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

TABLE OF CONTENTS ..... **II**

LIST OF TABLES ..... **V**

LIST OF FIGURES ..... **VIII**

LIST OF ABBREVIATIONS AND ACRONYMS ..... **XIII**

ABSTRACT ..... **XVI**

DECLARATION ..... **XVII**

ACKNOWLEDGEMENTS ..... **XVIII**

CHAPTER 1: GENERAL INTRODUCTION **1**

1.1 AN INTRODUCTION TO WHEAT 1

1.2 A COMPLEX GENOME 1

1.3 GENETIC TOOLS 4

1.4 MOLECULAR MARKERS 5

1.4.1 RFLPs *6*

1.4.2 RAPDs *7*

1.4.3 AFLPs *9*

1.4.4 SSRs *10*

1.4.5 SNPs *13*

1.5 IDENTIFYING AND LOCALIZING TRAITS OF INTEREST 15

1.6 ASSOCIATION MAPPING BASED ON LINKAGE DISEQUILIBRIUM 17

1.7 EXTENT AND PATTERNS OF LINKAGE DISEQUILIBRIUM IN HUMAN GENETICS 20

1.8 EXTENT OF LINKAGE DISEQUILIBRIUM IN PLANTS 22

1.9 MEASURING LINKAGE DISEQUILIBRIUM 26

1.10 FACTORS AFFECTING LINKAGE DISEQUILIBRIUM 29

1.11 PHYSICAL FACTORS AFFECTING LD 29

1.11.1 Mutations and recombinations *29*

1.12 POPULATION GENETIC FACTORS AFFECTING LD 31

1.12.1 Selection *31*

1.12.2 Population dynamics *33*

1.12.3 Mating systems *36*

1.13 SUMMARY 38

CHAPTER 2: SEQUENCE BASED ANALYSIS OF LINKAGE DISEQUILIBRIUM IN WHEAT **39**

2.1 INTRODUCTION 39
2.2 MATERIALS AND METHODS 42
2.2.1 Plant material 42
2.2.2 Nucleic acid isolation 44
2.2.3 PCR amplification of Pina and Pinb genes 45
2.2.4 Fragment purification and sequencing of Pina and Pinb genes 46
2.2.5 Data analysis 47
2.3 RESULTS 48
2.3.1 Sequence evaluation of Pina and Pinb genes from hexaploid wheat 48
2.3.2 LD across the hardness locus of Ae. tauschii 54
2.4 DISCUSSION 62

CHAPTER 3: POPULATION STRUCTURE WITHIN EXPERIMENTAL HEXAPLOID WHEAT POPULATIONS 67

3.1 INTRODUCTION 67
3.2 MATERIALS AND METHODS 69
3.2.1 Plant material 69
3.2.2 Microsatellite markers 70
3.2.3 PCR amplification microsatellite markers in hexaploid wheat 71
3.2.4 ABI Prism 3700 DNA analyzer 73
3.2.5 Microsatellite analysis 73
3.2.6 Cluster analysis 74
3.3 RESULTS 76
3.3.1 Establishing program parameters 76
3.3.2 Deciding on the most appropriate value of K 85
3.3.3 Identifying sub-populations in Australian and UK data sets 85
3.4 DISCUSSION 89

CHAPTER 4: GENOME WIDE LINKAGE DISEQUILIBRIUM IN HEXAPLOID WHEAT .. 97

4.1 INTRODUCTION 97
4.2 MATERIALS AND METHODS 99
4.2.1 Plant material and DNA isolation 99
4.2.2 Microsatellite Analysis 99
4.2.3 Linkage Disequilibrium Data Analysis 99
4.3 RESULTS 101
4.3.1 Extensive LD within South Australian hexaploid wheat germplasm 101
4.3.2 LD in the absence of population stratification 104
4.3.3 LD in the absence of rare alleles 109

4.3.4 Examination of LD within linkage groups and genomes *111*
4.3.4.1 LD within linkage groups 111
4.3.4.2 LD within genomes 117
4.3.4.3 LD across the Group 7 chromosomes 119
4.3.5 LD in a large European Population *124*
4.4 DISCUSSION 130
CHAPTER 5: GENERAL DISCUSSION 138
LITERATURE CITED 144
APPENDIX A. XXIII
APPENDIX B XXIX

List of Tables

Table 1.1 ……..…….4

A selection of genetic maps developed in *T. aestivum* sp. using a variety of molecular marker systems.

