located. There are several reasons to explain the absence of this feature in both the Australian and UK data sets. Firstly, plant varieties are known not to fit the assumption that they originated from populations with distinct ancestries. Despite the fact that there are individual breeding groups, each has been introgressed with material from other groups creating lines with very complex breeding histories. The *Structure* results presented here reflect this. Secondly, the lack of distinct sub-groups may be the result of inadequate analysis and may actually

TABLE 3.5 Pedigree information for 93 UK wheat lines, making up the largest sub-population based on the genetic clustering algorithm Structure.

Germplasm Name	Pedigree	Origin	Year
Adroit	Norman/Mercia//Moulin;	GBR	1992
Andante	Moulin/D-172-6-4;	GBR:Eng	1992
Anglia	Marksman*Clement	UK	1987
Apollo	Maris-Beacon/Clement//Kronjuwel;Maris-Beacon/Kronjuwel;	DEU	1984
Apostle	Alcedo/Avalon//Moulin;	GBR:Eng	1980
Aristocrat	Rendezvous/Moulin//Mercia;	GBR	1992
Aslan	Brigand*Talent	UK	1982
Avocet	Tjb-268-175/(Sib)Hobbit;	GBR:Eng	
Beaufort	Rendezvous/Haven//Fresco;	GBR:Eng	1993
Bilbo	TJB 268 175 x Hobbit-sib	UK	1986
Bounty	Tjb-30-148/Tl-365-A-25;Maris-Ploughman/Durin;	GBR:Eng	1979
Brigand	(S)Maris-Brigand;	GBR:Eng	1977
Brimstone	Tjb-54-218/Hobbit-30-2//Hustler;	GBR:Eng	1985
Brock	Hobbit-30-2/Talent;Talent/(Sib)Hobbit;	GBR:Eng	1985
Brutus	Not Recorded	GBR	1997
Buster	Brimstone/Parade;	GBR	1992
Capitole	Cappelle//(S-6)Hybride-80-3/Etoile-De-Choisy;	FRA	1964
Cappelle Desprez	Vilmorin-27/Hybride-Du-Joncquois;	FRA	1946
Caxton	Moulin/Riband;	GBR	1996
Chablis	Jerico/Tonic;	GBR:Eng	1994
Claire	WASP X FLAME	UK	1999
Consort	Riband(Sib)/Fresco/Riband;	GBR:Eng	1993
Craftsman	Virtue/Maris-Huntsman;	GBR:Eng	1987
Daphne	Fan/Late-Gluyas;	AUS:SA	1912
Dean	Disponent/Norman;	GBR:Eng	1989

Table 3.5 Continued

Germplasm Name	Pedigree	Origin	Year
Desprez 80	Vilmorin-23/Institut-Agronomique;	FRA	1934
Durin	Vilmorin 29 x vogel 8058 2 x cappelle 4 x ci 12633 x 4X cappelle 2 x Heine 110 x cappelle 3 x nord	UK	1976
Dwarf A	Selected at F5 from the same cross as Hobbit	GBR	1989
Eureka	(Des Domes*Vilmorin 27)*(Hybride de Joncquois*Providence)	FRA	1973
Fenman	((Maris Ranger x Durin) x Maris Beacon) x Hobbit 'sib'	UK	1973
Flambeau	Talent/Norda(N-8-15-D-1)//Feuvert/3/Caton/Rivoli;	FRA	1992
Flanders	Champlein/Fd-2816-348;	FRA	1986
Galahad	Joss-Cambier/Durin//(Sib)Hobbit;Hobbit(Sib)//Durin/Joss-Cambier;	GBR:Eng	1983
Hadrian	CB 296*Maris Templar	GBR	1987
Haven	Hedgehog/Norman//Moulin;	GBR:Eng	1988
Heine 7	Hybrid a courte paille x Svalofs Kronen	DEU	
Herald	Tjb-54-335/(Sib)Hobbit;	GBR	1976
Hereward	Norman/Disponent;Disponent/Norman;Norman(Sib)/Disponent;	GBR:Eng	1989
Hobbit	Professeur-Marchal//Marne-Desprez/Vg-9144/4/Ci-12633/4*Cappelle-Desprez//Heines-110/Cappelle-Desprez/3/Nord-Desprez;	GBR	1974
Hunter	Apostle/Haven;	GBR:Eng	1991
Hustler	Maris Huntsman*Durin 25	UK	1974
Joss Cambier	Heines-Vii/Tadepi//Cappelle-Desprez;Cambier-194/Tadepi//Cappelle-Desprez;194-C.H.Vii/Tadepi//Cappelle;	FRA	1966
Leo	Kristall/Marksman;	GBR:Eng	1986
Longbow	Tjb-268-175/Hobbit;	GBR:Eng	1981
Maestro	(Selkirk*Cappelle Desprez)*Hobbit-sib	GBR:Eng	1976
Mantle	Cappelle/H-2596//6003;Tjb-30-148/Durin//Virtue/3/Marksman;	GBR:Eng	1974
Maris Argent	Maris-Freeman/Tl-365-A-25;	GBR:Eng	1976
Maris Beacon	Ci-12633/5*Cappelle-Desprez/3/Hybrid-46/Cappelle-Desprez//Professeur-Marchal;Cappelle*4/Hybrid-46//2*Professeur-Marchal;Ci-12633/5*Cappelle//Hybrid-46/Cappelle/3/2*Professeur-Marchal;	GBR:Eng	1968
Maris Envoy	Ci-12633/5*Cappelle-Desprez//Heines-110/Cappelle/3/Nordeste/4/Viking;	GBR:Eng	1974
Maris Freeman	Maris-Widgeon/Maris-Ranger;	GBR:Eng	1974

Table 3.5 Continued

Germplasm Name	Pedigree	Origin	Year
	Ci-12633/5*Cappelle-Desprez//Hybrid-46/Cappelle-Desprez/3/2*Professeur-Marchal;Cappelle*4/Hybrid-46//2*Professeur-Marchal;	GBR:Eng	1971
Maris Nimrod	Ci-12633/Yeoman//5*Cappelle/3/Cappelle/Hybrid-46/4/2*Professeur-Marchal;Cappelle*4/Hybrid-46//2*Professeur-Marchal;	GBR:Eng	1971
Maris Ploughman	Cappelle*4/Hybrid-46//2*Maris-Widgeon/3/Viking;Ci-12633/Hybrid-46//Maris-Widgeon;	GBR:Eng	1972
Maris Ranger	Heines-Peko/Cappelle-Desprez;	GBR:Eng	1968
Maris Settler	Professeur-Marchal/Cappelle;	GBR:Eng	1972
Maris Teal	Hybrid-46/Minister;Hybrid-46/Ministre;	GBR:Eng	1972
Maris Templar	Ci-12633/5*Cappelle-Desprez//Heines-110/Cappelle-Desprez/3/Nord-Desprez/4/Viking;Heines-110/Cappelle-Desprez//Ci-12633/5*Cappelle-Desprez/3/Nord-Desprez/4/Viking;	GBR:Eng	1968
Maris Totem	Hybrid-46/Minister//Nord-Desprez;	GBR:Eng	1969
Maris Widgeon	Holdfast/Cappelle-Desprez;	GBR:Eng	1964
Marksman	Maris Huntsman*TL 365a/25(=Durin)	GBR	1977
Marne Desprez	Hybride-Du-Joncquois/Vilmorin-27//Hybride-Du-Joncquois/P.L.M.1;	FRA	1954
Mercia	Talent/Virtue//Flanders;	GBR:Eng	1984
Ministre	Benoist 40*Professeur Delos	BEL	
Mithras	Tjb-268-175/(Sib)Hobbit;	GBR:Eng	1980
Morell	Not Recorded	GBR	1997
Moulin	Yecora-70/Ciano-67(Cb-306-Y-70)//Maris-Widgeon/3/Maris-Hobbit;	GBR,FRA	1985
Norman	Tjb-268-175/(Sib)Hobbit;	GBR:Eng	1981
Norseman	Not Recorded	GBR	1988
Nsl92-5719			
Orqual	Capitole/Moisson//Horace/3/Thesee;	FRA	1991
Ostara	Not Recorded	GBR	1997
Pastiche	Jena/Norman;Iena/Norman;	GBR:Eng	1988
Professeur Marchal	Rimpaus-Bastard-Ii/Professeur-Delos//Professeur-Delos/Hybride-Du-Joncquois;Hybride-Du-Joncquois/Proffesseur-Delos//Bastard-Ii/Professeur-Delos;	BEL	1957
Renard	Hobbit-30-2/Talent;	GBR:Eng	1987

Table 3.5 Continued

Germplasm Name	Pedigree	Origin	Year
Rendezvous	Vpm-1/(Sib)Hobbit//Virtue;	GBR:Eng	1985
Ritmo	Hobbit/Line-1320//Wizard/3/Marksman/Virtue;Hobbit//Line-1320/Wizard/3/Marksman/4/Virtue;	NLD	1990
Samson	Sinew/Improved-Fife;	AUS:NSW	1899
Sarsen	Marksman/Armada;	GBR:Eng	1987
Shango	Fresco/Tiresius;	GBR	1994
Shire	Ploughman/Tl-36.30.6;	GBR	1979
Sickle	Brigand*(Maris Huntsman*TW 161)	GBR	1985
Sponsor		FRA	1994
Sportsman	Maris Envoy*Durin	GBR	1976
Steadfast	Little-Joss/Victor;	GBR:Eng	1952
Tadorna	(((Chinese 166*Panzer 3)*Heine 7)*(Teutonen*Hindukush 516)*Heine 7)*Me	DEU	1982
Texel	Capitole/Moisson//Horace/3/B-8111;	FRA	1992
Tipstaff	Maris-Ranger/Durin//Maris-Huntsman;	GBR	1976
Tsengrain			
Veritas	Not Recorded	GBR	1997
Villein	Maris-Ploughman/Hobbit-Sib-1;	GBR:Eng	1978
Virtue	Maris-Huntsman/Maris-Durin;	GBR:Eng	1979
Vivant	Boxer/Gawain;	GBR:Eng	1991

have required, for example, greater values of K tested against greater run lengths. However, this would require a significant increase in computing power, which was not available at the time this analysis was carried out. However, one simulation study performed by Rosenberg et al. (2001) to determine the success rate of the *Structure* algorithm as a function of the number of markers and the number of individuals, demonstrates that this is unlikely to be the cause. Rosenberg et al. (2001) determined that a minimum number of 12 to 15 markers were required to genotype 15 to 20 individuals in each of the proposed populations in order to successfully compartmentalise individuals into their obvious sub-group.

Additionally, a study on the bird species *Turdus halleri*, Taita thrush, by Pritchard and Przeworski (2001) saw the sampling of birds from geographically distinct populations, followed by genotypic analysis with seven SSR markers. This multi-locus genotype data was examined in *Structure* and nearly every individual was clustered within their obvious group. The exception was with four individuals, two who appeared to be migrants and two that were perhaps hybrids. Since this wheat experiment utilised a greater number of both genetic markers and individuals, and given the reliability of the algorithm to sort individuals into distinct sub-groups, as in the Taita thrush study, it is realistic to assume that if there were distinct sub-groups in these two larger populations, they would have been detected.

The decision to examine the K = 3 results from *Structure* was made in order to account for at least some of the population structure within each of the larger data sets. Accounting for some of the existing sub-structure will be important in attempting to determine the affect of this phenomenon on type-1 error rates in subsequent linkage disequilibrium studies. These *Structure* results have provided an initial division of much larger data sets into smaller ones which were then compared

against known pedigree data. As mentioned in Section 3.3.3 the pedigree data was consistent with the outcomes of the clustering algorithm analysis. In a genetic diversity study, Paull et al. (1998) examined RFLP data from 124 Australian wheat lines. The genetic distance analysis of these lines resulted in their placement into one of four groups. Comparing the clustering results observed by Paull et al. (1998) to those obtained from the *Structure* analysis of the 90 Australian wheat lines in this study, showed consistency between the studies. Of the 22 wheat lines making up the sub-population in this experiment, five were not included in the RFLP study, 14 grouped within the first two clusters of the dendrogram and the remaining three grouped within Group 3 of the RFLP study. These results confirm that *Structure* can be used as a tool to initially divide a larger population into small groups exhibiting less sub-structure. However existing pedigree data should not be overlooked.

The lines in the remaining two clusters from the K=3 results were divided in the same manner as the largest sub-group mentioned above. In the first sub-population a mere 25% of these lines originates from Australia with a larger portion of the lines having Mexican origins. Despite this observation, the second sub-population approximates the characteristics of the former sub-population since 80% of lines have Australian origins. However, in contrast to the third and largest sub-population identified earlier there is an equal distribution of the lines' origins between Queensland, Victoria, and New South Wales. Additional *Structure* runs result in the further division of the sub-populations although lines that originally grouped together have a tendency to stay together as opposed to combining with lines from other groups.

As discussed in this chapter, the application of genetic cluster analysis to wheat data sets is not entirely straightforward. In other studies where distinct genetic

clusters are available this approach has proven valuable (Pritchard et al. 2000; Rosenberg et al. 2001; Rosenberg et al. 2002). However in wheat breeding populations the intercrossing of lines from different groups has resulted in this lack of specific populations. Consequently, it is proposed in this study that cluster analysis, such as that implemented in the *Structure* software, be used as a primary tool for establishing smaller sub-populations based purely on genetic information. Subsequently, an evaluation of the lines in the sub-groups together with their pedigree information would result in the confident grouping of individuals into sub-populations with substantially reduced population structure.

Chapter 4: Genome Wide Linkage Disequilibrium in Hexaploid Wheat

4.1 Introduction

Over the past 20 to 30 years molecular markers have assisted in the identification and localisation of a number of economically important QTL and major gene loci in hexaploid wheat (Charmet et al. 2001; Kato et al. 2000; Prasad et al. 2003). QTL analysis has provided geneticists a starting point from which to target specific genes underlying a particular trait. Association genetics has been particularly effective in biomedical research and has led to a number of disease genes being identified and localized (Klein et al. 2005; Maraganore et al. 2005; Ozaki et al. 2002).

The premise of this mapping approach, as outlined in previous chapters, is the non-random inheritance of alleles at particular loci. One advantage of LD, or association mapping, comes as a result of studying inheritance of loci within relatively large natural populations, thus allowing geneticists to take advantage of historical recombination events that have occurred in the establishment and shaping of particular populations. This is particularly evident when compared to traditional linkage mapping methods, where familial inheritance of loci is studied and precision of gene localisation is restricted by the limited number of cross-over events in the establishment of the population. Since there are relatively few opportunities for recombination to occur in traditional QTL mapping populations, the segments of DNA harbouring a potential gene of interest may be several cM in some instances (Cuthbert et al. 2006; Roder et al. 2007; Ronin et al. 2003).

However in order for LD mapping to be successful there are several factors that need to be considered, in particular it is crucial that marker density be sufficient enough to capture the levels of LD (Clark et al. 2007; Wang et al. 2005). In human

genetics there has been considerable debate over what is the ideal marker density. Through computer simulations, Kruglyak (1999) estimated that useful LD would extend over 3 Kb, resulting in a required minimum marker density of 500,000 SNPs for whole genome analysis. There has since been significant empirical data suggesting that this initial estimate is not correct and that there are several factors that will affect these results, including the marker type used, genomic location studied and the population within which LD is measured (Abecasis et al. 2001; Reich et al. 2001; Service et al. 2001). Although there is substantial information surrounding LD in humans there is significantly less information on the extent and distribution of LD in plant systems.

As a consequence of this knowledge gap plant geneticists have started to measure LD. However, most of the research published to date is in diploid species such as maize (Remington et al. 2001; Thornsberry et al. 2001) and *Arabidopsis* (Hagenblad and Nordborg 2002), with limited information on the patterns and extent of LD in economically important crops such as wheat and other polyploids. The aim of this study was to examine the patterns and assess the extent of LD within the complex polyploid genome of wheat. Two populations from different parts of the world were assessed for these characteristics with results suggesting that LD extends beyond tens of centimorgans along each chromosome and that there is evidence of significant association between non-syntenic markers, or those located on different chromosomes. Results reported in this Chapter are encouraging with respect to the application of association mapping in hexaploid wheat.

4.2 Materials and Methods

4.2.1 Plant material and DNA isolation

Germplasm outlined in Section 2.2.1 was also used in this Chapter with the isolation of DNA described in Section 2.2.2

4.2.2 Microsatellite Analysis

Microsatellite markers used in this Chapter and the analysis is described in Sections 3.2.2 to 3.2.5.

4.2.3 Linkage Disequilibrium Data Analysis

Using un-phased diploid genotype data from the 96 wheat lines in this study, LD between marker pairs was calculated using the two-locus LD method as implemented in the software package PowerMarker (Liu and Muse 2004). LD is calculated as a weighted mean of the total allele-pair LD and is defined as follows:

$$LD = \sum_{u}\sum_{v} p_u p_v \left| LD_{uv} \right|,$$

where LD is Lewontin's D' (Lewontin 1964) defined as follows:

$$D' = \frac{D_{uv}}{D_{\max}},$$

with LD_{uv}:

$$D_{uv} = x_u - p_u q_v,$$

where x_u is the observed frequency of the gametes and $p_u q_v$ are the frequencies of the alleles and:

$$D_{\max} = \begin{bmatrix} \min[p_u q_v, (1-p_u)(1-q_v)]; D_{ij} < 0 \\ \min[p_u (1-q_v), (1-p_u) q_i]; D_{ij} > 0 \end{bmatrix}$$

Of the two most commonly used statistical methods to estimate LD (D' and r^2), D' was used in this study. D' allows for the examination of highly polymorphic markers, such as microsatellites, and although it is still affected by differences in allele frequencies it is decidedly less so than other measures such as r^2 (Delvin and Risch 1995). Examining the results in this way allows for a direct comparison with results obtained by (Maccaferri et al. 2005) and their recent evaluation of LD in the tetraploid durum wheat species. However, D' is strongly affected by small population size and can result in an upwards bias of total LD (Mohlke et al. 2001). Additionally, D' measures only recombination fractions whereas r^2 measures recombination and mutation history; as well it is indicative of how the markers and QTL of interest might be correlated (Gupta et al. 2005a). As such Gupta et al. (2005a) point out that the use of one measure over the other (D' vs. r^2), is largely dependant on the objective of the study. Since the objective of this genome wide study in hexaploid wheat using SSR markers was to examine the levels of LD within the genome and not to assess associations with QTL D' was chosen as the measure of LD.

Exact significance values (α) of the data were estimated by a Markov Chain Monte-Carlo approach. The Bonferroni correction, as explained previously, was not applied to each individual α value since this would result in a loss of nearly all statistical power to detect non-random associations between loci.

4.3 Results

4.3.1 Extensive LD within South Australian hexaploid wheat germplasm

Using 124 SSR markers, 96 Australian hexaploid wheat lines were examined in order to understand the patterns and extent of LD. This is equivalent to 439 linked pair-wise comparisons (or comparisons between markers located on the same chromosome) and 7,187 non-syntenic pair-wise comparisons. However, one marker comparison had many missing data points and was excluded from the study. The average number of alleles per SSR marker was 9.90 and ranged between two and 23 alleles.

LD between each marker pair, both syntenic and non-syntenic was calculated using the normalised D' value (Lewontin 1964). A summary of the syntenic and non-syntenic pairwise marker comparisons used in this experiment are summarised in Table 4.1. The marker pairs were separated into five main 'classes' (Maccaferri et al. 2005) based on the intermarker distances obtained from the consensus map from Somers et al. (2004). These marker 'classes' represent tightly linked (0 to 5 cM), linked (5 to 20 cM), loosely linked (20 to 50 cM), unlinked (>50 cM), and non-syntenic markers. The total number of SSR markers used in this experiment is also listed.

In three cases there is a reduction in the number of marker loci used since the pooling of rare alleles into a common class as well as the reduction of lines in the formation of the sub-population resulted in monomorphic scores. Figure 4.1 illustrates D' values as a function of map distance (cM) for all syntenic marker comparisons. It is clear from this scatter plot that LD does indeed decay with distance, *albeit* slightly from a mean D' level of approximately 0.56 to approximately

TABLE 4.1 Number of pair-wise comparisons performed in the LD analysis of the complete Australian wheat dataset (96 lines) as well as the largest Australian sub-population as determined from STRUCTURE in Section 3.3.3 (22 lines). Distinction is also made between pairwise comparisons using all of the markers with all alleles as well as with rare alleles (frequency <0.05) pooled. Marker 'classes' were arbitrarily chosen with the marker distances obtained from Somers et al. (2004).

	Number of SSR markers used	Number of pairwise marker comparisons in each 'class' of syntenic markers					Total pairwise comparisons
		0 to 5 cM	5 to 20 Cm	20 to 50cM	> 50 cM	Non-syntenic	
Entire Australian dataset	124	29	92	132	185	7187	7625
Entire Australian dataset with pooled rare alleles	123	27	91	131	185	7069	7503
Australian Sub-population	122	27	90	128	185	6950	7380
Australian Sub-population with pooled rare alleles	122	27	90	128	185	6950	7380

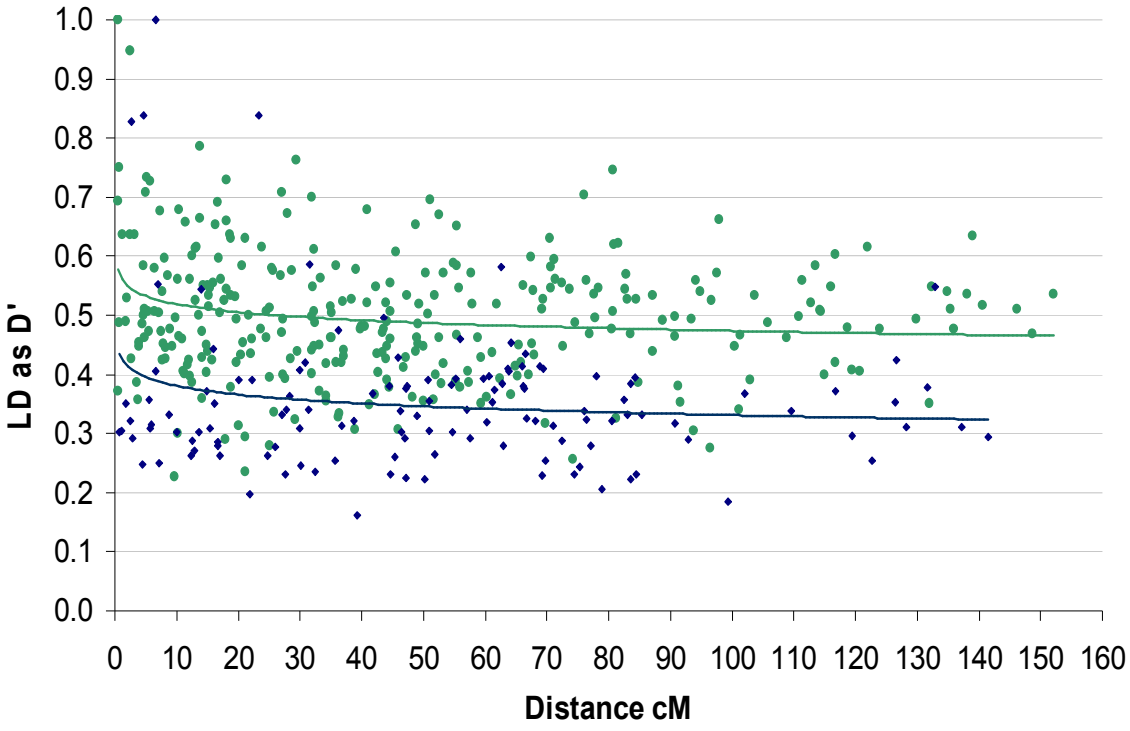


FIGURE 4.1 Scatter plot of LD (D') vs marker distance (cM) for all Australia wheat lines. Green data points are significant (P <0.0001) pairwise marker comparisons and the green trend line provides visualisation of the decay of LD with distance. The blue datapoints and trendline are non-significant pairwise marker comparisons.

0.45 over 50 cM. When D' is averaged between closely linked (<5 cM) as well as unlinked (>50 cM) markers the magnitude of this measure declined slightly from 0.5268 to 0.4400 respectively. The correlation coefficients of D' values with distance in cM were also calculated. On a genome wide level there is a significant negative correlation between LD levels and marker distance with a correlation coefficient (r) of -0.1132 at a significance level of P <0.025. This correlation was even more pronounced when only those markers separated by less than 50 cM were examined (r = -0.2173, P < 0.0005).

Figure 4.2 illustrates the distribution of the D' values in each of the five marker 'classes' and Figure 4.3 illustrates the distribution of percentage of pair-wise comparisons that are significant at the P < 0.0001 level. Examining the D' values estimated in the entire Australian population there is a clear and steady decline in the magnitude of the LD statistic across the five marker 'classes'. There is also a high proportion of pair-wise comparisons that are significant at the P < 0.0001 level when examining the population as a whole. For unlinked markers (separated by >50 cM) and non-syntenic markers, the estimated D' values remain high, averaging 0.4400 and 0.4205 respectively. Furthermore, a high percentage of significant pair-wise marker comparisons remains in these later two marker 'classes' (65% and 64% respectively) indicating high levels of LD along chromosomes as well as between chromosomes.

4.3.2 LD in the absence of population stratification

It has been well documented that in the presence of population structure, the levels of LD will be elevated and spurious allele associations will be detected. Due to the high level of significant LD in the results presented above, between both syntenic and non-syntenic marker pairs, it was clear that this could potentially be due to the effects of population structure and rare allele frequencies. In this study, the impact of

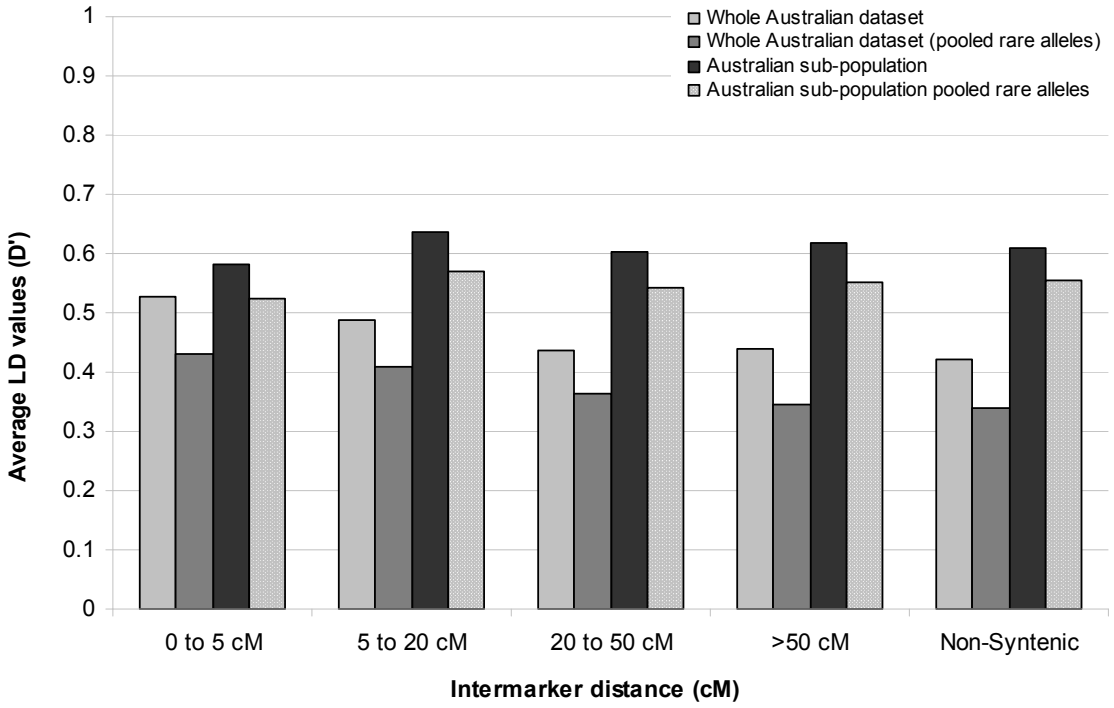


FIGURE 4.2 Bar graph illustrating the distribution of average D' values for each of the four experimental scenarios (outlined in the legend) within each of the five marker 'classes' (on the x-axis).

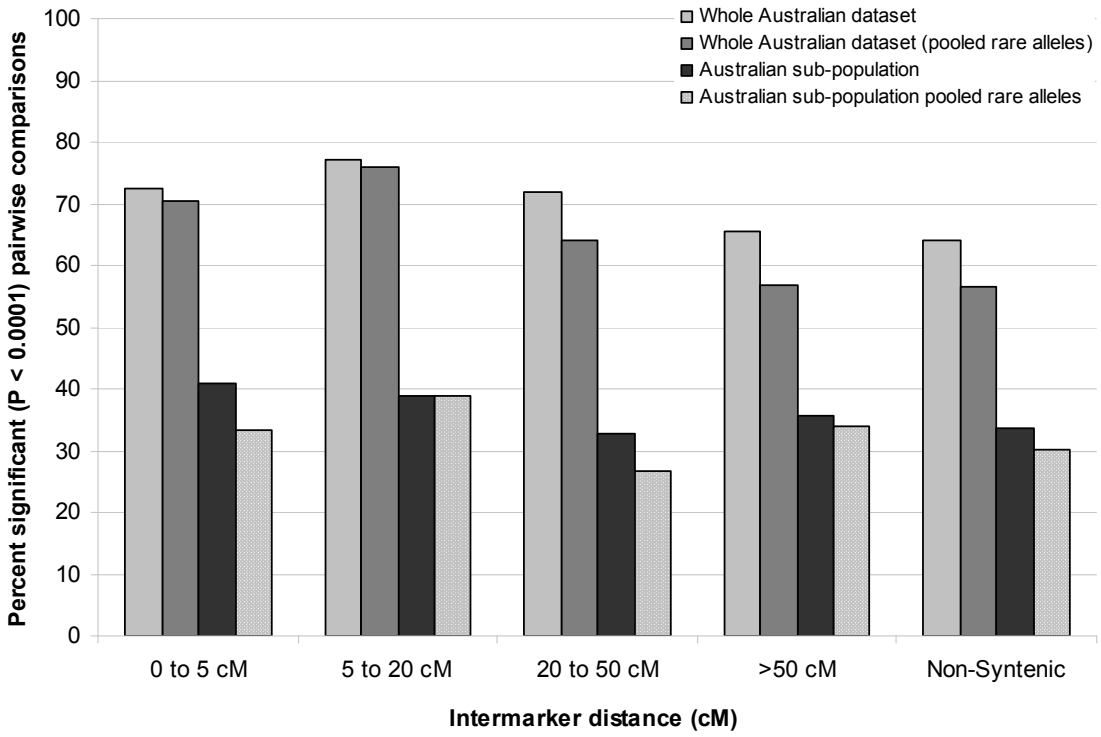


FIGURE 4.3 Bar graph illustrating the distribution of percent significant pairwise comparisons for each of the four experimental scenarios (outlined in the legend) within each of the five marker 'classes' (on the x-axis).

population structure on the overall values of D' was accounted for by the Structure analysis carried out in Chapter 3. Previous reports have found this to be an important factor where it has affected LD measurments (Remington et al. 2001).

Using the clustering program *Structure* (Pritchard et al. 2000) together with known pedigree data, the 96 Australian wheat accessions were sub-divided into several sub-groups as summarised in Chapter 3. Examination of the largest sub-population within the Australian germplasm, consisting of 22 individuals, revealed a higher D' genome average than with all lines together. In fact, the mean D' for all pairwise marker comparisons was 0.6232 for markers separated by less than 20 cM of which 39% of the pairwise comparisons were significant at the P <0.0001 level, and decayed only slightly to 0.6175 with markers separated by >50 cM, with 36% of the pairwise comparisons significant at the P < 0.0001 level. The green trendline superimposed on Figure 4.4 illustrates the tendency for LD to decrease with map distance from D' = ~0.78 to 0.7 over nearly 150 cM for significant pairwise comparisons (P <0.0001). This is in contrast to the non-significant pairwise comparisons that where there is a marginal tendency for LD to increase with distance. This increase in D' is most likely attributed to the small number of individuals (22) in this part of the study, which affects the D' calculations of LD in a positive way. The D' estimate for unlinked syntenic marker pairs was similar, yet slightly higher (D' = 0.6175) than that detected between non-syntenic marker comparisons in the absence of population structure where D' averaged 0.6094.

The decay of LD over map distance in this experiment reveals that in both cases, with and without population structure, there is a clear decrease in LD with map distance amongst significant pairwise marker comparisons. The magnitude of the D' values for significant pairwise comparisons increases in the absence of population

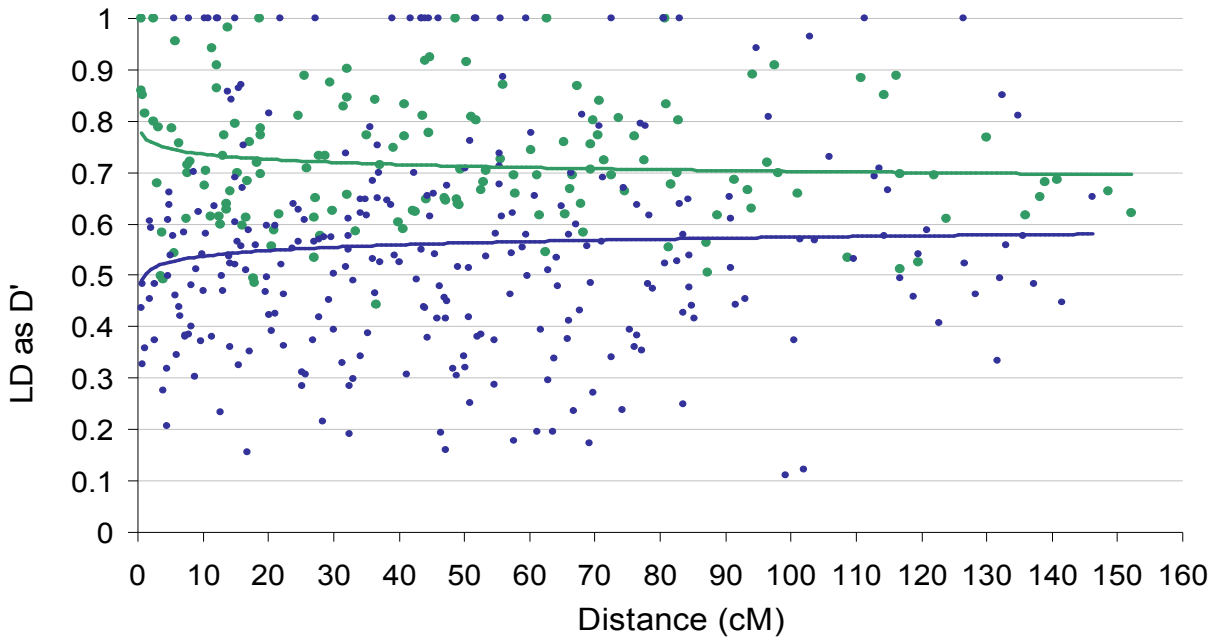


FIGURE 4.4 Scatter plot of LD (D') vs marker distance (cM) for the largest Australian sub-population (22 lines) as determined through Structure analysis. The green trend line provides visualisation of the decay of LD with map distance in the significant pair-wise marker comparisons (green data points). This is in contrast to the lack of LD decay with map distance amongst the non-significant pair-wise marker comparisons represented by the blue trend line and data points.

structure and is likely due to the small sample size. However the percentage of significant pair-wise comparisons has decreased by nearly half in each of the marker 'classes' as illustrated in Figure 4.3.

Compared to the 70.32% of all pair-wise comparisons significant at the P <0.0001 level for the 96 samples, there is a mere 35.81% when only one un-stratified population was examined. Of all markers separated by 5 to 20 cM, 38.89% of these pair-wise comparisons were significant at the P <0.0001 level, as were 32.81% of comparisons involving markers separated by 20 to 50 cM. For unlinked markers on the same chromosome, 35.68% were significant at the same level, in contrast to the 65.41% observed when all 96 lines were analysed as a whole. These results show that accounting for population structure has a positive affect on the magnitude of D' in this population but also a negative effect on the overall number of pair-wise comparisons that are significant at the P <0.0001 level. It is likely that the resulting small population size is contributing to the former observation.

4.3.3 LD in the absence of rare alleles

The empirical data from Section 4.3.2 suggests that there is an unexpectedly high rate of non-random assortment within this population. Furthermore, a large portion of this extensive LD appears to be attributed to population stratification. Another factor that is worth taking into account is the presence of rare or relatively new alleles in a sample. New alleles or mutations will have a positive affect on the extent to which LD is observed in a population as only a limited number of recombination events could have occurred to rendomise associations with other closely adjacent markers. With this in mind, D' values were re-evaluated for the whole population as well as for the sub-population after having chosen a cut off point for the rare allele frequency of <5%. This means that for every marker used, the frequency of each allele was

calculated and those having a less than 5% occurrence in the data set were grouped together and assigned a new allele name, which is the name of the most frequent rare allele.

As illustrated in Figure 4.2, markers on the same chromosome separated by 5 to 20 cM had a mean D' of 0.4102 when the 96 lines were examined as a whole in the absence of rare alleles. For those markers separated by more than 50 cM the mean D' decreased slightly to 0.3448. Amongst the non-syntenic marker pairs the mean D' value was slightly less at 0.3399.

When the frequency of significant D' values were examined, 63.82% of all syntenic pair-wise comparisons were significant at the $P < 0.0001$ level and only 56.42% of all non-syntenic marker-pairs were significant. Furthermore, 75.82% of syntenic marker pairs separated by 5 to 20 cM were significant at the $P < 0.0001$ level, as were 64.12% of the pair-wise comparisons between markers separated by 20 to 50 cM. An additional 56.76% of those pair-wise comparisons with markers separated by more than 50 cM were significant at the $P < 0.0001$ level.

When compared to the initial estimates of D' from the 96 lines studied, there is an overall decrease in the number of significant pair-wise comparisons of roughly 6.5 in the calculations involving all lines with pooled rare alleles (Figure 4.3). Specifically there were 8.7% fewer significant estimates of LD (D') for syntenic markers separated by more than 50 cM, which is a similar difference to that of the non-syntenic marker pairs where there was an observed decrease of 7.6%.

When estimating LD (D') in the absence of rare alleles within the sub-population of 22 individuals, the mean genome wide value increased to 0.5504 which is nearly 1.5 times greater than that for the population examined as a whole in the

absence of rare alleles. For markers separated by 5 to 20 cM, the D' value rose slightly to 0.5692 but declined to 0.5414 for markers separated by 20 to 50 cM (Figure 4.2). When considering only significant pairwise marker comparisons (P <0.0001), there is a tendancy for genome wide LD to decrease with map distance; this is illustrated by the green trend line in Figure 4.5.

Nearly 32% of all the pair-wise comparisons were significant at the $P < 0.0001$ level which is slightly less than the 36.07% that were significant when analysing the sub-population in the presence of rare alleles. In the non-syntenic marker pair evaluation, the mean D' values were higher than the syntenic marker pair analysis, with the average genome wide D' value of 0.5544 (30.19% significant at a level of P < 0.0001).

4.3.4 Examination of LD within linkage groups and genomes

Through examining the levels and patterns of genome wide LD, valuable information can be acquired. However it might be more appropriate to examine this phenomenon at specific regions of the genome since it is likely to vary from region to region, due to recombination hotspots and/or to the physical position on a chromosome. With this in mind LD within groups and genomes were examined.

4.3.4.1 LD within linkage groups

Initial analysis focused on each genomic group in the 96 Australian wheat lines, which provided detail of the distribution of LD decay with distance. Some groups declined much more rapidly than others (Figure 4.6). Chromosome Groups 4 and 5 demonstrate a significant negative correlation between their D' values and the distance between each marker as illustrated in Figure 4.6D and E, respectively. The observed r value, for those pair-wise marker comparisons specific to Group 4

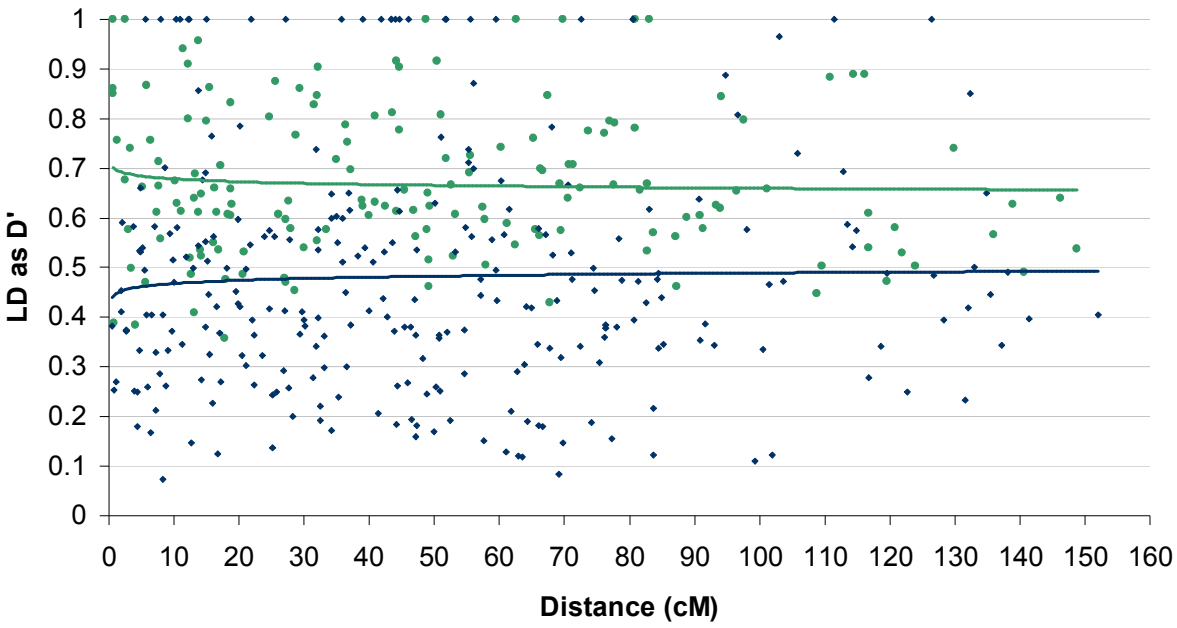


FIGURE 4.5 Scatter plot of LD (D') vs marker distance (cM) for Australian wheat lines in the absence of population structure and rare alleles pooled. The green trend line provides visualisation of the decay of LD with map distance in the significant pair-wise marker comparisons (green data points). Blue datapoints are non-significant pairwise marker comparisons with the blue trendline illustrating the lack of reduction in LD with map distance.

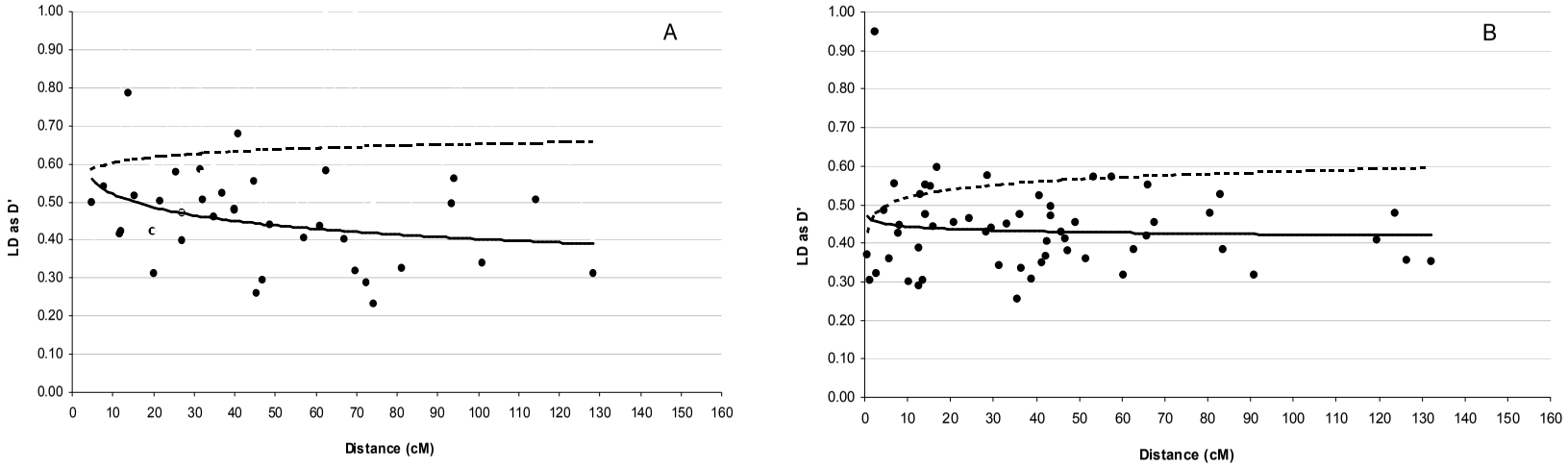


FIGURE 4.6 Scatter plot of LD (D') vs marker distance (cM) for groups 1 (A), 2 (B), 3 (C), 4 (D), 5 (E), 6 (F), and 7 (G) of the whole Australian wheat data set. The solid trend line provides visualisation of the decay of LD with distance across the whole dataset. The hashed trend line in each plot illustrates the decay (or lack there of) in LD with distance when population structure is accounted for and rare alleles pooled.

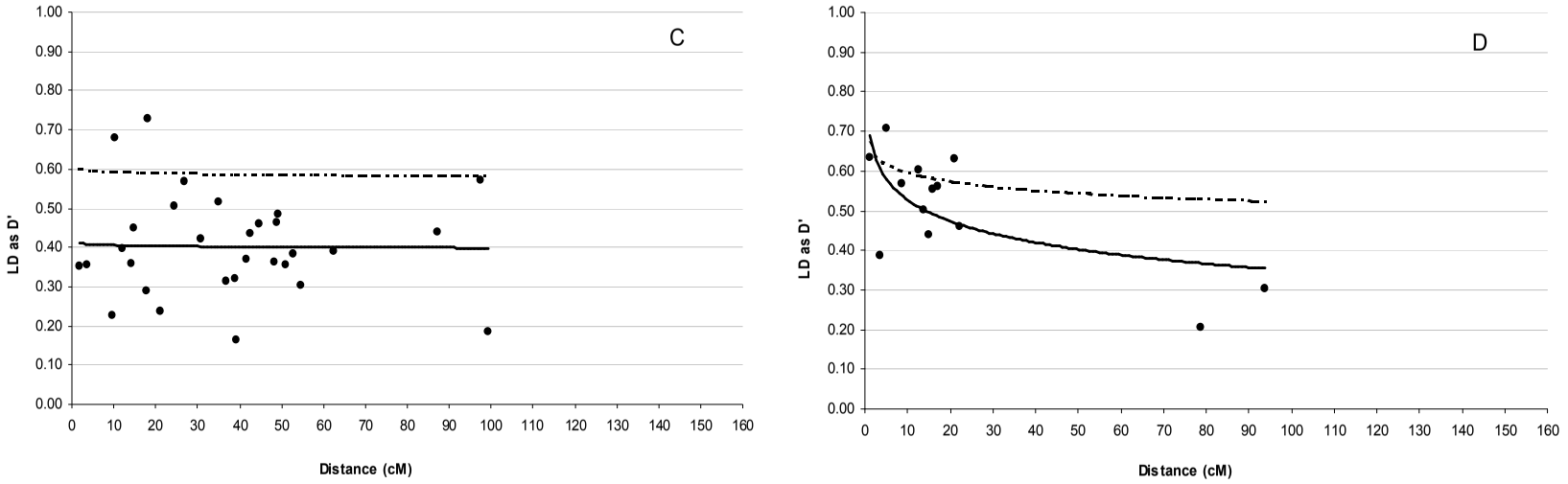


FIGURE 4.6 (Continued) Scatter plot of LD (D') vs marker distance (cM) for groups 1 (A), 2 (B), 3 (C), 4 (D), 5 (E), 6 (F), and 7 (G) of the whole Australian wheat data set. The solid trend line provides visualisation of the decay of LD with distance across the whole dataset. The hashed trend line in each plot illustrates the decay (or lack there of) in LD with distance when population structure is accounted for and rare alleles pooled.