Table 1.2 ……..…...6

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

Table 1.3 ……..…...23

Extent of linkage disequilibrium in plant studies.

Table 2.1……..……43

List of the Australian hexaploid wheat accessions used in evaluating sequence level LD in the Pina and Pinb genes.

Table 2.2 …….45

Gene specific primers used to amplify the full length Pina and Pinb genes in the hexaploid wheat varieties outlined in Table 2.1

Table 2.3 …….56

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

Table 3.1 ……....70

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

Table 3.2 ……72

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

Table 3.3 ……..……...75

Five experiments set up in Structure in order to determine the most appropriate parameter settings for the program.

Table 3.4 ……....88

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

Table 3.5 ……....92

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

Table 4.1……102

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).

Table 4.2 ……..…….125

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).

List of Figures

Figure 1.1……2

Flow chart illustrating the origin of hexaploid wheat, Triticum aestivum.

Figure 1.2……..37

Schematic representations of the effects of LD on association mapping strategies from (Rafalski 2002).

Figure 2.1 ……49

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).

Figure 2.2 ……50

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).

Figure 2.3 ……52

Consensus sequence of the Pina gene as determined from sequence analysis of 44 Australian hexaploid wheat lines.

Figure 2.4 ……53

Pinb gene consensus sequences as determined from sequence analysis of 62 Australian hexaploid wheat lines.

Figure 2.5……..55

Consensus sequence of 6 haplotypes of the Pina gene as determined from sequence analysis of 50 Ae. tauschii wheat lines**.**

Figure 2.6……..58

Consensus sequence of 4 haplotypes of the Pinb gene as determined from sequence analysis of 50 Ae. tauschii wheat lines.

Figure 2.7……..60

Scatter plot of LD (r2) as a function of distance in base pairs across the 447 bases in the Pinb gene of 50 Ae. tauschii lines.

Figure 2.8……..61

Scatter plot of LD (r2) as a function of distance in base pairs across the hardness locus of 50 Ae. tauschii lines.

Figure 3.1……..78

Experiment 1. Scatter plot of the log probability of the data versus the estimated K values (1 to 23) using 90 Australian wheat lines.

Figure 3.2……..……79

Experiment 2. Scatter plot of the log probability of the data versus the estimated K values (1 to 23) using 90 Australian wheat lines.

Figure 3.3……..80

Experiment 3. Scatter plot of the log probability of the data versus the estimated K values (1 to 23) using 90 Australian wheat lines.

Figure 3.4……..82

Experiment 4. Scatter plot of the log probability of the data versus the estimated K values (1 to 23) using 90 Australian wheat lines.

Figure 3.5……..83

Experiment 5. Scatter plot of the log probability of the data versus the estimated K values (1 to 23) using 90 Australian wheat lines.

Figure 3.6……..84

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.

Figure 3.7……..87

Estimation of population structure for the Australian (A) and UK (B) data sets for population estimations K = 3.

Figure 4.1……..103

Scatter plot of LD (D') vs marker distance (cM) for all Australia wheat lines.

Figure 4.2……..105

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……...……...106

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).

Figure 4.4……108

Scatter plot of LD (D') vs marker distance (cM) for the largest Australian sub-population (22 lines) as determined through Structure analysis.

Figure 4.5……112

Scatter plot of LD (D') vs marker distance (cM) for Australian wheat lines in the absence of population structure and rare alleles pooled.

Figure 4.6……113

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.

Figure 4.7……118

Scatter plot of LD (D') vs marker distance (cM) for genomes A, B, and D using the whole Australian wheat data set.

Figure 4.8……120

Bar graph illustrating the distribution of average D' values of each genome (A, B, and D).

Figure 4.9……121

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.

Figure 4.10……..122

Bar graph illustrating the distribution of average D' values of the group 7 chromosomes (7A, 7B, and 7D).

Figure 4.11……..123

Bar graph illustrating the distribution of percent significant pairwise comparisons of the group 7 chromosomes (7A, 7B, and 7D).

Figure 4.12…….….127

Scatter plot of LD (D') vs marker distance (cM) for whole UK data set.

Figure 4.13……..128

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……..129

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 data set.

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! ☺