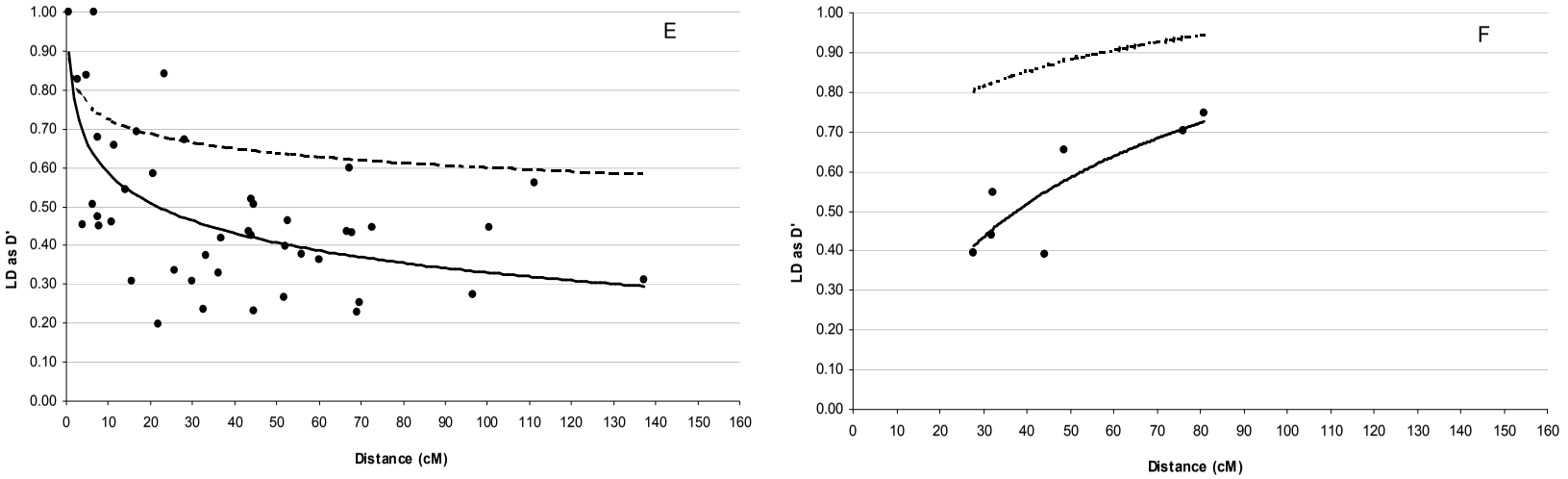


FIGURE 4.6 (Continued) Scatter plot of LD (D') vs marker distance (cM) for groups 1 (A), 2 (B), 3 (C), 4 (D), 5 (E), 6 (F), and 7 (G) of the whole Australian wheat data set. The solid trend line provides visualisation of the decay of LD with distance across the whole dataset. The hashed trend line in each plot illustrates the decay (or lack there of) in LD with distance when population structure is accounted for and rare alleles pooled.

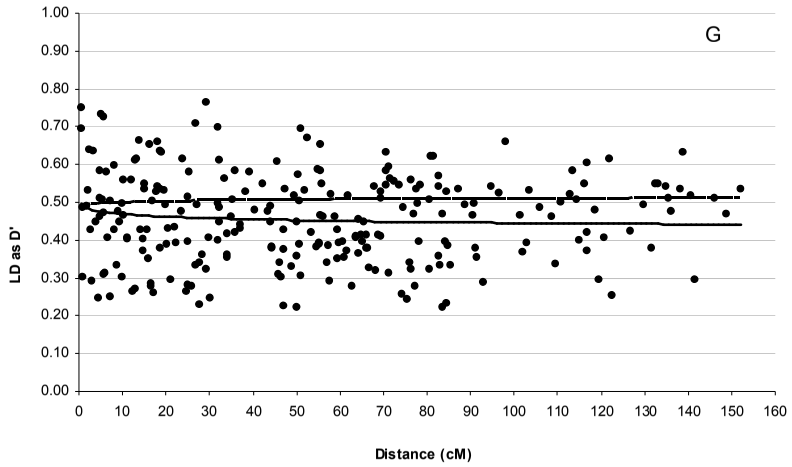


FIGURE 4.6 (Continued) Scatter plot of LD (D') vs marker distance (cM) for groups 1 (A), 2 (B), 3 (C), 4 (D), 5 (E), 6 (F), and 7 (G) of the whole Australian wheat data set. The solid trend line provides visualisation of the decay of LD with distance across the whole dataset. The hashed trend line in each plot illustrates the decay (or lack there of) in LD with distance when population structure is accounted for and rare alleles pooled.

chromosomes (r = -0.7617; *P* 0.0005 > 0.00025), was over 6.7-fold more negative than that observed in the genome wide average, despite the number of pair-wise comparisons in this group being amongst the lowest of all groups with a total of 13. The group with the highest number of pair-wise comparisons, with the exception of group 7, was the Group 2 chromosomes with 56 pair-wise comparisons. As with nearly all of the chromosome groups there was a negative correlation of D' over distance with an r value of -0.1201, equivalent to a 1-fold increase which is not significant (Figure 4.6B).

When taking population structure and rare alleles into account, the trend of decay with distance varies greatly between chromosome groups. Hashed trend lines in Figure 4.6 represent the trend of LD decay, (or lack thereof) with distance for chromosome Groups 1 to 7 when taking into account the genetic and evolutionary factors mentioned above. The most notable comparison is the overall increase of the D' values by 0.2 in most cases within the groups. There is also a lack of significant correlation between the decay of LD with distance. Where LD levels in chromosome Groups 4 and 5 were once significantly negatively correlated with distance there remains a slight negative correlation. However it is no longer significant with *P* values > 0.1 in both instances.

4.3.4.2 LD within genomes

Within each of the three wheat genomes a trend emerges that is comparable to what is observed when examining each of the chromosome groups. The scatter plot in Figure 4.7 reveals that when LD is estimated for all 96 Australian wheat lines, the B genome has the largest average D' value of the three (D' = 0.4843). The B genome also has the largest number of pair-wise comparisons that are significant at the *P* <0.0001 level (76.17%). There is a negative correlation of LD with distance across the B

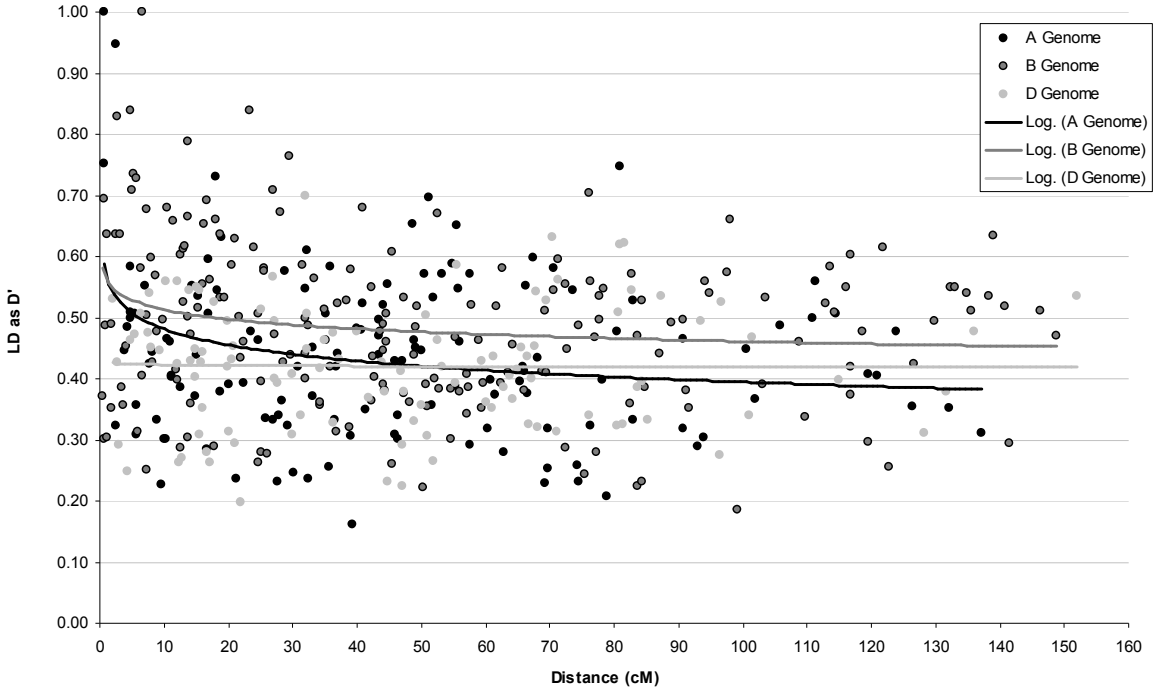


FIGURE 4.7 Scatter plot of LD (D') vs marker distance (cM) for genomes A, B, and D using the whole Australian wheat data set. The solid trend line provides visualisation of the decay of LD with distance across the whole dataset. The hashed trend line in each plot illustrates the decay (or lack there of) in LD with distance when population structure is accounted for and rare alleles pooled.

genome of r = -0.1758. When rare alleles are taken into account the B genome still has the highest D' value (D' = 0.3881). However the magnitude of this measure has decreased approximately 1.2-fold in each of the three genomes. In addition there is a reduction in the number of significant (P <0.0001) pair-wise comparisons across each of the three genomes. As with the other analyses, the magnitude of the D' values increase by approximately 0.2 when population structure is taken into account, while the percentage of significant pair-wise comparisons decreases (Figure 4.8).

4.3.4.3 LD across the Group 7 chromosomes

The Group 7 chromosomes were covered with the greatest number of molecular markers in an effort to determine appropriate marker spacing in hexaploid wheat that would facilitate the detection of meaningful levels of LD. A total of 40 SSR markers specific to Group 7 were included in this study. The average distance between markers was 10 cM with four, six, and five markers that lay within 5 cM of each other on chromosomes 7A, 7B, and 7D respectively. There were a total of 252 pair-wise comparisons within this group resulting in an average D' value of 0.4558. While the correlation coefficient of this group was slightly negative with an r value of -0.0138, it was not significant. Of all the Group 7 chromosomes the pair-wise comparisons specific to chromosome 7B had the highest negative correlation between LD and distance, with an r value of -0.2701. Group 7A had an r value of -0.0944 and group 7D appears to have no correlation (r = 0.1619). Figure 4.9 illustrates LD decay with distance for each of the Group 7 chromosomes.

Figures 4.10 and 4.11 illustrate the distribution of average D' values and significant P values for four marker 'classes' in each of the four experimental scenarios. In the absence of rare alleles the overall magnitude of the D' values decreased by approximately 0.1 in each case. The number of significant pair-wise

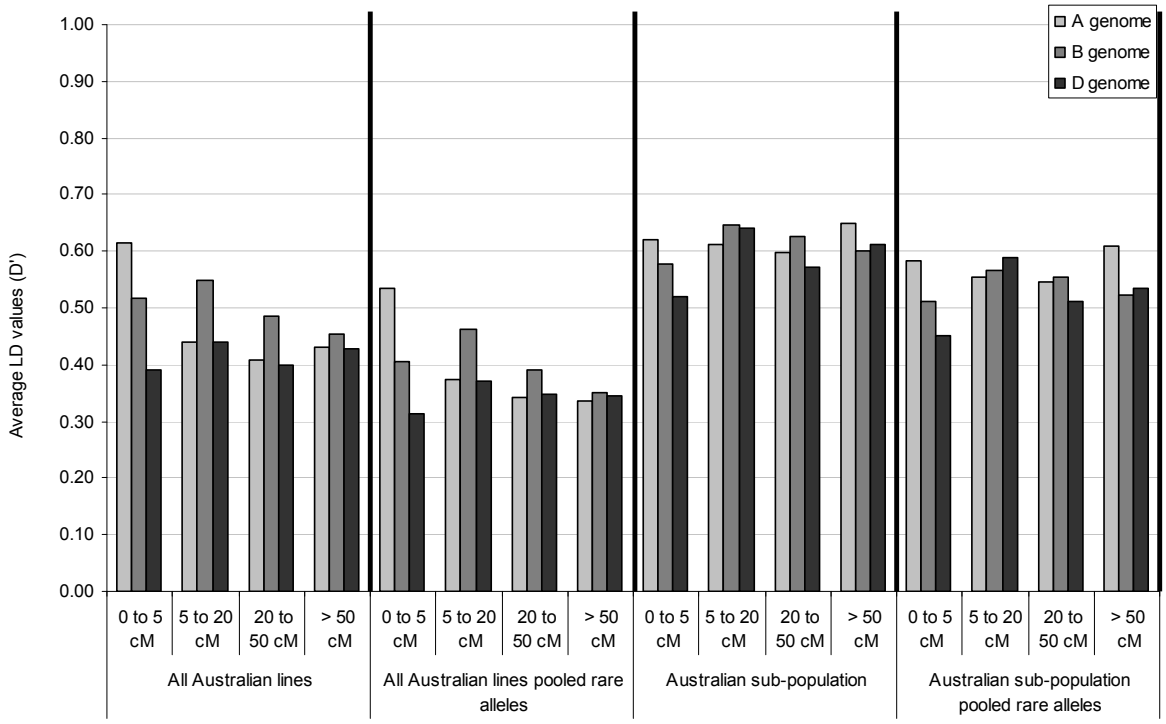


FIGURE 4.8 Bar graph illustrating the distribution of average D' values of each genome (A, B, and D). Results are presented for markers in each of the four marker 'classes' within each experimental scenario (named along the X axis).

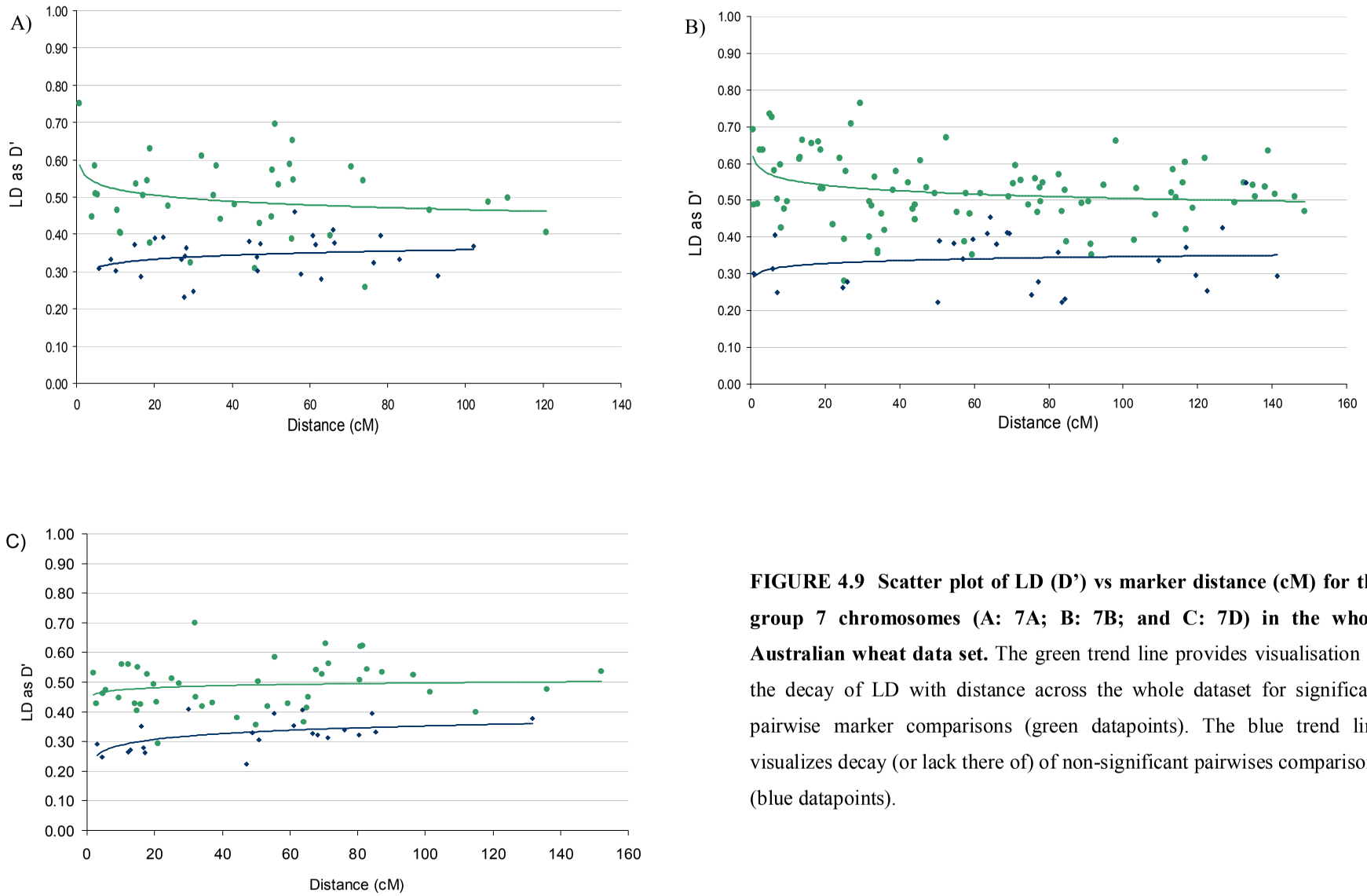


FIGURE 4.9 Scatter plot of LD (D') vs marker distance (cM) for the group 7 chromosomes (A: 7A; B: 7B; and C: 7D) in the whole Australian wheat data set. The green trend line provides visualisation of the decay of LD with distance across the whole dataset for significant pairwise marker comparisons (green datapoints). The blue trend line visualizes decay (or lack there of) of non-significant pairwises comparisons (blue datapoints).

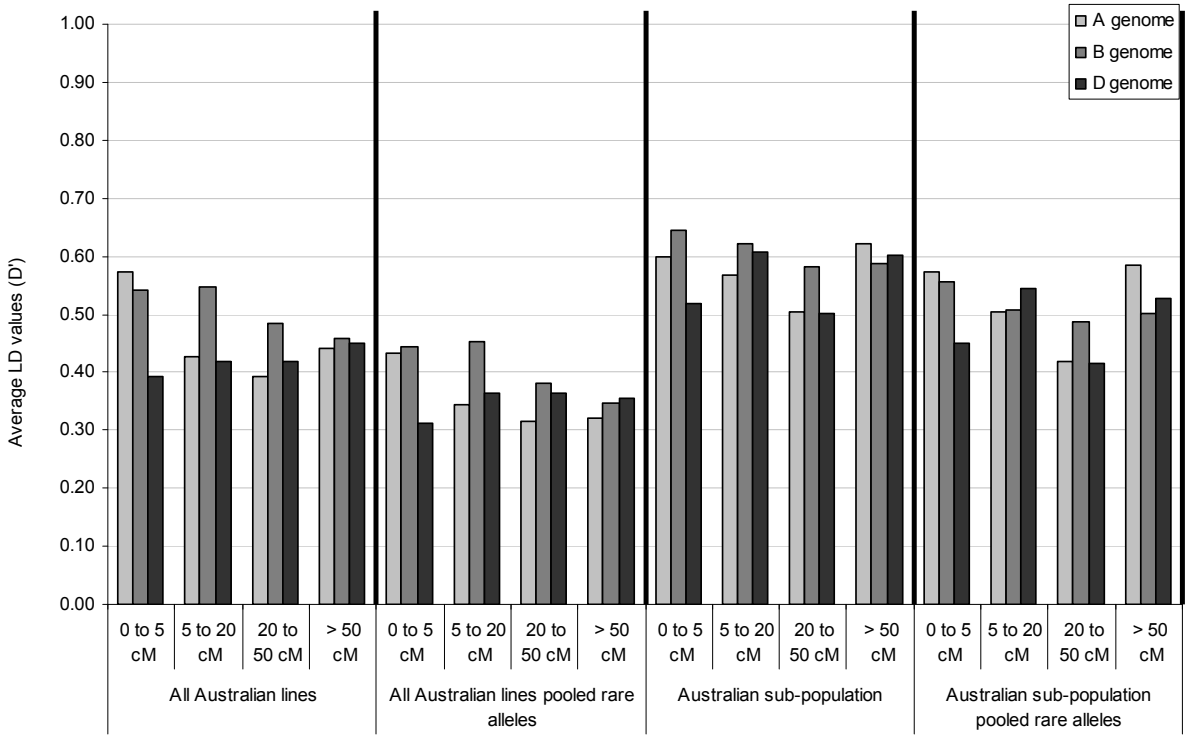


FIGURE 4.10 Bar graph illustrating the distribution of average D' values of the group 7 chromosomes (7A, 7B, and 7D). Results are presented for markers in each of the four marker 'classes' within each experimental scenario (named along the X axis).

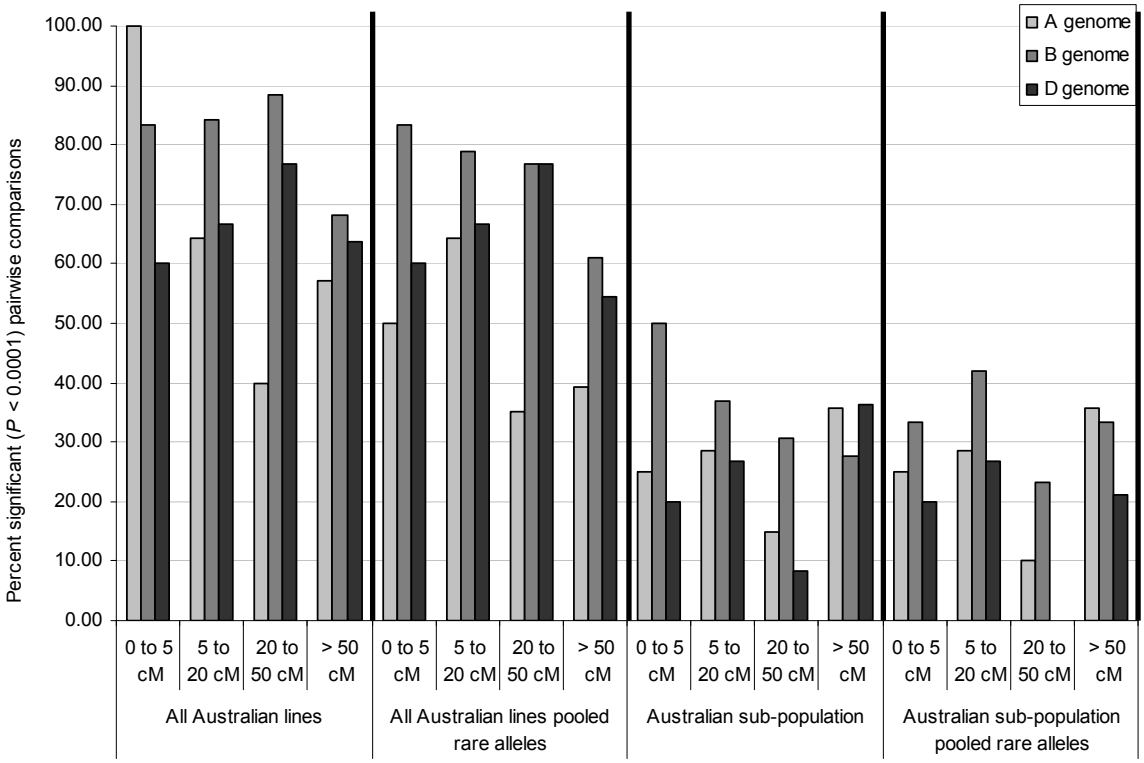


FIGURE 4.11 Bar graph illustrating the distribution of percent significant pairwise comparisons of the group 7 chromosomes (7A, 7B, and 7D). Results are presented for markers in each of the four marker 'classes' within each experimental scenario (named along the X axis).

comparisons was also reduced amongst the three genomes of chromosome group 7. When population structure was taken into account there was a marked increase in size of the D' values observed. Furthermore there was a greater reduction in the number of pair-wise comparisons that were significant at the P <0.0001 level.

4.3.5 LD in a large European Population

In contrast to the relatively small population size of the Australian hexaploid wheat study, an additional study employing over twice as many wheat lines but 44% fewer microsatellite markers was conducted. This was carried out to identify levels and patterns of LD in a different, geographically distinct population of hexaploid wheat with the anticipation of discovering differing levels of LD and thus potentially an additional population, which would have different applications in the association mapping of complex traits.

Two hundred and twenty five lines were genotyped with 70 microsatellite markers, as outlined in the materials and methods section, courtesy of Pauline Stephenson (JIC). These 70 microsatellite markers result in 97 syntenic pair-wise comparisons and 2,318 non-syntenic pair-wise comparisons. Table 4.2 provides a summary of the syntenic and non-syntenic pair-wise marker comparisons used in this experiment. Syntenic markers were examined in the same five marker 'classes' as those outlined in Section 4.3.1. This includes 11 (11%) markers separated by < 5 cM, a further 16 (16.5%) separated between 5 and 20 cM, 32 (33%) markers separated by 20 cM <50 cM, and a total of 38 (39%) markers that are more than 50 cM apart. There is a continual reduction in the number of marker loci used in each of the four experiments since pooling rare alleles into a common class and reducing the number of lines in the formation of the sub-population identified numerous monomorphic

TABLE 4.2 Number of pairwise comparisons performed in the LD analysis of the complete UK wheat dataset (225 lines) as well as the largest UK sub-population as determined from STRUCTURE in Section 3.3.3 (93 lines). Distinction is also made between pairwise comparisons using all of the markers with all alleles as well as with rare alleles (frequency <0.05) pooled. Marker 'classes' were arbitrarily chosen with the marker distances obtained from Somers et al. (2004).

		Number of pairwise marker comparisons in each 'class' of syntenic markers					
	Number of SSR markers used	0 to 5 cM	5 to 20cM	20 to 50cM	> 50 cM	Non-syntenic	Total pairwise comparisons
Entire UK dataset	70	11	16	32	38	2318	2415
Entire UK dataset with pooled rare alleles	69	9	16	31	37	2253	2346
UK Sub-population	65	9	14	26	31	1999	2079
UK Sub-population with pooled rare alleles	55	7	10	18	18	1432	1485

markers. When analysing the 225 wheat lines as a whole, the genome wide average of D' is 0.4486 and the LD decay with distance is only slight (illustrated in Figure 4.12). With respect to the non-syntenic marker pairs, D' averaged 0.3730 which is 1.2 times lower than that observed amongst the syntenic markers.

Figure 4.13 illustrates the distribution of the D' values in each of the five marker 'classes' and Figure 4.14 illustrates the percentage of pair-wise comparisons that are significant at the $P < 0.0001$ level. When examining the D' values estimated in the entire UK population, there is a clear and steady decline in the magnitude of the LD statistic across the five marker 'classes'. There are also a high proportion of pair-wise comparisons significant at the $P < 0.0001$ level when examining the population as a whole and appears to remain elevated across all five 'classes'. For unlinked markers (separated by >50 cM) and non-syntenic markers the estimated D' values remain high averaging 0.4375 and 0.3738 respectively. Furthermore, a high percentage of significant pair-wise marker comparisons in these later two marker 'classes' remains (73.7% and 67.6% respectively). Of all the syntenic pair-wise comparisons, 73.4% were significant at the $P <0.0001$ level, whereas 87.5% and 73.68% were significant at the same level for markers separated by 5 to 20 cM and >50 cM, respectively. This is in contrast to the non-syntenic marker pairs where only 67.62% were significant at that level.

When accounting for both rare alleles and population structure, the results are very similar to those obtained in the Australian dataset. The genome wide average of D', in the absence of rare alleles, decreased 1.2 times to 0.3782 followed by an increase by the same amount to 0.5240 when only population structure was accounted for. Again, the number of pair-wise comparisons significant at the $P <0.0001$ level, drastically decreased when taking these evolutionary and genetic factors into account

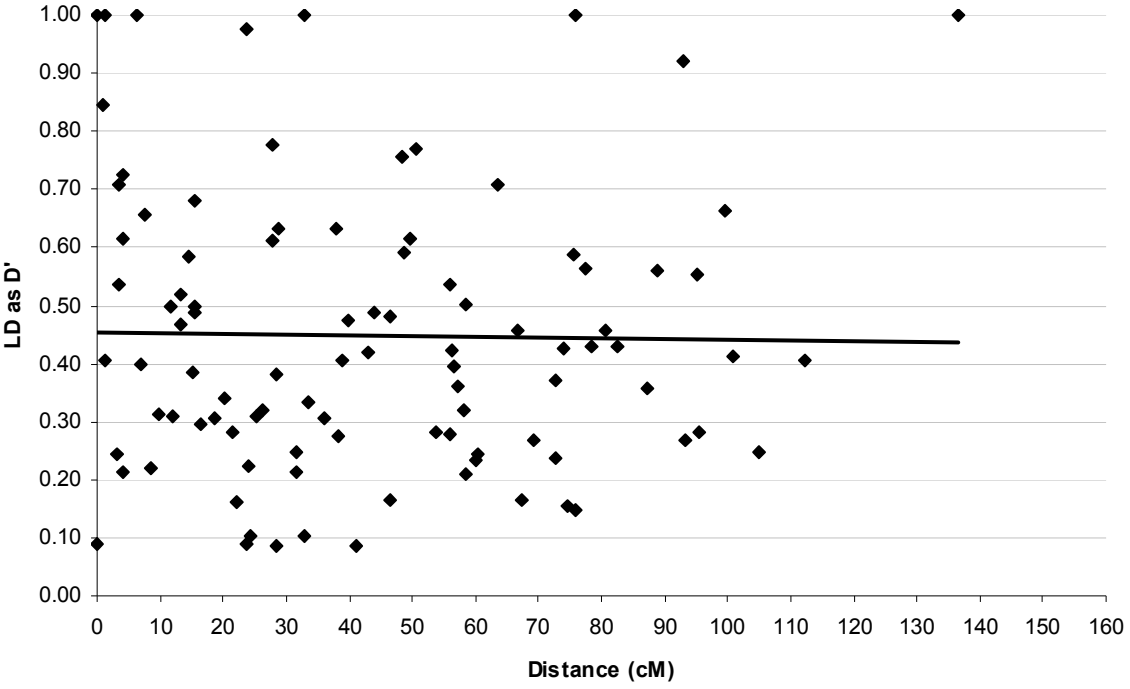


FIGURE 4.12 Scatter plot of LD (D') vs marker distance (cM) for whole UK data set. The trend line provides visualisation of the decay of LD with distance in the 97 pair-wise comparisons.

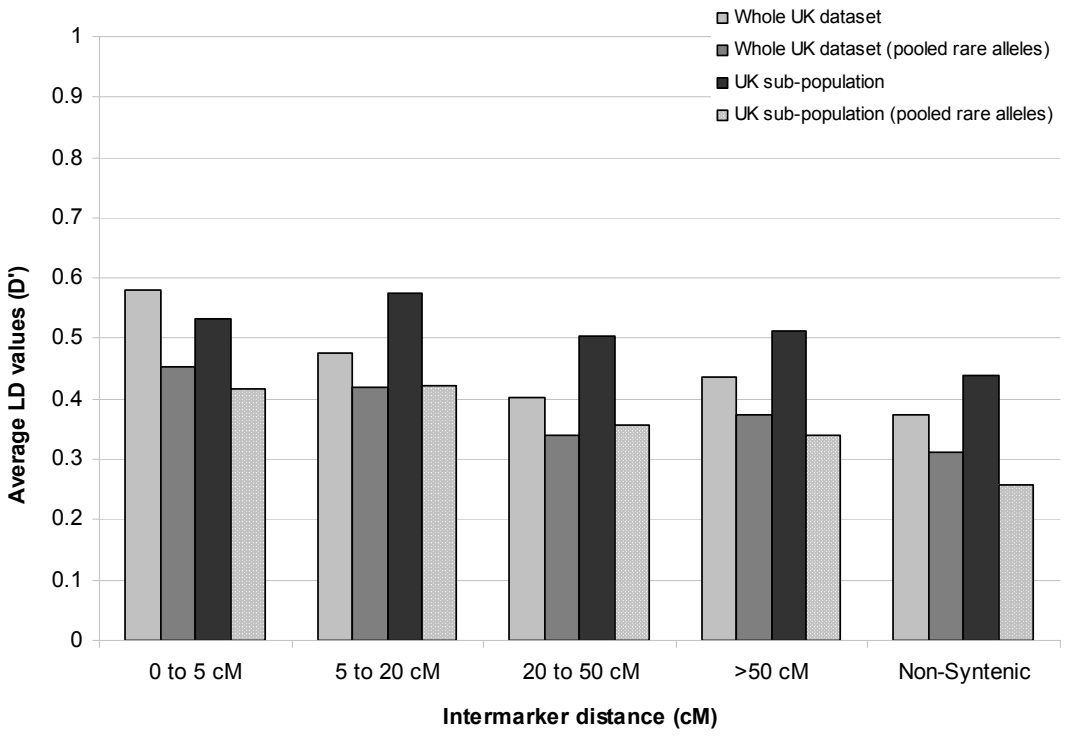


FIGURE 4.13 Bar graph illustrating the distribution of average D' values for each of the four experimental scenarios (outlined in the legend) within each of the five marker 'classes' (on the x- axis) for the UK wheat data set.

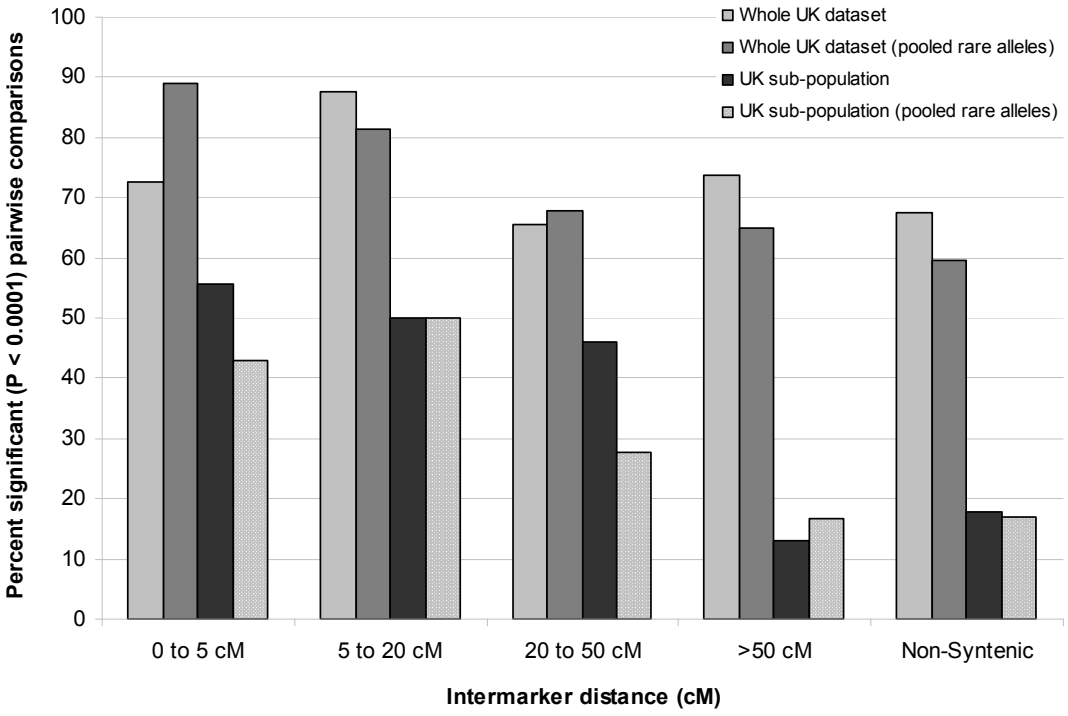


FIGURE 4.14 Bar graph illustrating the distribution of percent significant pair-wise comparisons for each of the four experimental scenarios (outlined in the legend) within each of the five marker 'classes' (on the x-axis) for the UK wheat dataset.

(Figure 4.14). It would appear that population structure has a greater negative effect on the level of significant comparisons in the UK data set when compared to the Australian data set. This may be due to the size of the initial data set, the size of the resulting data set and the number and number of alleles of markers used in each study.

4.4 Discussion

In this study SSR markers were used to estimate the genome wide levels of LD in two experimental hexaploid wheat populations and to evaluate the effects of two evolutionary factors, population structure and rare alleles. By taking into account population stratification and rare alleles, as well as attributing some of the results simply to the type of population examined, results provide a preliminary assessment on LD levels and structure in this economically important crop and the consequences of accounting for evolutionary and genomic factors in the analysis.

The levels of LD in both the Australian and UK experimental wheat populations are extensive. When calculating D' for the 96 Australian lines the D' value decreases from 0.56 to 0.45 over 50cM, and is even less pronounced when examining the 225 UK lines. The levels of LD in wheat are expected to be elevated due to the mating system, self-pollinated. Furthermore, the extent of LD observed in this study specifically, is likely due to the fact that a majority of the lines are locally adapted to Australian (or UK) environments and the selection pressure applied during breeding cycles. The direct selection for traits or allele combinations by breeders leads to population bottlenecks as well as a small founding population, which will increase the overall frequency of alleles and reduce the number of alleles in a population thus increasing the similarity of individuals, reducing the effective population size and increasing LD (Crepieux et al. 2004).

Additionally, a majority of the Australian lines included in this study were also examined in a study by Parker et al. (2002) who were seeking to understand the impact that semi-dwarf germplasm introduction had on the genetic diversity in Australian wheat breeding programs. The 'old' pre-semi-dwarf varieties were found to be less genetically diverse than the 'new' post-semi-dwarf varieties. A majority of the lines included in this LD study originate from breeding programs using the former varieties, which results in lines with very similar lineage (Parker et al. 2002), supporting the findings of elevated genome wide levels of LD in this study.

In a recent publication, Maccaferri et al. (2005) examined the levels of LD in the tetraploid wheat species *T. durum*. They also observed extensive LD between syntenic and non-syntenic markers. Maccaferri et al. (2005) report that within their data set as a whole (134 accessions), the highest LD value (D' = 0.67) was obtained for markers with an intermarker distance of <10 cM and was reduced (D' = 0.31) for markers separated by >50 cM. This has a similar trend to the decay of LD observed in hexaploid wheat where LD decayed from D' = 0.53 (markers <5 cM) to D' = 0.44 (markers >50 cM). The reduction in LD values is more prominent in the durum data set and this may be accounted for by the type of population examined.

In *Arabidopsis,* recent studies have found that LD decays to insignificant levels at approximately 250 Kb, equivalent to 1 cM (Hagenblad and Nordborg 2002; Nordborg et al. 2002). However the populations used in these studies were largely global which typically resulted in less extensive LD than those of locally adapted populations alone. When local populations of *Arabidopsis* are examined, significant genome-wide LD extends beyond 50 to 100 cM (Nordborg et al. 2002). As with these *Arabidopsis* populations, the durum population used by Maccaferri et al. (2005) was chosen to represent the majority of the genetic diversity in the improved gene pools.

Examination of local versus global populations may provide an explanation of the LD levels seen in both the *Arabidopsis* and the durum wheat experiments.

The SSR marker system may also provide some explanation for the extensive LD levels observed. The age of the alleles present in the study will positively affect the level of allelic association as such new mutations that occur in the population will result in complete LD between it and the surrounding chromosome region (Rafalski and Morgante 2004). Providing there are limited generations within this hexaploid wheat population there will be exaggerated levels of association, which are not necessarily indicative of the species as a whole. Since SSR markers were used in this study and are known to have a higher mutation rate, primarily due to strand slippage during DNA replication and unequal crossing over, (Payseur and Nachman 2000; Schlotterer 2000; Vigouroux et al. 2002) than most other molecular marker systems, such as SNPs, this may be another factor contributing to the elevated levels of LD seen in this study.

By accounting for evolutionary factors such as the presence of rare alleles (those with a frequency <5%), and population structure in both the Australian and UK data sets, the positive impact they had on the magnitude of D' and the negative impact on the number of pair-wise comparisons that proved to be significant was evident (Figures 4.2 and 4.3). The presence of population structure amongst a number of lines in a study can greatly increase the probability of type 1 error rate and thus skew allelic associations as was previously noted in the *Dwarf8* study by (Thornsberry et al. 2001). The effect of population sub-structure within this maize population was evaluated and the largest sub-population, as determined by *Structure* was analysed independently for D' and resulted in higher D' values. However the detection of allele association over distance remained comparable.

In the durum wheat study, Maccaferri et al. (2005) also observed inflated LD values when examining sub-populations consisting of fewer individuals. Although for tightly linked markers the LD values remained at similar levels there does appear to be an average increase in the magnitude of D' approximately 1.2-fold. Since D' is known to be upwardly influenced by small sample size, the substantial increase in the magnitude of D' in the Australian sub-population (Figure 4.2) is likely to be due to the fact that only 22 individuals were examined. In the UK dataset there was a slight increase in the D' values when accounting solely for population-structure but that was reduced when the rare alleles were taken into account. Perhaps the most striking feature of the results obtained from the UK data set is the reduction in the number of significant pair-wise comparisons within the sub-population amongst unlinked syntenic markers from 74% to 17% (Figure 4.14).

The upward bias of D' values as a function of sample size was calculated by McRae et al. (2002) through computer simulations while examining LD in domesticated sheep. McRae et al. (2002) suggest that their LD estimates were upwardly biased from 0.025 to 0.05 while using sample sizes of 482 and 276, respectively. While the 96 Australian wheat lines in this study comprise a sample with fewer individuals than those in the sheep study (482), the 22 wheat lines in the largest sub-population undoubtedly demonstrate exaggerated LD values because of this upward bias. Assuming a similar trend for the impact of sample size on D' values, D' values from the Australian wheat data would be upwardly biased by approximately 0.1 and above 0.2 for the 96 individuals and Sub-population 1, respectively.

The UK data set also had a slight upward bias in the magnitude of D'. Comparing the D' values observed in this study with the results from simulations of

McRae et al. (2002), both the complete UK data set (225 lines) and the sub-population (93 lines) demonstrate an estimated upward bias of 0.2. However, the study of LD in *Arabidopsis* was carried out on populations which were much smaller than the total population size described in this Chapter (Nordborg et al. 2002), stressing that there is merit in estimating levels of LD in populations which are available in order to gain a starting point from which further research can be conducted.

Populations for which LD levels extend over great physical distances will not be useful in the fine mapping of QTL in any species. However, they may provide a useful tool in identifying broad regions of the genome controlling a particular trait of interest. Once a region has been identified, populations where LD breaks down rapidly (i.e. global populations or including wild species in the study) could be examined to precisely locate the gene of interest. This two-tier approach to localise genes controlling traits of interest is possibly the most efficient way of using previously generated marker data.

Although genome wide averages of D' provide an initial impression of the D' levels in an experimental population of an organism as a whole, it provides little evidence to the levels of LD decay in specific regions of the genome and how, or indeed if, this approach will be useful in fine mapping complex traits of interest in particular regions of the genome. The analyses relating to genome and group specific estimations of LD gave some insight into how the patterns vary around the genome. Overall, the B genome demonstrated higher levels of LD and maintained those elevated levels over a distance of 100 cM, with the D genome demonstrating the lowest average D' values and decayed the least (0.01) over 100 cM (Figure 4.7). Examinations of individual chromosome groups provided more information with

respect to LD levels in specific regions of the genome. Overall the Group 6 chromosomes had the highest average D' values and appeared to have a positive correlation with marker distance. This is possibly due to the small number of Group 6 specific markers used in this analysis, which contributed to a mere seven pair-wise comparisons or the presence of alien chromosome segments on this group. A study by Paull et al. (1998) lists a number of disease resistance genes introgressed into hexaploid wheat, from related species and exotic germplasm, on chromosome groups 2, 6, and 7. Since recombination works to reduce allelic association, as discussed previously, and there is a lack of recombination within alien chromosome segments, as demonstrated by (Paull et al. 1998), this would explain the elevated levels of LD estimated in these chromosome groups.

Groups 4 and 5 had the second and third highest average D' values respectively and also demonstrated the most rapid rate of decay than anywhere else in the genome. Group 4 chromosomes declined from an approximate D' value of 0.7000 to 0.3700 from 0 cM to 95 cM (the most distal markers) whereas Group 5 showed a decline in LD from D' values of 0.9000 to 0.3300 over 100 cM. A summary of the decay values and those of the other group specific analyses are shown in Figure 4.6.

It would appear from the results presented in this study that the marker coverage used in the UK data set (on average 1 marker every 35 cM) is sufficient to detect LD in locally adapted germplasm pools since the increase in map saturation as seen in the Australian data set, particularly the Group 7 data set, did not offer any more power in detection. Likewise, the number of lines available in the UK germplasm experiment potentially offered a more realistic view of the magnitude of D' in these populations, since this measure of LD is greatly affected by small population sizes.

The results presented here suggest that over the past 100 years of wheat breeding recombination and genome mixing has occurred at only a low rate. There appears to be a transfer of large chromosome segments or linkage blocks at each generation. This also implies that a much simpler breeding strategy consisting largely of backcrossing has been used to introgress traits of interest. Understandably, more research is required, particularly in different populations in order to assess the general rate of LD decay in hexaploid wheat. Since levels of LD do decay with distance, even in these populations, it is possible for association mapping to be used in the detection of QTL in hexaploid wheat. What remains to be seen is the potential of association mapping for the fine mapping of traits.

A study by Paull et al. (1998) demonstrated the effectiveness of association mapping in the detection a number of stem-rust, leaf-rust, and yellow-rust resistance genes in 23 important Australian wheat varieties. Paull et al. (1998) were able to successfully identify RFLP markers associated with six segments of chromatin harbouring disease resistant genes. In one example an RFLP marker proved to be associated with a gene complex mapped to an 'alien segment' on chromosome 2A. This gene complex was transferred to a wheat variety through five backcrosses. This variety was subsequently backcrossed to another line an additional three times to produce the resistant variety, 'Trident'. This gene complex consisting of three genes remained in-tact despite the number of backcross generations, as too did a locus conferring CCN (Cereal Cyst Nematode) resistance (Paull et al. 1998). These results suggest that alien chromatin introgressed into hexaploid wheat does not tend to recombine. This has direct implications for association mapping and LD analysis since this lack of recombination will present itself in the form of large segments that remain in tact and are transferred from generation to generation as large linkage blocks. This will result in extensive LD and it is therefore, important to identify and

account for these introgressed segments when studying the extent of LD in particular genomic regions.

In this Chapter, genome wide distribution as well as patterns of LD was examined, with results indicating that it is extensively distributed throughout the genome between both linked and unlinked markers. This preliminary study is the first step in gaining an understanding of the pattens and distribution of LD in hexaploid wheat, with the next to examine the patterns and distribution of this genetic phenomenon in other populations in order to identify those with less LD to enable fine mapping of complex traits. This may involve work with populations of diploid and tetraploid wild progenitors of bread wheat.

Chapter 5: General Discussion

LD is the genetic phenomenon described as the non-random association of markers or alleles within particular haplotypes. In recent times there has been an explosion of interest in this area of population genetics as LD has been shown to be a powerful tool in the fine mapping of traits in higher eukaryotes Furthermore, it can be used to elucidate population histories and unravel aspects of evolutionary biology. The ongoing development of resources in molecular biology such as EST libraries, fine genetic and physical mapping, and high throughput assays, has facilitated the examination of gene structure, function, and patterns of diversity at the sequence level.

In plant species, LD and association mapping studies have been undertaken in some major crops, such as maize (Remington et al. 2001) and barley (Caldwell 2004), and is gaining momentum in other plant species as well such as forrest species (Cogan et al. 2006; Kumar et al. 2004; Thumma et al. 2005). The overall aim of this study was to obtain a preliminary understanding of the patterns and distribution of LD in the economically important crop, *T. aestivum*. There were two approaches used in reaching this goal. In Chapter 3 two genes at the *Ha* locus were sequenced and analysed for diversity within 44 *T.aestivum* accessions. These two genes represent a total of 894 bp of sequence spanning a region of approximately 33.1 Kb (Chantret et al. 2004). Complete sequencing of these genes revealed that there was a considerable lack of diversity amongst the lines studied and as such sequence variation was examined in the two-puroindoline genes (Pina and Pinb) and the closely related, physically linked, grain softness protein gene, Gsp-1, in a more diverse set of *Ae. tauschii* germplasm. The concatenated sequences, equating to nearly 70 Kb, revealed that LD was maintained over this distance; and in fact remained above $r^2 = 0.02$. Through the examination of this diverse data set it logical to speculate that LD in

hexaploid wheat may extend beyond 70 Kb in some regions of the genome. Direct sequencing of these genes provided an accurate and direct method of determining the genetic variation within this region and gave a platform for which LD could be examined.

To determine if an alternate marker system would resolve different patterns of LD while targeting larger intervals of the genome, microsatellite markers were employed. The microsatellite markers used in this study were chosen based on map location in order to provide uniform genome coverage in order to determine if LD decays in this experimental population over cM distances as opposed to bp distances.

LD was estimated on a large scale based on the microsatellite marker data. Prior to this however, genetic and evolutionary factors such as rare alleles and population structure were taken into account as these are known to positively affect the extent of observed LD (Jorde et al. 2000). The clustering algorithm in the software *Structure* was evaluated to test the appropriateness of this method in assigning individuals to sub-populations based on the genetic composition of each line. Despite being effective in doing the same in other organisms such as humans (Rosenberg et al. 2002), chickens (Rosenberg et al. 2001) and maize (Remington et al. 2001), the structure of the wheat population used here was not as straightforward.

Further to the study by Paull et al. (1998), Parker et al. (2002) examined the genetic diversity of the same 124 Australian wheat lines using microsatellite markers and pedigree data. They sought to understand the impact that the introduction of semi-dwarf germplasm had on the subsequent diversity within Australian wheat breeding programs. Dividing 101 of the initial 124 wheat lines into ‘New’ versus ‘Old’ revealed that; in general, the ‘Old’ pre-semi-dwarf lines were less genetically

diverse than the 'New' post-semi-dwarf varieties. Examining the regional Australian breeding programs and the parent material used in this study, revealed a number of the breeding programs produced lines that were very closely related. In particular the New South Wales, Victorian and South Australian-Waite breeding programs had considerable common germplasm in the varieties they have recently produced.

The majority of lines included in this LD study are from the South Australian-Waite breeding program, with others originating from Victoria and New South Wales, which produce lines with similar lineage. These observations by Parker et al. (2002) explain why there were no distinct sub-groups identified when using *Structure* to account for population structure. Furthermore, older, less diverse varieties such as Ghurka, Dirk, and Gabo-Aus, as examples, were included in this data set. In fact, roughly one half of the lines evaluated for LD were released prior to the 1970's and approximately two thirds prior to the 1980's, which approximately coincides with the introduction of semi-dwarf varieties. The results presented in this study infer that (and according to Parker et al. (2002)), the set of lines examined in this study were less genetically diverse than if a more recent set of lines were studied.

Having accounted for population structure (Chapter 3), rare alleles, and having generated a large quantity of genotypic data (Chapter 4), an opportunity was created to examine LD on a genome wide scale. The results presented in Chapter 4 represent the first study of long range LD levels and patterns in hexaploid wheat. Examining the experimental population as a whole revealed extensive LD both between markers on the same chromosome as well as between markers on different chromosomes. These results are a product of the nature of the lines used in this study, as discussed above, since lines that are more homogeneous will have LD extending over greater distances. The greatest decline in significant LD estimations was

observed when the population structure and rare alleles were taken into account. The expected negative correlation between the number of significant pair wise calculations and distance was observed. However this may be attributed to the low number of lines used in the analysis.

Genome wide averages in bread wheat, although important, are not necessarily indicative of the patterns distributed throughout the genome. This has direct consequences on QTL mapping, since recombination rates, which work to reduce LD, vary drastically throughout the genome. This holds particularly true for regions containing introgressed alien chromatin derived from wild relatives. Alien chromatin does not readily recombine with recipient DNA and as a result extensive LD is likely to be observed across these regions (Paull et al. 1998). Paull et al. (1998), list a number of disease resistance genes that have been introgressed into chromosome groups 2, 6, and 7 in Australian germplasm from exotic germplasm and closely related species such as *T.boeoticum*, *A. elongatum*, *T. timopheevi*, *Ae. ventricosa* and *S. cereale*. The locations of these introgressed alien chromosome segments are consistent with the group specific LD results obtained in Chapter 4. The decrease in D' values with distance were observed to a lesser extent in these three chromosome groups when compared with the others. These findings suggest that the more prominent LD within these groups can be attributed to those chromosome segments that do no readily recombine and thus remain intact from generation to generation. From this analysis it is clear that there are very different results with respect to the rate of decay of LD over distance depending on the region under study.

The empirical data collected from this experiment led to the question: what are the practical applications of this technique in today's breeding programs? Since the resolution with which a QTL may be mapped is reliant on the extent of LD

around candidate genes, it is currently difficult to determine how this technique will offer any benefit over traditional QTL mapping methods. There have been several published reports stating the merits of utilising association mapping (LD) in conjunction with traditional QTL mapping (linkage analysis) in order to increase the resolution with which a gene is mapped and to also increase the statistical power of the analysis (Garris et al. 2003; Wu and Zeng 2001). Although there are benefits to both methods of trait mapping, association mapping allows for the use of large natural populations and as a result there is no need to establish and manage new populations. Additionally, previously accumulated phenotypic and marker data can be used thus reducing time in developing populations and generating maps.

Wheat lines from plant breeding programs, such as those used in this study, exhibit extensive LD. In such cases it will be virtually impossible to identify individual genes since the region that is in LD extends over large regions, on the order of cM. There is therefore a need to identify reference populations in hexaploid wheat which display varying levels of LD. Once identified, these populations may be used with breeding populations in a two-tiered approach to association mapping. Firstly, populations with extensive LD, such as a data set consisting of locally adapted lines, could be scanned with markers across the genome such that broad regions associated with a particular phenotype are identified. This strategy was implemented recently by Breseghello and Sorrells (2006), on their study of association mapping around the kernel morphology and milling quality loci. In this study 95 cultivars representing elite breeding stock of soft winter wheat were examined. LD was shown to decay within 1 cM in one region and at approximately 5 cM in another. This later, broad region could be targeted in a subsequent study utilizing a population with greatly reduced LD, such as a global population, in order to define the individual gene responsible for the phenotype.

Elucidating the genes that control agronomically important traits has long been the work of geneticists and breeders. Marker assisted selection (MAS) has already been adopted into many breeding programs with plant breeders realising the advantages of using such a technology (Gupta et al. 2005a; Hayden et al. 2004). Even so, there is still a high cost to marker development and deployment of such markers to a large number of populations. The development of new marker technologies in plants, SNPs in particular, will provide a vast source of polymorphic markers for the use in high-resolution genetic mapping and enable the use of high-throughput detection methods (Kwok and Chen 2003). This may prove to be particularly useful to target the underlying genetic control of new traits such as disease and insect resistance, water utilitzation and nitrogen use efficiency, as examples. As more SNP markers become available and patterns of LD emerge, identification of minimum marker sets, or "tagging" SNPs (Halldorsson et al. 2004), to search for marker allele associations will be identified; this was one of the main goals of the Hapmap project in humans and has proven successful (Consortium 2003; Consortium 2005).

As the current marker technologies are enhanced and the understanding of population and genetic dynamics is ameliorated, there will be growing enthusiasm surrounding the applications of LD and association mapping; specifically in how these tools may enable the precise location and subsequent isolation of individual genes that control agronomically important traits. It is still not clear what the far-reaching benefits of these methods in QTL mapping will be. However with continued research in this field, combined with the integration of markers in breeding programs, these benefits should soon be realised.

Literature Cited

Abecasis, G.R., E. Noguchi, A. Heinzmann, J.A. Traherne, S. Bhattacharyya, N.I. Leaves, G.G. Anderson, Y. Zhang, N.J. Lench, A. Carey, L.R. Cardon, M.F. Moffatt, and W.O.C. Cookson. 2001. Extent and distribution of linkage disequilibrium in three genomic regions. *American Journal of Human Genetics* **68:** 191-197.

Akbari, M., P. Wenzl, V. Caig, J. Carling, L. Xia, S. Yang, G. Uszynski, V. Mohler, A. Lehmensiek, H. Kuchel, M.J. Hayden, N. Howes, P.J. Sharp, P. Vaughan, B. Rathmell, E. Huttner, and A. Kilian. 2006. Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. *Theoretical and Applied Genetics* **113:** 1409-1420.

Akhunov, E.D., A.W. Goodyear, S. Geng, L.-L. Qi, B. Echalier, B.S. Gill, Miftahudin, J.P. Gustafson, G. Lazo, S. Chao, O.D. Anderson, A.M. Linkiewicz, J. Dubcovsky, M.L. Rota, M.E. Sorrells, D. Zhang, H.T. Nguyen, V. Kalavacharla, K. Hossain, S.F. Kianian, J. Peng, N.L.V. Lapitan, J.L. Gonzalez-Hernandez, J.A. Anderson, D.-W. Choi, T.J. Close, M. Dilbirligi, K.S. Gill, M.K. Walker-Simmons, C. Steber, P.E. McGuire, C.O. Qualset, and J. Dvorak. 2003. The Organization and Rate of Evolution of Wheat Genomes Are Correlated With Recombination Rates Along Chromosome Arms. *Genome Research* **13:** 753-763.

Aranzana, M.i.a.J.e., S. Kim, K. Zhao, E. Bakker, M. Horton, K. Jakob, C. Lister, J. Molitor, C. Shindo, C. Tang, C. Toomajian, B. Traw, H. Zheng, J. Bergelson, C. Dean, P. Marjoram, and M. Nordborg. 2005. Genome-Wide Association Mapping in Arabidopsis Identifies Previously Known Flowering Time and Pathogen Resistance Genes. *PLoS Genetics* **1:** e60 EP -.

Arumuganathan, K. and E.D. Earle. 1991. Nuclear DNA Content of Some Important Plant Species. *Plant Molecular Biology Reporter* **9:** 211-215.

Balfourier, F., V. Roussel, P. Strelchenko, F. Exbrayat-Vinson, P. Sourdille, G. Boutet, J. Koenig, C. Ravel, O. Mitrofanova, M. Beckert, and G. Charmet. 2007. A worldwide bread wheat core collection arrayed in a 384-well plate. *Theoretical and Applied Genetics* **114:** 1265-1275.

Bass, C., R. Hendley, M.J. Adams, K.E. Hammond-Kosack, and K. Kanyuka. 2006. The Sbm1 locus conferring resistance to Soil-borne cereal mosaic virus maps to a gene-rich region on 5DL in wheat. *Genome* **49:** 1140-1148.

Batley, J., G. Barker, H. O'Sullivan, K.J. Edwards, and D. Edwards. 2003. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiology* **132:** 84-91.

Borevitz, J.O. and M. Nordborg. 2003. The Impact of Genomics on the Study of Natural Variation in Arabidopsis. *Plant Physiology* **132:** 718-725.

Boyko, E., R. Kalendar, V. Korzun, J. Fellers, A. Korol, A.H. Schulman, and B.S. Gill. 2002. A high-density cytogenetic map of the *Aegilops tauschii* genome incorporating retrotransposons and defense-related genes: insights into cereal chromosome structure and function. *Plant Molecular Biology* **48:** 767-790.

Brennan, C.S., N. Harris, D. Smith, and F.R. Shewry. 1996. Structural differences in the mature endosperms of good and poor malting barley cultivars. *Journal of Cereal Science* **24:** 171-177.

Breseghello, F. and M.E. Sorrells. 2006. Association Mapping of Kernel Size and Milling Quality in Wheat (*Triticum aestivum* L.) Cultivars. *Genetics* **172:** 1165-1177.

Buckler IV, E.S. and J.M. Thornsberry. 2002. Plant molecular diversity and applications to genomics. *Current Opinion in Plant Biology* **5:** 107-111.

Caldwell, K.S. 2004. An Evaluation of the Patterns of Nucleotide Diversity and Linkage Disequilibrium at the Regional Level in Hordeum vulgare. In *School of Agriculture and Wine*, pp. 183. The University of Adelaide, Adelaide.

Chalmers, K.J., A.W. Campbell, J. Kretschmer, A. Karakousis, P.H. Henschke, S. Pierens, N. Harker, M. Pallotta, G.B. Cornish, M.R. Shariflou, L.R. Rampling, A. McLauchlan, G. Daggard, P.J. Sharp, T.A. Holton, M.W. Sutherland, R. Appels, and P. Langridge. 2001. Construction of three linkage maps in bread wheat, *Triticum aestivum*. *Australian Journal of Agricultural Research* **52:** 1089-1119.

Chantret, N., A. Cenci, F. Sabot, O. Anderson, and J. Dubcovsky. 2004. Sequencing of the Triticum monococcum Hardness locus reveals good microcolinearity with rice. *Molecular Genetics and Genomics* **271:** 377-386.

Chantret, N., J. Salse, F. Sabot, S. Rahman, A. Bellec, B. Laubin, I. Dubois, C. Dossat, P. Sourdille, P. Joudrier, M.-F. Gautier, L. Cattolico, M. Beckert, S. Aubourg, J. Weissenbach, M. Caboche, M. Bernard, P. Leroy, and B. Chalhoub. 2005. Molecular Basis of Evolutionary Events That Shaped the Hardness Locus in Diploid and Polyploid Wheat Species (Triticum and Aegilops). *THE PLANT CELL Plant Cell* **17:** 1033-1045.

Chao, S., P.J. Sharp, A.J. Worland, E.J. Warham, R.M.D. Koebner, and M.D. Gale. 1989. RFLP-based genetic linkage maps of wheat homoeologous group 7 chromosomes. *Theoretical and Applied Genetics* **78:** 495-504.

Charmet, G., N. Robert, M.R. Perretant, G. Gay, P. Sourdille, C. Groos, S. Bernard, and M. Bernard. 2001. Marker assisted recurrent selection for cumulating QTLs for bread-making related traits. *Euphytica* **119:** 89-93.

Ching, A., K.S. Caldwell, M. Jung, M. Dolan, O.S. Smith, S. Tingey, M. Morgante, and A. Rafalski. 2002. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. In *BMC Genetics*.

Clark, A.G., E. Boerwinkle, J. Hixson, and C.F. Sing. 2007. Determinants of the success of whole-genome association testing. *Genome* **15:** 1463-1467.

Clark, R.M., E. Linton, J. Messing, and J.F. Doebley. 2004. Inaugural Article: Pattern of diversity in the genomic region near the maize domestication gene tb1 10.1073/pnas.2237049100. *Proceedings of the National Academy of Sciences* **101:** 700-707.

Cogan, N., R. Ponting, A. Vecchies, M. Drayton, J. George, P. Dracatos, M. Dobrowolski, T. Sawbridge, K. Smith, G.n. Spangenberg, and J. Forster. 2006. Gene-associated single nucleotide polymorphism discovery in perennial ryegrass (Lolium perenne L.). *Molecular Genetics and Genomics* **276:** 101-112.

Consortium, T.I.H. 2003. The International HapMap Project. **426:** 789-796.

Consortium, T.I.H. 2005. A haplotype map of the human genome. **437:** 1299-1320.

Crepieux, S., C. Lebreton, B. Servin, and G. Charmet. 2004. Quantitative Trait Loci (QTL) Detection in Multicross Inbred Designs: Recovering QTL Identical-by-Descent Status Information From Marker Data 10.1534/genetics.104.028993. *Genetics* **168:** 1737-1749.

CSIRO, P.I. 2005. Rapid ID system separates wheat from the chaff. CSIRO.

Cuthbert, P.A., D.J. Somers, J. Thomas, S. Cloutier, and A. Brule-Babel. 2006. Fine mapping *Fhb1*, a major gene controlling fusarium head blight resistance in bread wheat (*Triticum aestivum* L.). *Theoretical and Applied Genetics* **112:** 1465-1472.

Daly, M.J., J.D. Roioux, S.F. Schaffner, T.J. Hudson, and E.S. Lander. 2001. High-resolution haplotype structure in the human genome. *Nature Genetics* **29:** 229-232.

Dawson, E., G.R. Abecasis, S. Bumpstead, Y. Chen, S. Hunt, D.M. Beare, J. Pabial, T. Dibling, E. Tinsley, S. Kirby, D. Carter, M. Papaspyridonos, S. Livingstone, R. Ganske, E. Lohmussaar, J. Zernant, N. Tonisson, M. Remm, R. Magi, T. Puurand, J. Vilo, A. Kurg, K. Rice, P. Deloukas, R. Mott, A. Metspalu, D.R. Bentley, L.R.

Cardon, and I. Dunham. 2002. A first-generation linkage disequilibrium map of human chromosome 22. **418:** 544-548.

del Bosque-Plata, L., C.A. Aguilar-Salinas, M.T. Tusie-Luna, S. Ramirez-Jimenez, M. Rodriguez-Torres, M. Auron-Gomez, E. Ramirez, M.L. Velasco-Perez, A. Ramirez-Silva, F. Gomez-Perez, C.L. Hanis, T. Tsuchiya, I. Yoshiuchi, N.J. Cox, and G.I. Bell. 2004. Association of the calpain-10 gene with type 2 diabetes mellitus in a Mexican populaiton. *Molecular Genetics and Metabolism* **81:** 122-126.

Delvin, B. and N. Risch. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29:** 311-322.

Devos, K.M. and M.D. Gale. 1997. Comparative genetics in the grasses. *Plant Molecular Biology* **35:** 3-15.

Devos, K.M. and M.D. Gale. 2000. Genome Relationships: The Grass Model in Current Research. *The Plant Cell* **12:** 637-646.

Dooner, H.K. and I.M. Martinez-Ferez. 1997. Recombination occurs uniformly within the *bronze* gene, a meiotic recombination hotspot in the maize genome. *The Plant Cell* **9:** 1633-1646.

Dunham, A. L.H. Matthews J. Burton J.L. Ashurst K.L. Howe K.J. Ashcroft D.M. Beare D.C. Burford S.E. Hunt S. Griffiths-Jones M.C. Jones S.J. Keenan K. Oliver C.E. Scott R. Ainscough J.P. Almeida K.D. Ambrose D.T. Andrews R.I.S. Ashwell A.K. Babbage C.L. Bagguley J. Bailey R. Bannerjee K.F. Barlow K. Bates H. Beasley C.P. Bird S. Bray-Allen A.J. Brown J.Y. Brown W. Burrill C. Carder N.P. Carter J.C. Chapman M.E. Clamp S.Y. Clark G. Clarke C.M. Clee S.C.M. Clegg V. Cobley J.E. Collins N. Corby G.J. Coville P. Deloukas P. Dhami I. Dunham M. Dunn M.E. Earthrowl A.G. Ellington L. Faulkner A.G. Frankish J. Frankland L. French P. Garner J. Garnett J.G.R. Gilbert C.J. Gilson J. Ghori D.V. Grafham S.M. Gribble C. Griffiths R.E. Hall S. Hammond J.L. Harley E.A. Hart P.D. Heath P.J. Howden E.J. Huckle P.J. Hunt A.R. Hunt C. Johnson D. Johnson M. Kay A.M. Kimberley A. King G.K. Laird C.J. Langford S. Lawlor D.A. Leongamornlert D.M. Lloyd C. Lloyd J.E. Loveland J. Lovell S. Martin M. Mashreghi-Mohammadi S.J. McLaren A. McMurray S. Milne M.J.F. Moore T. Nickerson S.A. Palmer A.V. Pearce A.I. Peck S. Pelan B. Phillimore K.M. Porter C.M. Rice S. Searle H.K. Sehra R. Shownkeen C.D. Skuce M. Smith C.A. Steward N. Sycamore J. Tester D.W. Thomas A. Tracey A. Tromans B. Tubby M. Wall J.M. Wallis A.P. West S.L. Whitehead D.L. Willey L. Wilming P.W. Wray M.W. Wright L. Young A. Coulson R. Durbin T. Hubbard J.E. Sulston S. Beck D.R. Bentley J. Rogers and M.T. Ross. 2004. The DNA sequence and analysis of human chromosome 13. **428:** 522-528.

Dunning, A., M., F. Durocher, C.S. Healey, M.D. Teare, S.E. McBride, F. Carlomagno, C.-F. Xu, E. Dawson, S. Rhodes, S. Ueda, E. Lai, R.N. Luben, E.J. Van Rensburg, A. Mannermaa, V. Kataja, G. Rennart, I. Dunham, I. Purvis, D. Easton, and B.A.J. Ponder. 2000. The extent of linkage disequilibrium in four populations with distinct demographic histories. *American Journal of Human Genetics* **67:** 1544-1554.

Dvorak, J., M.-C. Luo, and Z.-L. Yang. 1998. Restriction Fragment Length Polymorphism and Divergence in the Genomic Regions of High and Low Recombination in Self-Fertilizing and Cross-Fertilizing Aegilops Species. *Genetics* **148:** 423-434.

Eaves, I.A., T.R. Merriman, R.A. Barber, S. Nutland, E. Tuomilehto-Wolf, J. Tuomilehto, F. Cucca, and J.A. Todd. 2000. The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nature Genetics* **25:** 320-323.

Eujayl, I., M.E. Sorrells, M. Baurn, P. Wolters, and W. Powell. 2002. Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. *Theoretical and Applied Genetics* **104:** 399-407.

Excoffier, L., G. Laval, and S. Schneider. 2005. Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* **1:** 47-50.

Falush, D., M. Stephens, and J.K. Pritchard. 2003. Inference of population structure using mulitlocus genotype data: Linked loci and correlated allele frequencies. *Genetics*.

FAO. 2006. Food Outlook-No.2: Global Market Analysis.

Faris, J.D., K.M. Haen, and B.S. Gill. 2000. Saturation Mapping of a Gene-Rich Recombination Hot Spot Region in Wheat. *Genetics* **154:** 823-835.

Feltus, F.A., J. Wan, S.R. Schulze, J.C. Estill, N. Jiang, and A.H. Paterson. 2004. An SNP Resource for Rice Genetics and Breeding Based on Subspecies Indica and Japonica Genome Alignments. *Genome Research*
Genome Res. **14:** 1812-1819.

Flint-Garcia, S.A., J.M. Thornsberry, and E.S. Buckler IV. 2003. Structure of Linkage Disequilibrium in Plants. *Annual Reviews of Plant Biology* **54:** 357-374.

Fu, H. and H.K. Dooner. 2002. Intraspecific violation of genetic colinearity and its implications in maize. *Proceedings of the National Academy of Sciences* **99:** 9573-9578.

Fu, Y.B., G.W. Peterson, J.K. Yu, L. Gao, J. Jia, and K.W. Richards. 2006. Impact of plant breeding on genetic diversity of the Canadian hard red spring wheat germplasm as revealed by EST-derived SSR markers. *Theoretical and Applied Genetics* **112:** 1239-1247.

Gabriel, S.B., S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S.N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E.S. Lander, M.J. Daly, and D. Altshuler. 2002. The Structure of Haplotype Blocks in the Human Genome. *Science* **296:** 2225-2229.

Gardiner, J.M., E.H. Coe, S. Melia-Hancock, D.A. Hoisington, and S. Chao. 1993. Development of a core RFLP map in maize using an immortalized F_2 population. *Genetics* **134:** 917-930.

Garris, A.J., S.R. McCouch, and S. Kresovich. 2003. Population Structure and Its Effect on Haplotype Diversity and Linkage Disequilibrium Surrounding the xa5 Locus of Rice (Oryza sativa L.). *Genetics* **165:** 759-769.

Gautier, M.-F., M.-E. Aleman, A. Guirao, D. Marion, and P. Joudrier. 1994. *Triticum aestivum* puroindolines, two basic cystine-rich seed proteins: cDNA sequence analysis and developmental gene expression. *Plant Molecular Biology* **25:** 43-57.

Gautier, M.-F., P. Cosson, A. Guirao, R. Alary, and P. Joudrier. 2000. Puroindoline genes are highly conserved in diploid ancestor wheats and related species but absent in tetraploid Triticum species. *Plant Science* **153:** 81-91.

Genome, S.P.R. 2005. The map-based sequence of the rice genome. **436:** 793-800.

Gill, K.S., B.S. Gill, T.R. Endo, and E.V. Boyko. 1996a. Identification and High-Density Mapping of Gene-Rich Regions in Chromosome Group 5 of Wheat. *Genetics* **143:** 1001-1012.

Gill, K.S., B.S. Gill, T.R. Endo, and T. Taylor. 1996b. Identification and High-Density Mapping of Gene-Rich Regions in Chromosome Group 1 of Wheat. *Genetics* **144:** 1883-1891.

Giroux, M.J. and C.F. Morris. 1997. A glycine to serine change in puroindoline b is associated with wheat grain hardness and low levels of starch-surface friabilin. *Theoretical and Applied Genetics* **95:** 857-864.

Giroux, M.J. and C.F. Morris. 1998. Wheat grain hardness results from highly conserved mutations in the friabilin components puroindoline a and b. *Proceedings of the National Academy of Sciences*
PNAS **95:** 6262-6266.

Gonzalez-Martinez, S.C., E. Ersoz, G.R. Brown, N.C. Wheeler, and D.B. Neale. 2006. DNA Sequence Variation and Selection of Tag Single-Nucleotide Polymorphisms at Candidate Genes for Drought-Stress Response in Pinus taeda L. 10.1534/genetics.105.047126. *Genetics* **172:** 1915-1926.

Gupta, P.K. 2002. Molecular markers and QTL analysis in crop plants. *Current Science* **83:** 113-114.

Gupta, P.K., H.S. Balyan, K.J. Edwards, P. Isacc, V. Korzun, M. Roder, M.-F. Gautier, P. Joudrier, A.R. Schlatter, J. Dubcovsky, R.C. De la Pena, M. Khairallah, G.A. Penner, M.J. Hayden, P.J. Sharp, B. Keller, R.C.C. Wang, J.P. Hardouin, P. Jack, and P. Leroy. 2002. Genetic mapping of 66 new microsatellite (SSR) loci in bread wheat. *Theoretical and Applied Genetics* **105:** 413-422.

Gupta, P.K., S. Rustgi, and P.L. Kulwal. 2005a. Linkage disequilibrium and association studies in higher plants: Present status and future prospects. *Plant Molecular Biology* **57:** 461-485.

Gupta, R.B., R.K. Varshney, P.C. Sharma, and B. Ramesh. 1999. Molecular markers and their applications in wheat breeding. *Plant Breeding* **118:** 369-390.

Gupta, S.K., A. Charpe, S. Koul, K.V. Prabhu, and Q.M. Haq. 2005b. Development and validation of molecular markers linked to an Aegilops umbellulata-derived leaf-rust-resistance gene, Lr9, for marker-assisted selection in bread wheat. *Genome* **48:** 823-830.

Hagenblad, J. and M. Nordborg. 2002. Sequence Variation and Haplotype Structure Surrounding the Flowering Time Locus *FRI* in *Arabidopsis thaliana*. *Genetics* **161:** 289-298.

Halldorsson, B.V., V. Bafna, R. Lippert, R. Schwartz, F.M. De La Vega, A.G. Clark, and S. Istrail. 2004. Optimal Haplotype Block-Free Selection of Tagging SNPs for Genome-Wide Association Studies
10.1101/gr.2570004. *Genome Res.* **14:** 1633-1640.

Hass, B.L., J.C. Pires, R. Porter, R.L. Phillips, and S.A. Jackson. 2003. Comparative genetics at the gene and chromosome levels between rice *(Oryza sativa)* and wildrice *(Zizania palustris)*. *Theoretical and Applied Genetics* **107:** 773-782.

Hayden, M.J., H. Kuchel, and K.J. Chalmers. 2004. Sequence tagged microsatellites for the *Xgwm533* locus provide new diagnostic markers to select for the presence of stem rust resistance gene Sr2 in bread wheat (*Triticum aestivum* L.). *Theoretical and Applied Genetics* **109:** 1641-1647.

Hayden, M.J. and P.J. Sharp. 2001. Sequence-tagged microsatellite profiling (STMP): a rapid technique for developing SSR markers. *Nucl. Acids Res.* **29:** e43-.

Hayden, M.J., P. Stephenson, A.M. Logojan, D. Khatkar, C. Rogers, J. Elsden, R.M.D. Koebner, J.W. Snape, and P.J. Sharp. 2006. Development and genetic mapping of sequence-tagged microsatellites (STMs) in bread wheat (Triticum aestivum L.). *Theoretical and Applied Genetics* **113:** 1271-1281.

He, Q., M.K. Viljanen, and J. Mertsola. 1994. Effects of thermocyclers and primers on the reproducibility of banding patterns in randomly amplified polymorphic DNA analysis. *Molecular and Cellular Probes* **8:** 155-160.

Horikawa, Y., N. Oda, N.J. Cox, X. Li, M. Orho-Melander, M. Hara, Y. Hinokio, T.H. Lindner, H. Mashima, P.E.H. Schwarz, L. del Bosque-Plata, Y. Horikawa, Y. Oda, I. Yoshiuchi, S. Colilla, K.S. Polonsky, S. Wei, P. Concannon, N. Iwasaki, J. Schulze, L.J. Baier, C. Bogardus, L. Groop, E. Boerwinkle, C.L. Hanis, and G.I. Bell. 2000. Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genetics* **26:** 163-175.

Hospital, F., L. Moreau, F. Lacoudre, A. Charcosset, and A. Gallais. 1997. More on the efficiency of marker-assisted selection. *Theoretical and Applied Genetics* **95:** 1181-1189.

http://wheat.pw.usda.gov/ggpages/SSR/WMC/. 2004. The Wheat Microsatellite Consortium.

Huang, X.Q., S. Cloutier, L. Lycar, N. Radovanovic, D.G. Humphreys, J.S. Noll, D.J. Somers, and P.D. Brown. 2006. Molecular detection of QTLs for agronomic and quality traits in a doubled haploid population derived from two Canadian wheats (Triticum aestivum L.). *Theoretical and Applied Genetics* **113:** 753-766.

Information, N.C.f.B. 2005.

Initiative, T.A. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796-815.

Jahier, Abelard, Tanguy, Dedryver, Rivoal, Khatkar, and Bariana. 2001. The Aegilops ventricosa segment on chromosome 2AS of the wheat cultivar 'VPM1' carries the cereal cyst nematode resistance gene Cre5 doi:10.1046/j.1439-0523.2001.00585.x. *Plant Breeding* **120:** 125-128.

Jannoo, N., L. Grivet, A. Dookun, A. D'Hont, and J.C. Glaszmann. 1999. Linkage disequilibrium among modern sugarcane cultivars. *Theoretical and Applied Genetics* **99:** 1053-1060.

Jeffreys, A.J., L. Kauppi, and R. Neumann. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics* **29:** 217-222.

Jia, L., M.T. Clegg, and T. Jiang. 2003. Excess non-synonymous substitutions suggest that positive selection episodes occurred during the evolution of DNA-binding domains in the *Arabidopsis* R2R3-MYB gene family. *Plant Molecular Biology* **52:** 627-642.

Johnson, G.C.L., L. Esposito, B.J. Barratt, A.N. Smith, J. Heward, G. Di Genova, H. Ueda, H.J. Cordell, I.A. Eaves, F. Dudbridge, R.C.J. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S.C.L. Gough, D.G. Clayton, and J.A. Todd. 2001. Haplotype tagging for the identification of common disease genes. *Nature Genetics* **29:** 233-237.

Jones, C.J., K.J. Edwards, S. Castaglione, M.O. Winfield, F. Sala, C. van de Wiel, G. Bredemeijer, B. Vosman, M. Matthes, A. Daly, R. Brettschneider, P. Bettini, M. Buiatti, E. Maestri, A. Malcevschi, N. Marmiroli, R. Aert, G. Volckaert, J. Rueda, R. Linacero, A. Vazquez, and A. Karp. 1997. Reproducibility of RAPD, AFLP and SSR markers in plants by a network of European laboratories. *Molecular Breeding* **3:** 381-390.

Jorde, L.B. 2000. Linkage disequilibrium and the search for complex disease genes. *Genome Research* **10:** 1435-1444.

Jorde, L.B., W.S. Watkins, J. Kere, D. Nyman, and A.W. Eriksson. 2000. Gene mapping in isolated populations: new roles for old friends? *Human Heredity* **50:** 57-65.

Kato, K., H. Miura, and S. Sawada. 2000. Mapping QTLs controlling grain yield and its components on chromosome 5A of wheat. *Theoretical and Applied Genetics* **101:** 1114-1121.

Klein, R.J., C. Zeiss, E.Y. Chew, J.-Y. Tsai, R.S. Sackler, C. Haynes, A.K. Henning, J.P. SanGiovanni, S.M. Mane, S.T. Mayne, M.B. Bracken, F.L. Ferris, J. Ott, C.

Barnstable, and J. Hoh. 2005. Complement Factor H Polymorphism in Age-Related Macular Degeneration
10.1126/science.1109557. *Science* **308:** 385-389.

Knowler, W., R. Williams, D. Pettitt, and A. Steinberg. 1988. $Gm^{3-5,13,14}$ and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *American Journal of Human Genetics* **43:** 520-526.

Kraft, T., M. Hansen, and N.O. Nilsson. 2000. Linkage disequilibrium and fingerprinting in sugar beet. *Theoretical and Applied Genetics* **101:** 323-326.

Kruglyak, L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* **22:** 139-144.

Kuchel, H., P. Langridge, L. Mosionek, K. Williams, and S.P. Jefferies. 2006. The genetic control of milling yield, dough rheology and baking quality of wheat. *Theoretical and Applied Genetics* **112:** 1487-1495.

Kumar, S., C. Echt, P.L. Wilcox, and T.E. Richardson. 2004. Testing for linkage disequilibrium in the New Zealand radiata pine breeding population. *Theoretical and Applied Genetics* **108:** 292-298.

Kwok, P.-Y. and X. Chen. 2003. Detection of Single Nucleotide Polymorphisms. *Current Issues in Molecular Biology* **5:** 43-60.

La Rota, M. and M.E. Sorrells. 2004. Comparative DNA sequence analysis of mapped wheat ESTs reveals the complexity of genome relationships between rice and wheat. *Functional and Integrative Genomics* **4:** 34-46.

Langridge, P., E.S. Lagudah, T.A. Holton, R. Appels, P.J. Sharp, and K.J. Chalmers. 2001. Trends in genetic and genome analyses in wheat: a review. *Australian Journal of Agricultural Research* **52:** 1043-1077.

Lee, S.J., G.A. Penner, and K.M. Devos. 1995. Characterization of loci containing microsatellite sequences among Canadian wheat cultivars. *Genome* **38:** 1037-1040.

Levinson, G. and G. Gutman. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* **4:** 203-221.

Lewontin, R.C. 1960. The Evolutionary Dynamics of Complex Polymorphisms. *Evolution* **14:** 458-472.

Lewontin, R.C. 1964. The interaction of selection and linkage. I. General considerations; Heterotic models. *Genetics* **49:** 49-67.

Lewontin, R.C. 1988. On Measures of Gametic Disequilibrium. *Genetics* **120:** 849-852.

Li, W., D.F. Zhang, Y.M. Wei, Z.H. Yan, and Y.L. Zheng. 2006. Genetic diversity of Triticum turgidum L. based on microsatellite markers. *Genetica* **42:** 397-402.

Lillemo, M., M.C. Simeone, and C.F. Morris. 2002. Analysis of puroindoline a and b sequences from *Triticum aestivum* cv. 'Penawawa' and related diploid taxa. *Euphytica* **126:** 321-331.

Liu, K., M. Goodman, S. Muse, J.S. Smith, E. Buckler, and J. Doebley. 2003. Genetic Structure and Diversity Among Maize Inbred Lines as Inferred From DNA Microsatellites. *Genetics* **165:** 2117-2128.

Liu, K. and S. Muse. 2004. PowerMarker: new genetic data analysis software. Free program distributed by the author over the internet from http://statgen.ncsu.edu/powermarker/.

Liu, S., X. Zhang, M.O. Pumphrey, R.W. Stack, B.S. Gill, and J.A. Anderson. 2006. Comple microlinearity among wheat, rice, and barley revealed by fine mapping of the genomic region harboring a major QTL for resistance to Fusarium head blight in wheat. *Functional and Integrated Genomics* **6:** 83-89.

Liu, X.M., A.K. Fritz, J.C. Reese, G.E. Wilde, B.S. Gill, and M.S. Chen. 2005. H9, H10, and H11 compose a cluster of Hessian fly-resistance genes in the distal gene-rich region of wheat chromosome 1AS. *Theoretical and Applied Genetics* **110:** 1473.

Lonjou, C., W. Zhang, A. Collins, W.J. Tapper, N. Maniatis, and N.E. Morton. 2003. Linkage disequilibrium in human populations. *Proceedings of the National Academy of Sciences* **100:** 6069-6074.

Lotti, C., S. Salvi, A. Pasqualone, R. Tuberosa, and A. Blanco. 2000. Integration of AFLP markers into an RFLP-based map of durum wheat. *Plant Breeding* **119:** 393-401.

Lupton, F.G.H. 1987. History of wheat breeding. In *Wheat Breeding Its scientific basis* (ed. F.G.H. Lupton), pp. 51-70. Chapman and Hall Ltd, Cambridge.

Maccaferri, M., M.C. Sanguineti, E. Noli, and R. Tuberosa. 2005. Population structure and long-range linkage disequilibrium in a durum wheat elite collection. *Molecular Breeding* **15:** 271-290.

Mackay, I. and W. Powell. 2007. Methods for linkage disequilibrium mapping in crops. *Trends in Plant Science* **12:** 57-63.

Macllwain, C. 2002. World leaders heap praise on human genome landmark. *Nature* **406:** 983-985.

Maraganore, D.M., M. de Andrade, T.G. Lesnick, K.J. Strain, M.J. Farrer, W.A. Rocca, P.V.K. Pant, K.A. Frazer, D.R. Cox, and D.G. Ballinger. 2005. High-Resolution Whole-Genome Association Study of Parkinson Disease. *Am J Hum Genet* **77:** 685-693.

Masojc, P. 2002. The application of molecular markers in the process of selection. *Cellular and Molecular Biology Letters* **7:** 499-509.

Massa, A.N., C.F. Morris, and B.S. Gill. 2004. Sequence diversity of Puroindoline-a, Puroindoline-b, and the Grain Softness Protein genes in *Aegilops tauschii* Coss. *Crop Science* **44:** 1808-1816.

McRae, A.F., J.C. McEwan, K.G. Dodds, T. Wilson, A.M. Crawford, and J. Slate. 2002. Linkage Disequilibrium in Domestic Sheep. *Genetics* **160:** 1113-1122.

Michelmore, R.W., I. Paran, and R.V. Kesseli. 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. *Proceedings of the National Academy of Sciences* **88:** 9828-9832.

Mohlke, K.L., E.M. Lange, T.T. Valle, S. Ghosh, V.L. Magnuson, K. Silander, R.M. Watanabe, P.S. Chines, R.N. Bergman, J. Tuomilehto, F.S. Collins, and M. Boehnke. 2001. Linkage Disequilibrium Between Microsatellite Markers Extends Beyond 1 cM on Chromosome 20 in Finns
10.1101/gr.173201. *Genome Res.* **11:** 1221-1226.

Morris, C.F. 2002. Puroindolines: the molecular genetic basis of wheat grain hardness. *Plant Molecular Biology* **48:** 633-647.

Morris, C.F., M. Lillemo, M.C. Simeone, M.J. Giroux, S.L. Babb, and K.K. Kidwell. 2001. Prevalence of Puroindoline Grain Hardness Genotypes among Historically Significant North American Spring and Winter Wheats. *Crop Sci* **41:** 218-228.

Nelson, J.C., M.E. Sorrells, A.E. Van Deynze, Y.H. Lu, M. Atkinson, M. Bernard, P. Leroy, J.D. Faris, and J.A. Anderson. 1995. Molecular Mapping of Wheat: Major Genes and Rearrangements in Homoeologous Groups 4, 5, and 7. *Genetics* **141:** 721-731.

Nordborg, M. 1997. Structured Coalescent Processes on Different Time Scales. *Genetics* **146:** 1501-1514.

Nordborg, M., J.O. Borevitz, J. Bergelson, C.C. Berry, J. Chory, J. Hagenblad, M. Kreitman, J.N. Maloof, T. Noyes, P.J. Oefner, E.A. Stahl, and D. Weigel. 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics* **30:** 190-193.

Ohashi, J., I. Naka, J. Patarapotikul, H. Hananantachai, G. Brittenham, S. Looareesuwan, A.G. Clark, and K. Tokunaga. 2004. Extended linkage disequilibrium surrounding the Hemoglobin E variant due to malarial selection. *American Journal of Human Genetics* **74:** 1198-1208.

Ozaki, K., Y. Ohnishi, A. Iida, A. Sekine, R. Yamada, T. Tsunoda, H. Sato, H. Sato, M. Hori, Y. Nakamura, and T. Tanaka. 2002. Functional SNPs in the lymphotoxin-[alpha] gene that are associated with susceptibility to myocardial infarction. **32:** 650-654.

Palaisa, K.A., M. Morgante, S. Tingey, and A. Rafalski. 2004. Long-range patterns of diversity and linkage disequilibrium surrounding the maize *Y1* gene are indicative of an asymmetric selective sweep. *Proceedings of the National Academy of Sciences* **101:** 9885-9890.

Palaisa, K.A., M. Morgante, M. Williams, and A. Rafalski. 2003. Contrasting Effects of Selection on Sequence Diversity and Linkage Disequilibrium at Two Phytoene Synthase Loci. *The Plant Cell* **15:** 1795-1806.

Parker, G., P. Fox, P. Langridge, K. Chalmers, B. Whan, and P. Ganter. 2002. Genetic diversity within Australian wheat breeding programs based on molecular and pedigree data. *Euphytica* **124:** 293-306.

Parker, G.D., K.J. Chalmers, A.J. Rathjen, and P. Langridge. 1998. Mapping loci associated with flour colour in wheat (*Triticum aestivum* L.). *Theoretical and Applied Genetics* **97:** 238-245.

Parsian, A., B. Racette, Z.H. Zhang, M. Rundle, and J.S. Perlmutter. 2004. Association of variations in monoamine oxidases A and B with Parkinson's disease subgroups. *Genomics* **83:** 454-460.

Paull, J.G., K.J. Chalmers, A. Karakousis, J.M. Kretschmer, S. Manning, and P. Langridge. 1998. Genetic diversity in Australian wheat varieties and breeding material based on RFLP data. *Theoretical and Applied Genetics* **96:** 435-446.

Payseur, B.A. and M.W. Nachman. 2000. Microsatellite variation and recombination rate in the human genome. *Genetics* **156:** 1285-1298.

Pestsova, E., M.W. Ganal, and M.S. Roder. 2000. Isolation and mapping of microsatellite markers specific for the D genome of bread wheat. *Genome* **43:** 689-697.

Pestsova, E.G., A. Borner, and M.S. Roder. 2006. Development and QTL assessment of Triticum aestivum-Aegilops tauschii introgression lines. *Theoretical and Applied Genetics* **112:** 634-647.

Peterson, R.F. 1965. *WHEAT Botany, Cultivation, and Utilization*. Interscience Publishers Inc, New York.

Prasad, M., N. Kumar, P.L. Kulwal, M.S. Roder, H.S. Balyan, H.S. Dhaliwal, and P.K. Gupta. 2003. QTL analysis for grain protein content using SSR markers and validation studies using NILs in bread wheat. *Theoretical and Applied Genetics* **106:** 659-667.

Pritchard, J.K. and M. Przeworski. 2001. Linkage Disequilibrium in Humans: Models and Data. *American Journal of Human Genetics* **69:** 1-14.

Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155:** 945-959.

Quarrie, S.A., A. Steed, C. Calestani, A. Semikhodskii, C. Lebreton, C. Chinoy, N. Steele, D. Pljevljakusic, E. Waterman, J. Weyen, J. Schondelmaier, D.Z. Habash, P. Farmer, L. Saker, D.T. Clarkson, A. Abugalieva, M. Yessimbekova, Y. Turuspekov, S. Abugalieva, R. Tuberosa, M.C. Sanguineti, P.A. Hollington, R. Aragues, A. Royo, and D. Dodig. 2005. A high-density genetic map of hexaploid wheat (Triticum aestivum L.) from the cross Chinese Spring x SQ1 and its use to compare QTLs for grain yield across a range of environments. *Theoretical and Applied Genetics* **110:** 865-880.

Rafalski, A. 2002. Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology* **5:** 94-100.

Rafalski, A. and M. Morgante. 2004. Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. *Trends in Genetics* **20:** 103-111.

Ravel, C., S. Praud, A. Murigneux, A. Canaguier, F. Sapet, D. Samson, F. Balfourier, P. Dufour, B. Chalhoub, D. Brunel, M. Beckert, and G. Charmet. 2006. Single-nucleotide polymorphism frequency in a set of selected lines of bread wheat (*Triticum aestivum* L.). *Genome* **49:** 1131-1139.

Reich, D.E., M. Cargill, S. Bolk, J. Ireland, P.C. Sabetli, D.J. Richter, T. Lavery, R. Kouyoumjian, S.F. Farhadian, R. Ward, and E.S. Lander. 2001. Linkage disequilibrium in the human genome. *Nature* **411:** 199-204.

Reif, J.C., X.C. Xia, A.E. Melchinger, M.L. Warburton, D.A. Hoisington, D. Beck, M. Bohn, and M. Frisch. 2004. Genetic diversity determined within and among CIMMYT maize populations of tropical, subtropical and temperate germplasm by SSR markers. *Crop Science* **44:** 326-334.

Remington, D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt, J. Doebley, S. Kresovich, M.M. Goodman, and E.S. Buckler IV. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences* **98:** 11479-11484.

Rhone, B., A.L. Raquin, and I. Goldringer. 2007. Strong linkage disequilibrium near the selected Yr17 resistance gene in a wheat experimental population. *Theoretical and Applied Genetics* **114:** 787-802.

Roder, M., X.-Q. Huang, and A. Borner. 2007. Fine mapping of the region on wheat chromosome 7D controlling grain weight. *Functional & Integrative Genomics* **Online First**.

Roder, M.S., V. Korzun, K. Wendehake, J. Plaschke, M.-H. Tixier, P. Leroy, and M.W. Ganal. 1998. A Microsatellite Map of Wheat. *Genetics* **149:** 2007-2023.

Ronin, Y., A. Korol, M. Shtemberg, E. Nevo, and M. Soller. 2003. High-Resolution Mapping of Quantitative Trait Loci by Selective Recombinant Genotyping. *Genetics* **164:** 1657-1666.

Rosenberg, N.A., T. Burke, K. Elo, M.W. Feldman, P.J. Freidlin, M.A.M. Groenen, J. Hillel, A. Maki-Tanila, M. Tixier-Boichard, A. Vignal, K. Wimmers, and S. Weigend. 2001. Empirical Evaluation of Genetic Clustering Methods Using Multilocus Genotypes From 20 Chicken Breeds. *Genetics* **159:** 699-713.

Rosenberg, N.A., J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, and M.W. Feldman. 2002. Genetic structure of human populations. *Science* **298:** 2381-2385.

Rostoks, N., S. Mudie, L. Cardle, J. Russell, L. Ramsay, A. Booth, J. Svensson, S. Wanamaker, H. Walia, E. Rodriguez, P. Hedley, H. Liu, J. Morris, T. Close, D. Marshall, and R. Waugh. 2005. Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. *Molecular Genetics and Genomics* **274:** 515-527.

Rozas, J. and R. Rozas. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15:** 174-175.

Sandhu, D. and K.S. Gill. 2002. Structural and functional organization of the '1S0.8 gene-rich region' in the Triticeae. *Plant Molecular Biology* **48:** 791-804.

Sardesai, N., J.A. Nemacheck, S. Subramanyam, and C.E. Williams. 2005. Identification and mapping of H32, a new wheat gene conferring resistance to Hessian fly. *Theoretical and Applied Genetics* **111:** 1167-1173.

Schlotterer, C. 2000. Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109:** 365-371.

Service, S.K., R.A. Ophoff, and N.B. Freimer. 2001. The genome-wide distribution of background linkage disequilibrium in a population isolate. *Hum. Mol. Genet.* **10:** 545-551.

Shepard, K.A. and M.D. Purugganan. 2003. Molecular population genetics of the Arabidopsis *CLAVATA2* region: The genomic scale of variation and selection in a selfing species. *Genetics* **163:** 1083-1095.

Shifman, S., M. Bronstein, M. Sternfeld, A. Pisante-Shalom, E. Lev-Lehman, A. Weizman, I. Reznik, B. Spivak, N. Grisaru, L. Karp, R. Schiffer, M. Kotler, R.D. Strous, M. Swartz-Vanetik, H.Y. Knobler, E. Shinar, J.S. Beckmann, B. Yakir, N. Risch, N.B. Zak, and A. Darvasi. 2002. A Highly Significant Association between a COMT Haplotype and Schizophrenia. *American Journal of Human Genetics* **71:** 1296-1302.

Smith, P.H., J. Hadfield, N.J. Hart, R.M. Koebner, and L.A. Boyd. 2007. STS markers for the wheat yellow rust resistance gene Yr5 suggest a NBS-LRR-type resistance gene cluster. *Genome* **50:** 259-265.

Soltis, D.E. and P.S. Soltis. 1999. Polyploidy: recurrent formation and genome evolution. *TREE* **14:** 348-352.

Somers, D.J., G. Fedak, J. Clarke, and W. Cao. 2006. Mapping of FHB resistance QTLs in tetraploid wheat. *Genome* **49:** 1586-1593.

Somers, D.J., P. Isacc, and K. Edwards. 2004. A high-density microsatellite consensus map for bread wheat (*Triticum aestivum* L.). *Theoretical and Applied Genetics* **109:** 1105-1114.

Somers, D.J., R. Kirkpatrick, M. Moniwa, and A. Walsh. 2003. Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. *Genome* **46:** 431-437.

Song, Q.J., J.R. Shi, S. Singh, E.W. Fickus, J.M. Costa, J. Lewis, B.S. Gill, R. Ward, and P.B. Cregan. 2005. Development and mapping of microsatellite (SSR) markers in wheat. *Theoretical and Applied Genetics* **110:** 550-560.

Squirrell, J., P.M. Hollingsworth, M. Woodhead, J. Russell, A.J. Lowe, M. Gibby, and W. Powell. 2003. How much effort is required to isolate nuclear microsatellites from plants? *Molecular Ecology* **12:** 1339-1348.

Stam, P. 1993. Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. *The Plant Journal* **5:** 739-744.

Tanksley, S.D., M.W. Ganal, J.P. Prince, M.C. de Vincente, M.W. Bonierbale, P. Broun, T.M. Fulton, J.J. Giovannoni, S. Grandillo, G.B. Martin, R. Messeguer, J.C. Miller, L. Miller, A.H. Paterson, O. Pineda, M.S. Roder, R.A. Wing, W. Wu, and N.D. Young. 1992. High density molecular linkage map of the tomato and potato genomes. *Genetics* **132:** 1141-1160.

Tanksley, S.D. and S.R. McCouch. 1997. Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* **277:** 1063-1066.

Tenaillon, M.I., M.C. Sawkins, L.K. Anderson, S.M. Stack, J. Doebley, and B.S. Gaut. 2002. Patterns of Diversity and Recombination Along Chromosome 1 of Maize (Zea mays ssp. mays L.). *Genetics* **162:** 1401-1413.

Tenaillon, M.I., M.C. Sawkins, A.D. Long, R.L. Gaut, J.F. Doebley, and B.S. Gaut. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). *PNAS* **98:** 9161-9166.

Tenesa, A., A.F. Wright, S.A. Knott, A.D. Carothers, C. Hayward, A. Angius, I. Perisco, G. Maestrale, N.D. Hastie, M. Pirastu, and P.M. Visscher. 2003. Extent of linkage disequilibrium in a Sardinian sub-isolate: sampling and methodological considerations. *Human Molecular Genetics* **13:** 25-33.

Thornsberry, J.M., M.M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E.S. Buckler IV. 2001. *Dwarf8* polymorphisms associate with variation in flowering time. *Nature Genetics* **28:** 286-289.

Thumma, B.R., M.F. Nolan, R. Evans, and G.F. Moran. 2005. Polymorphisms in Cinnamoyl CoA Reductase (CCR) Are Associated With Variation in Microfibril Angle in Eucalyptus spp.
10.1534/genetics.105.042028. *Genetics* **171:** 1257-1265.

Torada, A., M. Koike, K. Mochida, and Y. Ogihara. 2006. SSR-based linkage map with new markers using an intraspecific population of common wheat. *Theoretical and Applied Genetics* **112:** 1042-1051.

Vigouroux, Y., J.S. Jaqueth, Y. Matsuoka, O.S. Smith, W.D. Beavis, S.C. Smith, and J. Doebley. 2002. Rate and pattern of mutation at microsatellit loci in maize. *Molecular Biology and Evolution* **19:** 1251-1260.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, M. Kuiper, and M. Zabeau. 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* **23:** 4407-4414.

Wang, D.G., J.-B. Fan, C.-J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M.S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T.J. Hudson, R. Lipshutz, M. Chee, and E.S. Lander. 1998. Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science* **280:** 1077-1082.

Wang, W.Y.S., B.J. Barratt, D.G. Clayton, and J.A. Todd. 2005. Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics* **6:** 109-118.

Watanabe, N., A.S. Akond, and M.M. Nachit. 2006. Genetic mapping of the gene affecting polyphenol oxidase activity in tetraploid durum wheat. *Journal of Applied Genetics* **47:** 201-205.

Weil, C.F. 2002. Finding the crosswalks on DNA. *PNAS* **99:** 5763-5765.

Weiss, K.M. and A.G. Clark. 2002. Linkage disequilibrium and the mapping of complex human traits. *Trends in Genetics* **18:** 19-24.

Whitt, S.R., L.M. Wilson, M.I. Tenaillon, B.S. Gaut, and E.S. Buckler IV. 2002. Genetic diversity and selection in the maize starch pathway. *Proceedings of the National Academy of Sciences* **99:** 12959-12962.

William, H.M., R.P. Singh, J. Huerta-Espino, G. Palacios, and K. Suenaga. 2006. Characterization of genetic loci conferring adult plant resistance to leaf rust and stripe rust in spring wheat. *Genome* **49:** 977-990.

William, M., R.P. Singh, J. Huerta-Espino, S. Ortiz Islas, and D. Hoisington. 2003. Molecular marker mapping of leaf rust resistance gene *Lr46* and its association with stripe rust resistance gene *Yr29* in wheat. *Phytopathology* **93:** 153-159.

Williams, J., G. K., A. Kubelik, R., K.J. Livak, J.A. Rafalski, and S.V. Tingey. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research* **18:** 6531-6535.

Williams, K.J., J.G. Lewis, P. Bogacki, M.A. Pallotta, K.L. Willsmore, H. Kuchel, and H. Wallwork. 2003. Mapping of a QTL contributing to cereal cyst nematode tolerance and resistance in wheat. *Australian Journal of Agricultural Research* **54:** 731-737.

Williams, K.J., K.L. Willsmore, S. Olson, M. Matic, and H. Kuchel. 2006. Mapping of a novel QTL for resistance to cereal cyst nematode in wheat. *Theoretical and Applied Genetics* **112:** 1480-1486.

Wu, R. and Z.-B. Zeng. 2001. Joint Linkage and Linkage Disequilibrium Mapping in Natural Populaitons. *Genetics* **157:** 899-909.

Xu, X., G. Bai, B.F. Carver, G.E. Shaner, and R.M. Hunger. 2005. Molecular characterization of slow leaf-rusting resistance in wheat. *Crop Science* **45**.

Young, N.D., D. Zamir, M.W. Ganal, and S.D. Tanksley. 1988. Use of isogenic lines and simultaneous probing to identify DNA markers tightly linked to the *Tm-2a* gene in tomato. *Genetics* **120:** 579-585.

Yu, J. and E.S. Buckler. 2006. Genetic association mapping and genome organization of maize. *Current Opinion in Biotechnology*
Plant biotechnology/Food biotechnology **17:** 155-160.

Yu, J., S. Hu, J. Wang, G.K.-S. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, M. Cao, J. Liu, J. Sun, J. Tang, Y. Chen, X. Huang, W. Lin, C. Ye, W. Tong, L. Cong, J. Geng, Y. Han, L. Li, W. Li, G. Hu, X. Huang, W. Li, J. Li, Z. Liu, L. Li, J. Liu, Q. Qi, J. Liu, L. Li, T. Li, X. Wang, H. Lu, T. Wu, M. Zhu, P. Ni, H. Han, W. Dong, X. Ren, X. Feng, P. Cui, X. Li, H. Wang, X. Xu, W. Zhai, Z. Xu, J. Zhang, S. He, J. Zhang, J. Xu, K. Zhang, X. Zheng, J. Dong, W. Zeng, L. Tao, J. Ye, J. Tan, X. Ren, X. Chen, J. He, D. Liu, W. Tian, C. Tian, H. Xia, Q. Bao, G. Li, H. Gao, T. Cao, J. Wang, W. Zhao, P. Li, W. Chen, X. Wang, Y. Zhang, J. Hu, J. Wang, S. Liu, J. Yang, G. Zhang, Y. Xiong, Z. Li, L. Mao, C. Zhou, Z. Zhu, R. Chen, B. Hao, W.

Zheng, S. Chen, W. Guo, G. Li, S. Liu, M. Tao, J. Wang, L. Zhu, L. Yuan, and H. Yang. 2002. A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica). *Science* **296:** 79-92.

Zeder, M.A., E. Emshwiller, B.D. Smith, and D.G. Bradley. 2006. Documenting domestication: the intersection of genetics and archaeology. *Trends in Genetics* **22:** 139-155.

Zhu, Q., X. Zheng, J. Luo, B.S. Gaut, and S. Ge. 2007. Multilocus Analysis of Nucleotide Variation of Oryza sativa and Its Wild Relatives: Severe Bottleneck during Domestication of Rice 10.1093/molbev/msm005. *Mol Biol Evol* **24:** 875-888.

Zhu, Y.L., Q.J. Song, D.L. Hyten, C.P. Van Tassell, L.K. Matukumalli, D.R. Grimm, S.M. Hyatt, E.W. Fickus, N.D. Young, and P.B. Cregan. 2003. Single-nucleotide polymorphisms in soybean. *Genetics* **163:** 1123-1134.

Appendix A. **List of Australian hexaploid wheat germplasm used in estimating population structure and linkage disequilibrium.**

Germplasm Name	Source	Pedigree	Origin	Year
Andes	Tamworth	Kentana/Frontana//Mayo-48,	MEX	1969
Anza	Tamworth	Lerma-Rojo-64//Norin-10/Brevor/3/3*Andes-Enano,	MEX,USA:Calf	1971
Aroona	Waite	Ww-15/Raven,	AUS:SA	1981
Aus10894	Tamworth	Lv-Afg	AUS	1975
Banks	Waite	Pwth/(Sib)Condor//2*Condor,	AUS:NSW	1979
Barunga	Waite	Bt-Schomburgk/Molineux;Halberd/Aroona//3*Schomburgk/3/2*Molineux;	AUS:SA	1997
Beulah	Tamworth	Cook*2/Millewa//Tm-56,	AUS:Vic	1993
Bindawarra	Waite	Mexico-120/Koda//Raven,	AUS:SA	1980
Bluebird	Tamworth	Ciano-67(Sib)//Sonora-64/Klein-Rendidor/3/Ii-8156,	MEX	1969
Bobin	Tamworth	Thew/Steinwedel,	AUS:NSW	1925
Bowie	Tamworth	Renacimiento/Kenya-C-10862, Renacimiento//Kenya/Gular	USA:Tx	1953
Camm	Tamworth			
Cascades	Waite	Aroona*3//(Ausen-Vii-95)Tadorna/Inia-66,	AUS:WA	1994
CD87	Waite			
Ciano67	Tamworth	Pitic-62/(Sib)Chris//Sonora-64;	MEX	1967
Cocamba	Waite	Aus-10894/4*Condor,	AUS:Vic	1984
Condor	Waite	Ww-80/2*Ww-15,	AUS:NSW	1973
Cook	Waite	Timgalen/(Sib)Condor//Condor, Timgalen/2*Condor	AUS:Qld	1977
Cunningham	Waite	3-Ag-3/4*Condor//Cook,	AUS:Qld	1990
Dagger	Waite	Rac-111/Insignia,	AUS:SA	1983
Daruma	Tamworth	Lv-Gunma;(S)Anzeunbaengimil;	JPN	1900

Appendix A. **List of Australian hexaploid wheat germplasm used in estimating population structure and linkage disequilibrium.**

Germplasm Name	Source	Pedigree	Origin	Year
Dirk48	Tamworth	Gabo/4*Dirk;	AUS:SA	1951
Excalibur	Waite	Rac-177(Sr26)/Uniculm-492//Rac-311-S,	AUS:SA	1990
Federation	Waite	Yandilla/Purple-Straw, Purple-Straw/Yandilla, Yandilla-King/Purple-Straw, Purplestraw-14-A/Yandilla	AUS:NSW	1901
Festiguay	Waite	Festival/Uruguay-C-10837, Festival/(Ury)Ci-10837	AUS:NSW	1963
Frame	Tamworth	Molineux/3*Dagger,	AUS:SA	1997
Frontana	Tamworth	Fronteira/Mentana,	BRA:Rio	1930
Fultz	Tamworth	(S)Lancaster,	USA:Penn	1871
Gabo-Aus	Waite	Bobin(S)/(Tr.Dr)Gaza//(S)Bobin;Gular/(Tr.Dr)Gaza//Gular;Bobin*2/Gaza;Bobin-W-39//Bobin-W-39/Gaza;	AUS:NSW	1951
Gamenya	Waite	Kenya-117-A/2*Gabo//Mentana/6*Gabo, Gabo/3/Gabo*5/Mentana//Gabo*2/Kenya-117-A, Gabo/4/Gabo*5/Mentana//2*Gabo/3/Kenya-117-A,	AUS:NSW	1958
Gaza	Tamworth	Lv-Pal,	ISR	1973
Ghurka	Waite	Gallipoli/3/Currawa//Indian-4-E/Federation, Indian-H/Federation//Currawa, Zaff/Yandilla-King,	AUS:Vic	1924
Goldmark	Tamworth	Pavon-76(Sib)/Tm-56,	AUS:Vic	1998
H45	Tamworth			
Halberd	Waite	Scimitar/Kenya-C-6042//Bobin/3/Insignia-49,	AUS:SA	1969
Hartog	Waite	Vicam-71//Ciano-67(Sib)/Siete-Cerros-66/3/Kalyansona/Bluebird,	AUS:Qld	1982
Heron	Waite	Ranee/Doubbi//Ranee(R.D.R.)/3/4*Insignia-49,	AUS:NSW	1958

Appendix A. **List of Australian hexaploid wheat germplasm used in estimating population structure and linkage disequilibrium.**

Germplasm Name	Source	Pedigree	Origin	Year
Insignia	Waite	Ghurka/Ranee,	AUS:Vic	1946
Insignia49	Tamworth	Gabo/5*Insignia;	AUS:SA	1951
Janz	Waite	3-Ag-3/4*Condor//Cook,	AUS:Qld	1989
Kalyansona	Tamworth	Penjamo-62(Sib)/Gabo-55,	IND	1967
Katepwa	Waite	Neepawa*6/Rl-2938/3/Neepawa*6//Ci-8154/2*Frocor,	CAN:MB	1981
KenyaSupremo	Tamworth			
Kenya C6042	Tamworth			
Kite	Waite	Norin-10/Brevor(Sel.14)//4*Eureka-2/3/Thatcher/Agel/3*Falcon/4/Thatcher/Ag.El(T-A)//4*Falcon/5/Thatcher/Ag.El(T-A)//5*Falcon,	AUS:NSW	1973
Kloka	Tamworth	Rumkers-Erli/Kloka-309, (S)Weihenstephaner-43-48	DEU	1965
Kukrirac820	Waite			
LermaRojo	Tamworth	Lerma-50/Yaqui-48//Maria-Escobar*2/Supremo-211, Lerma-50/3/Yaqui-48/Maria-Escobar//Supremo-211	MEX	1955
Machete	Waite	Mec-3/2*Gabo(Rac-177)//Madden,	AUS:SA	1985
Meering	Waite	(S)Condor, Ww-80/2*Ww-15	AUS:Vic	1984
Mengavi	Tamworth	Eureka,Aus/Ci-12632//2*Gabo/3/Mentana/6*Gabo,	AUS:NSW	1958
Millewa	Waite	Sonora-64/Yaqui-50-Enano//Gaboto/Ii-8156,	AUS:Vic	1979
Molineux	Waite	Pitic-62/Festiguay//2*Warigal,	AUS:SA	1988
Norin10	Waite	Daruma/Fultz//Turkey-Red;	JPN	
Norin10Brevor	Waite	Norin-10/Brevor,		

Appendix A. **List of Australian hexaploid wheat germplasm used in estimating population structure and linkage disequilibrium.**

Germplasm Name	Source	Pedigree	Origin	Year
Olympic	Waite	Baldmin/Quadrat;	AUS:Vic	1956
OpataM85	Waite	Bluejay(Sib)/Jupateco-73;	MEX	1985
Orfed	Tamworth	Oro/Federation,	USA:Wash	1943
Oxley	Waite	Ww-15*2/Ww-30, Penjamo-62/4*Gabo-56//Tezanos-Pintos-Precoz/Nainari-60/4/2*Lerma-Rojo//Norin-10/Brevor-14/3/3*Andes, Penjamo-62/4*Gabo-56//Tezanos-Pintos-Precoz/Nainari-60/3/Ww-15,	AUS:Qld	1974
PavonS	Tamworth	Vicam-71//Ciano-67/Siete-Cerros-66/3/Kalyansona/Bluebird, Vicam-71/3/Ciano-67*2//Sonora-64/Klein-Rendidor/4/Siete-Cerros-66	MEX	1977
Pitic62	Waite	Yaktana-54//(Sel.26-1-C)Norin-10/Brevor,	MEX	1962
Rac177	Waite	Mec-3/2*Gabo;	AUS:SA	1977
Ranee	Waite	Indian-F/Federation,	AUS:Vic	1924
Raven	Waite	Mayo-48/Uruguay-1084//Orfed/3/4*Dirk-48, Mayo/4*Dirk-48	AUS:NSW	1963
Rosella	Waite	Farro-Lungo/Heron//2*Condor/3/(Sib)Quarrion, Farro-Lungo/Heron(19-Fm-74)//3*Condor/Ta-3-Pnb-3-P/Ww-33-G/Condor*2/Ww-33-B	AUS:NSW	1985
Sabre	Tamworth	Gabo/3/Nabawa/Dan//Dundee/4/Dundee/5/Kenya-C-6402;Gabo//Nabawa/Dan/3/Dundee/4/Dundee/Kenya-C-6042;	AUS:SA	1952
Scimitar	Tamworth	Nabawa/Egyptian-4,	AUS:SA	1930
Siete Cerros	Waite	Penjamo-62(Sib)/Gabo-55,	MEX	1966
Silverstar	Waite	Pavon-76(Sib)/Tm-56,	AUS:Vic	1998
Sonora64	Tamworth	Yaktana-54//Norin-10/Brevor/3/2*Yaqui-54,	MEX	1964
Spear	Waite	Rac-111/Insignia,	AUS:SA	1983

Appendix A. **List of Australian hexaploid wheat germplasm used in estimating population structure and linkage disequilibrium.**

Germplasm Name	Source	Pedigree	Origin	Year
Steinwedel	Tamworth	(S)Champlains-Hybrid, (S)Farmers-Friend	AUS:SA	1890
Stiletto	Waite	Veranopolis/3*Rac-177//3*Spear/3/Dagger,	AUS:SA	1993
Sunco	Waite	Sun-9-E-27*4/3-Ag-14//Ww-15/3/3*Cook,	AUS:NSW	1986
Sunstate	Tamworth	Hartog*4//Cook*5/Vpm-1,	AUS:NSW	1993
Sunvale	Tamworth	Cook*2/Vpm-1//3*Cook,	AUS:NSW	1994
Supremo	Tamworth	Surpresa//Hope/Mediterranean, Mediterranean/Hope(41-146)//Surpresa	USA:Tx	1948
Synthetic	Waite	Tr.Ti-Var.Typica/(Tr.Ta)Ae.Squarrosa-Var.Strangulata,	BGR	
Tasman	Waite	Torres/3/Gaboto/Siete-Cerros-66//Bluebird/Ciano-67;11-Ibwsn-45/Torres;	AUS:Qld	1993
Tatiara	Waite	Mexico-120/Koda//Raven/3/Mengavi/Siete-Cerros-66,	AUS:SA	1988
Timgalen	Tamworth	Aguilera/Kenya//Marroqui/3/Supremo/4/Gabo/5/Winglen, Aguilera/Kenya//Marroqui/Supremo/3/Gabo/4/Winglen	AUS:NSW	1967
Trident	Waite	Vpm-1/5*Cook//4*Spear, Vpm-1/4*Spear, Spear*4/Vpm-1,	AUS:SA	1993
Trintecino	Tamworth	Alfredo-Chaves-3.21/Alfredo-Chaves-4.21;	BRA	1936
TZPP	Tamworth			
UNICULM492	Tamworth			
Uraguay1064	Tamworth			
Veranapolis	Waite			
VPM1	Tamworth	Ae.Ve/Tr.Ca//3*Marne;	FRA	1981
WAREMEK	Tamworth	Mexico-120/Koda,	AUS:SA	1971
Warigal	Waite	Ww-15/Raven,	AUS:SA	1978
Wariquam	Tamworth	Mexico-120/Quadrat,	AUS:SA	1971

Appendix A. **List of Australian hexaploid wheat germplasm used in estimating population structure and linkage disequilibrium.**

Germplasm Name	Source	Pedigree	Origin	Year
Winglen	Tamworth	Kenya-C-6042/Gular//Winter-Minflor/3/Celebration;	AUS:NSW	1957
WW15	Waite	Lerma-Rojo-64//(Selection-14)Norin-10/Brevor/3/3*Andes-Enano,	AUS:NSW	1969
WW80	Waite	Penjamo-62/4*Gabo-56//Tezanos-Pintos-Precoz/Nainari-60;	AUS:NSW	
Yaktana54	Tamworth	Yaqui-48/Kentana-48//Frontana;	MEX	1954
Yaqui54	Tamworth	Yaqui-48/Timstein//Kenya-C-9906;Mayo-48/E-101//Timstein;Yaqui-50//Timstein/Kenya;	MEX	1954

Appendix B. List of UK hexaploid wheat germplasm used in estimating population structure and linkage disequilibrium.

Germplasm Name	Pedigree	Origin	Year
Adroit	Norman/Mercia//Moulin;	GBR	1992
Aintree	Sona 227*TB 306/45/2/10(=Jufy 1*Stella)	GBR:Eng	1981
Alta	Not Recorded	UK	1997
Altria	Gawain/Aquila//Recital;	FRA	1995
Andante	Moulin/D-172-6-4;	GBR:Eng	1992
Anduril	Sandown*Sicco	GBR:Eng	1981
Anfield	Svenno/Fasen//Svenno/Jufy-1;	GBR:Eng	1976
Angas	Schomburgk*3//Aroona/Moro;	AUS:SA	1991
Anglia	Marksman*Clement	UK	1987
Apollo	Maris-Beacon/Clement//Kronjuwel;Maris-Beacon/Kronjuwel;	DEU	1984
Apostle	Alcedo/Avalon//Moulin;	GBR:Eng	1980
Aquila	Tadorna/Carstens-854;	GBR:Eng	1972
Argent	Maris-Freeman/(Tl-365-A-25)Maris-Durin;	GBR	1979
Aristocrat	Rendezvous/Moulin//Mercia;	GBR	1992
Arminda	Carstens-854/Ibis,Deu;	NLD:Van-Der-Have,FRA:Benoist	1976
Arpege	Not Recorded	FRA	
Aslan	Brigand*Talent	UK	1982
Atou	Cappelle/Garnet;	FRA	1971
Avalon	Maris-Ploughman/Bilbo;Maris-Widgeon(Der)/Bilbo;Tjb-30-148/Bilbo;	GBR:Eng	1980
Avans	Tw-238-62/Kadett/Nemares;	SWE	1993
Avocet	Tjb-268-175/(Sib)Hobbit;	GBR:Eng	

Appendix B. **List of UK hexaploid wheat germplasm used in estimating population structure and linkage disequilibrium.**

Germplasm Name	Pedigree	Origin	Year
Axona	Hpg-522-66/Maris-Dove;	NLD	1983
Aztec	Not Recorded		
Baldus	Sicco/4/(Sel.)Sicco/3/N-66/Mgh-653//Kolibri;Sicco/4/(Sel.)Sicco//N-66/Mgh-653/3/Kolibri;	NLD	1992
Batavia	Brochis(Sib)/Banks;	AUS:Qld	1991
Beaufort	Rendezvous/Haven//Fresco;	GBR:Eng	1993
Beaver	Hedgehog/Norman//Moulin;	GBR:Eng	1989
Bersee	Hybrid des Allies x Vilmorin 23	FRA	
Bezostaya	Lutescens17 x Skorospelka2	RUS	
Bilbo	TJB 268 175 x Hobbit-sib	UK	1986
Blaze	Hussar/Beaver;	GBR	1998
Booty	Armada/Maris-Brigand;	GBR:Eng	1987
Bounty	Tjb-30-148/Tl-365-A-25;Maris-Ploughman/Durin;	GBR:Eng	1979
Boxer	Griffin/Rpb-181-70-D;	GBR:Eng	1987
Brigadier	Squadron/Rendezvous;	GBR	1992
Brigand	(S)Maris-Brigand;	GBR:Eng	1977
Brimstone	Tjb-54-218/Hobbit-30-2//Hustler;	GBR:Eng	1985
Brock	Hobbit-30-2/Talent;Talent/(Sib)Hobbit;	GBR:Eng	1985
Brutus	Not Recorded	GBR	1997
Bryden	Not Recorded	GBR	1997
Buchan	Beaver*Hussar	GBR	1995
Buster	Brimstone/Parade;	GBR	1992

Appendix B. **List of UK hexaploid wheat germplasm used in estimating population structure and linkage disequilibrium.**

Germplasm Name	Pedigree	Origin	Year
Cadenza	Axona/Tonic;Tonic/Axona;	GBR:Eng	1992
Camp Remy	Gu-362/Atou//Hardi;	FRA	1980
Capitole	Cappelle//(S-6)Hybride-80-3/Etoile-De-Choisy;	FRA	1964
Cappelle Desprez	Vilmorin-27/Hybride-Du-Joncquois;	FRA	1946
Carmen	Not available		
Carstens 8	Lv-Russian/Carstens-V//Minhardi/3/Carstens-Vi;Korchow/Carstens-V//Minhardi/3/Stamm/4/Minturki/Stamm//Kladener-Br/Stamm;Dummelweizen/Carstens-V//Carstens-Vi;	DEU:Lubeck	1952
Caxton	Moulin/Riband;	GBR	1996
Chablis	Jerico/Tonic;	GBR:Eng	1994
Champlein	Yga-Blondeau/Tadepi;Tadepi/Yga;White-Victoria/Chiddam-D-Automne-A-Epi-Rouge;	FRA	1959
Cheyenne	Crimean (CI 1435) Selection	USA	1977
Chinese Spring	CS/CNO.E//HORK/3/CS/CNO.E/4/2*CS	CHA	1999
Ciano 67	Pitic-62/(Sib)Chris//Sonora-64;	MEX	1967
Cinnabar	Hedgehog/Norman;	GBR	1985
Claire	WASP X FLAME	UK	1999
Consort	Riband(Sib)/Fresco/Riband;	GBR:Eng	1993
Corrigin	Tincurrin*2//Gamenya/Iassul;	AUS:WA	1989
Craftsman	Virtue/Maris-Huntsman;	GBR:Eng	1987
Cunningham	3-Ag-3/4*Condor//Cook;	AUS:Qld	1990
Dale	Not available	USA:Oreg	1900

Appendix B. **List of UK hexaploid wheat germplasm used in estimating population structure and linkage disequilibrium.**

Germplasm Name	Pedigree	Origin	Year
Daphne	Fan/Late-Gluyas;	AUS:SA	1912
Dean	Disponent/Norman;	GBR:Eng	1989
Demeter	(S)Mendel;	SWE	1950
Des Domes	K 8*Syekus	FRA	
Desprez 80	Vilmorin-23/Institut-Agronomique;	FRA	1934
Dicklow	(S)Surprise;(S)Californian-Club;	USA:Utah	1912
Durin	Vilmorin 29 x vogel 8058 2 x cappelle 4 x ci 12633 x 4X cappelle 2 x Heine 110 x cappelle 3 x nord	UK	1976
Dwarf A	Selected at F5 from the same cross as Hobbit	GBR	1989
Elland	(Highbury*Timmo)*(Sandown*Sicco)	GBR:Eng	1985
Encore	Apostle/Haven;	GBR:Eng	1993
Eureka	(Des Domes*Vilmorin 27)*(Hybride de Joncquois*Providence)	FRA	1973
Excalibur	Rac-177(Sr26)/Uniculm-492//Rac-311-S;	AUS:SA	1990
Favorite	Not Recorded		1999
Fenman	((Maris Ranger x Durin) x Maris Beacon) x Hobbit 'sib'	UK	1973
Festival	TH/VIL27//PQQ/HYBRIDE 40/5/TH/VIL27//FTU/3/CAP/4/CPN/3/TH/VIL27//FTU	FRA	1981
Flambeau	Talent/Norda(N-8-15-D-1)//Feuvert/3/Caton/Rivoli;	FRA	1992
Flame	Taurus/Moulin;	GBR	1995
Flanders	Champlein/Fd-2816-348;	FRA	1986
Flomar	Florence/Marquis;	USA:Wash	1933
Fresco	Moulin/Monopol;	GBR:Eng	1988
Galahad	Joss-Cambier/Durin//(Sib)Hobbit;Hobbit(Sib)//Durin/Joss-Cambier;	GBR:Eng	1983

Appendix B. **List of UK hexaploid wheat germplasm used in estimating population structure and linkage disequilibrium.**

Germplasm Name	Pedigree	Origin	Year
Garnet	Preston-A/Riga-M;Riga/Preston;Riga-M/Preston;	CAN:ON	1926
Gawain	Maris-Brigand//Maris-Huntsman/Durin;	GBR:Eng	1985
Hadrian	CB 296*Maris Templar	GBR	1987
Hardi	Cappelle-Desprez*2/Thatcher;	FRA	1969
Harlequin	Pelicano(Sib)/(Sib)Crane//(Sib)Tildillo/3/Pinguino(Sib)/Parana-66-270;	MEX	1978
Haven	Hedgehog/Norman//Moulin;	GBR:Eng	1988
Haydock	Sandon/Sicco;	GBR:Eng	1981
Hedgehog	Chile-13573-Lt/Mildress//Joss-Cambier/3/Cwm-100-1;	GBR	1976
Heine 7	Hybrid a courte paille x Svalofs Kronen	DEU	
Herald	Tjb-54-335/(Sib)Hobbit;	GBR	1976
Hereward	Norman/Disponent;Disponent/Norman;Norman(Sib)/Disponent;	GBR:Eng	1989
Hickory	Hunter/Caldwell;	USA:Ark	1993
Highbury	Jufy-I/Svenno(306)//Sona-227;Sona-227//Svenno/Jufi-I;Svenno/Jufy-I//Sona-227;	GBR:Eng	1968
Hobbit	Professeur-Marchal//Marne-Desprez/Vg-9144/4/Ci-12633/4*Cappelle-Desprez//Heines-110/Cappelle-Desprez/3/Nord-Desprez;	GBR	1974
Holdfast	Yeoman/White-Fife;	GBR:Eng	1936
Hope	Yaroslav-Emmer/Marquis;Vernal-Emmer(Tr.Dm)/Marquis;	USA:Sth Dkta	1927
Hornet	Norman/Hedgehog;	GBR:Eng	1986
Hotspur	Highbury//Sirius/Ciano-67;	GBR:Eng	1981
Hunter	Apostle/Haven;	GBR:Eng	1991
Hussar	Squadron/Rendezvous;	GBR	1991

Appendix B. **List of UK hexaploid wheat germplasm used in estimating population structure and linkage disequilibrium.**

Germplasm Name	Pedigree	Origin	Year
Hustler	Maris Huntsman*Durin 25	UK	1974
Hyper	Pacific-Bluestem/Prelude;	USA:Wash	1929
Ibis	Merlin*(Heine 7*((Teutonen*DHE 516)*((Chinese 165*Pansar 3)*Heine 4)))	DEU	1986
Indian	Not Recorded	USA	1994
Iona	Hobbit(Sib)-2/Ploughman;Ploughman/Tl-363-30-5;	GBR:Eng	1979
Jena	Not available		
Joss Cambier	Heines-Vii/Tadepi//Cappelle-Desprez;Cambier-194/Tadepi//Cappelle-Desprez;194-C.H.Vii/Tadepi//Cappelle;	FRA	1966
Jufy 1	Jubile/Fylgia-De-Printemps;Jubilegem/Fylgia,S;	BEL	1954
Kinsman	((CI 12633*C.Desprez)*(Hybrid 46*C.Desprez)*Professeur Marchal)*Maris	GBR	1975
Koga 2	(Heines Kolben*Garnet)*(Heines Kolben*Raeckes White Chaff)	DEU	1955
Lark	Not available	AUS	1992
Leo	Kristall/Marksman;	GBR:Eng	1986
Lillimur	Ibwsn-12-127/Asn-3-61;	AUS:Vic	1990
Little Joss	Squareheads-Master/Ghirka;	GBR:Eng	1908
Longbow	Tjb-268-175/Hobbit;	GBR:Eng	1981
Lutescens 055 74	Siete-Cerros/Saratovskaya-29;	RUS:Tyumen	1977
Maestro	(Selkirk*Cappelle Desprez)*Hobbit-sib	GBR:Eng	1976
Mandate	Norman/Hedgehog;	GBR:Eng	1980
Mantle	Cappelle/H-2596//6003;Tjb-30-148/Durin//Virtue/3/Marksman;	GBR:Eng	1974
Mara	Autonomia A*Aquila sib	ITA	
Mardler	Maris Ranger*(Durin 25*Maris Huntsman)	GBR	1976

Appendix B. **List of UK hexaploid wheat germplasm used in estimating population structure and linkage disequilibrium.**

Germplasm Name	Pedigree	Origin	Year
Maris Argent	Maris-Freeman/Tl-365-A-25;	GBR:Eng	1976
Maris Beacon	Ci-12633/5*Cappelle-Desprez/3/Hybrid-46/Cappelle-Desprez//Professeur-Marchal;Cappelle*4/Hybrid-46//2*Professeur-Marchal;Ci-12633/5*Cappelle//Hybrid-46/Cappelle/3/2*Professeur-Marchal;	GBR:Eng	1968
Maris Butler	Koga-Ii/Hybrid-46;	GBR:Eng	1972
Maris Dove	H-8810-47//Heines-Koga-Ii;	GBR:Eng	1971
Maris Ensign	Breustedts-Teutonen/(W)Cappelle-Desprez;	GBR:Eng	1968
Maris Envoy	Ci-12633/5*Cappelle-Desprez//Heines-110/Cappelle/3/Nordeste/4/Viking;	GBR:Eng	1974
Maris Freeman	Maris-Widgeon/Maris-Ranger;	GBR:Eng	1974
Maris Fundin	Vilmorin-29/Vogel-8058//Cappelle-Desprez/4/Ci-12633/4*Cappelle-Desprez//Heines-110/Cappelle-Desprez/3/Nord-Desprez;	GBR:Eng	1975
Maris Halberd	Els/4*Jufy-I;	GBR:Eng	1976
Maris Huntsman	Ci-12633/5*Cappelle-Desprez//Hybrid-46/Cappelle-Desprez/3/2*Professeur-Marchal;Cappelle*4/Hybrid-46//2*Professeur-Marchal;	GBR:Eng	1971
Maris Nimrod	Ci-12633/Yeoman//5*Cappelle/3/Cappelle/Hybrid-46/4/2*Professeur-Marchal;Cappelle*4/Hybrid-46//2*Professeur-Marchal;	GBR:Eng	1971
Maris Pinion	Minister/Els;	GBR:Eng	1975
Maris Ploughman	Cappelle*4/Hybrid-46//2*Maris-Widgeon/3/Viking;Ci-12633/Hybrid-46//Maris-Widgeon;	GBR:Eng	1972
Maris Ranger	Heines-Peko/Cappelle-Desprez;	GBR:Eng	1968
Maris Settler	Professeur-Marchal/Cappelle;	GBR:Eng	1972
Maris Teal	Hybrid-46/Minister;Hybrid-46/Ministre;	GBR:Eng	1972

Appendix B. **List of UK hexaploid wheat germplasm used in estimating population structure and linkage disequilibrium.**

Germplasm Name	Pedigree	Origin	Year
Maris Templar	Ci-12633/5*Cappelle-Desprez//Heines-110/Cappelle-Desprez/3/Nord-Desprez/4/Viking;Heines-110/Cappelle-Desprez//Ci-12633/5*Cappelle-Desprez/3/Nord-Desprez/4/Viking;	GBR:Eng	1968
Maris Totem	Hybrid-46/Minister//Nord-Desprez;	GBR:Eng	1969
Maris Widgeon	Holdfast/Cappelle-Desprez;	GBR:Eng	1964
Marksman	Maris Huntsman*TL 365a/25(=Durin)	GBR	1977
Marne Desprez	Hybride-Du-Joncquois/Vilmorin-27//Hybride-Du-Joncquois/P.L.M.1;	FRA	1954
Marquis	Hard-Red-Calcutta/Red-Fife;	CAN:ON	1907
Mercia	Talent/Virtue//Flanders;	GBR:Eng	1984
Ministre	Benoist 40*Professeur Delos	BEL	
Mithras	Tjb-268-175/(Sib)Hobbit;	GBR:Eng	1980
Monopol	Pantus/Admiral,Deu;	DEU	1975
Morell	Not Recorded	GBR	1997
Moulin	Yecora-70/Ciano-67(Cb-306-Y-70)//Maris-Widgeon/3/Maris-Hobbit;	GBR,FRA	1985
N.S. Rana	Not Recorded	YUG	1999
Newbury	(TB 306*Sona 227)*(Sirius*Ciano 67)	GBR	1981
Newmarket	Tw-161/Maris-Huntsman;	GBR:Eng	1977
Norman	Tjb-268-175/(Sib)Hobbit;	GBR:Eng	1981
Norseman	Not Recorded	GBR	1988
Nsl92-5719	Not available		
Oregon Zimmerman	(S)Surprise;	USA:Oreg	1921

Appendix B. **List of UK hexaploid wheat germplasm used in estimating population structure and linkage disequilibrium.**

Germplasm Name	Pedigree	Origin	Year
Orqual	Capitole/Moisson//Horace/3/Thesee;	FRA	1991
Ostara	Not Recorded	GBR	1997
Pastiche	Jena/Norman;Iena/Norman;	GBR:Eng	1988
Patience	Stuart*Galahad	GBR	1986
Peacock	Tw-161/Maris-Huntsman(Tw-275)//Norman;	GBR:Eng	1985
Peko	Peragis*Heines Kolben	DEU	
Poros	Stamm-3151-45/(Hadmerslebener-Iv)Trumpf;	DEU:Hohenthurm	1966
Professeur Marchal	Rimpaus-Bastard-Ii/Professeur-Delos//Professeur-Delos/Hybride-Du-Joncquois;Hybride-Du-Joncquois/Proffesseur-Delos//Bastard-Ii/Professeur-Delos;	BEL	1957
Recital	R-267(Mex)/4/V-81-12/Heine-Vii//Nordeste/Tadorna/3/9369;9369/R-267;	FRA	1986
Reeves	Bodallin//Gamenya/Inia-66;	AUS:WA	1989
Renard	Hobbit-30-2/Talent;	GBR:Eng	1987
Rendezvous	Vpm-1/(Sib)Hobbit//Virtue;	GBR:Eng	1985
Renown	Squareheads-Master/(Swe)Squarehead;	GBR:Eng	1960
Rialto	Haven(Sib)/(Sib)Fresco;	GBR:Eng	1993
Riband	Norman/Tw-275;	GBR:Eng	1987
Ritmo	Hobbit/Line-1320//Wizard/3/Marksman/Virtue;Hobbit//Line-1320/Wizard/3/Marksman/4/Virtue;	NLD	1990
Samson	Sinew/Improved-Fife;	AUS:NSW	1899
Sandown	Sona-227/Hobbit;Hobbit/Sona-227;	GBR:Eng	1980
Sarsen	Marksman/Armada;	GBR:Eng	1987

Appendix B. **List of UK hexaploid wheat germplasm used in estimating population structure and linkage disequilibrium.**

Germplasm Name	Pedigree	Origin	Year
Sava	Fortunato*2/(Ci-13170)Redcoat;	YUG:Serbia	1967
Selkirk	Mcmurachy/Exchange//3*Redman,Can;	CAN:MB	1953
Sentry	Maris-Ploughman/(Sib)Hobbit;	GBR:Eng	1979
Shango	Fresco/Tiresius;	GBR	1994
Shiraz	Jerico/Axona;	GBR:Eng	1994
Shire	Ploughman/Tl-36.30.6;	GBR	1979
Shrike	Ww-15/M-1238-2//Kite/3/Ww-15/4/Condor/5/Ww-33-G;	AUS:NSW	1990
Sicco	Ring//Opal/Selkirk;	NLD	1973
Sickle	Brigand*(Maris Huntsman*TW 161)	GBR	1985
Sideral	Arminda/Festival;	FRA	1990
Soissons	Iena(Jena)/(Hybride-Naturel)Hn-35;	FRA	1987
Solitaire	Sandown/Sicco;	GBR:Eng	1985
Sona 227	Penjamo-62(Sib)/Gabo-55;	IND	1967
Spark	Moulin/Tonic;	GBR	1991
Sponsor	Not available	FRA	1994
Sportsman	Maris Envoy*Durin	GBR	1976
Steadfast	Little-Joss/Victor;	GBR:Eng	1952
Stella	(Desprez 80*Professeur Delos)*(Bastard 2*Professeur Delos)	BEL	
Sumei-3	Not available		
Sunbri	Cook*2/Vpm-1//3*Cook;	AUS:NSW	1990
Supreme	(Selection from) Red Bobs	USA	1994
Surprise	Chile-Club/Michigan-Club;	USA:Wash	1899

Appendix B. **List of UK hexaploid wheat germplasm used in estimating population structure and linkage disequilibrium.**

Germplasm Name	Pedigree	Origin	Year
Synthetic	Tr.Ti-Var.Typica/(Tr.Ta)Ae.Squarrosa-Var.Strangulata;	GBR	1999
T. Spelta – Grey	Not Recorded		1999
Tadorna	(((Chinese 166*Panzer 3)*Heine 7)*(Teutonen*Hindukush 516)*Heine 7)*Me	DEU	1982
Talent	Champlein/3/Thatcher/Vilmorin-27//Fortunato;	FRA	1973
Teutonen	Heines Kolben*R 1004(=Noe*Squarehead)	DEU	1952
Texel	Capitole/Moisson//Horace/3/B-8111;	FRA	1992
Thatcher	Marquis/(Tr.Dr)Iumillo//Marquis/Kanred;	USA:MN	1934
Timstein	Steinwedel/Tr.Ti;Steinwedel/Tr.Dr;Bobin(S)/(Tr.Dr)Gaza//Bobin;Steinwedel/Gaza;	USA:MN	1939
Tipstaff	Maris-Ranger/Durin//Maris-Huntsman;	GBR	1976
Token	Bounty,Gbr/Armada//Flanders;Flanders/Bounty//Armada;	GBR:Eng	1985
Touchstone	Norman*Musket	GBR	1997
Tremie	S-32/Moulin;Se-32/Moulin;	FRA	1992
Troy	Tas-894-5-3/Sicco;	GBR	1993
Tsengrain	Not available		
Turnpike	TGS 335/1/191*QY 7/73/2/8	GBR	1985
Veritas	Not Recorded	GBR	1997
Villein	Maris-Ploughman/Hobbit-Sib-1;	GBR:Eng	1978
Vilmorin 27	Dattel//Japhet/Parsel/3/Hatif-Inversable/Bon-Fermier;Dattel/Allies//Hatif-Inversable/Bon-Fermier;	FRA	1928
Vilmorin 29	Vilmorin-23/Allies;Vilmorin-23/Hybride-Des-Allies;	FRA	1929
Virtue	Maris-Huntsman/Maris-Durin;	GBR:Eng	1979
Vivant	Boxer/Gawain;	GBR:Eng	1991

Appendix B. **List of UK hexaploid wheat germplasm used in estimating population structure and linkage disequilibrium.**

Germplasm Name	Pedigree	Origin	Year
Wembley	Hobbit/Sona-227//Sicco;Sandown/Sicco;	GBR:Eng	1985
Wizard	Hobbit(Sib)*2/Maris-Freeman;Maris-Freeman/Tl-363-30-5;	GBR:Eng	1983
Yeoman	Browick/Red-Fife;	GBR:Eng	1916