An Evaluation of the Patterns of Nucleotide Diversity and Linkage Disequilibrium at the Regional Level in *Hordeum vulgare*.

A thesis presented for the degree of Doctor of Philosophy

by

Katherine Selby Caldwell

School of Agriculture and Wine
The University of Adelaide

Genome Dynamics Program
The Scottish Crop Research Institute

June 2004

LIST OF TABLES

Table 4.2.

Estimates of nucleotide polymorphism within different germplasm samples across the *hina* gene region. Number of haplotypes, number of mutations (η), and gene length were calculated after the omission of indels. Length of regions including indels are denoted in parentheses. Tajima's D statistic significance levels are denoted: * P<0.5 and **P<0.1.

Table 4.3.

Estimates of nucleotide polymorphism within different germplasm samples across the *hinb-1* gene region. Number of haplotypes, number of mutations (η), and gene length were calculated after the omission of indels. Length of regions including indels are denoted in parentheses. Tajima's D statistic significance levels are denoted: * P<0.5 and **P<0.1.

Table 4.4.

Estimates of nucleotide polymorphism within different germplasm samples across the *hinb-2* gene region. Number of haplotypes, number of mutations (η), and gene length were calculated after the omission of indels. Length of regions including indels are denoted in parentheses. Tajima's D statistic significance levels are denoted: * P<0.5 and **P<0.1.

Table 4.5.

Estimates of nucleotide polymorphism within different germplasm samples across the *PG2* gene region. Number of haplotypes, number of mutations (η), and gene length were calculated after the omission of indels. Length of regions including indels are denoted in parentheses. Tajima's D statistic significance levels are denoted: * P<0.5 and **P<0.1.

Table 4.6.

Estimated levels of diversity for each gene region within the different barley genepools and the level of diversity (%) within the cultivated material with respect to the diversity levels within the landrace and wild samples.

Table 4.7.

McDonald-Krietman Test statistics and significance values.

Table 4.8.

HKA Test statistics and significance values. Test was performed using all sites (red), silent sites (black), and synonymous sites (blue).

LIST OF FIGURES

Figure 3.4.

Structure of the genes located within the barley contig, the colinear rice region, and their closest *Arabidopsis* homologs (Table 3.1). Intron phase is indicated by the number above each intron.

Figure 4.1.

Multiple alignment of a 1802 bp region of *GSP*. Only the positions of polymorphisms are indicated. Positions are labeled by both the distance away from the A in the start codon (position 1) and by the position in the consensus of the alignment. The haplotype (H) and haplotype group (HG) designations, the number of lines within each haplotype, and haplotype frequencies (f) are included. Positions of nonsynonymous change are indicated (*). Positions fixed within a given subset although polymorphic in the sample as a whole have been grayed.

Figure 4.2.

Multiple protein sequence alignment of *GSP*.

Figure 4.3.

Multiple alignment of a 1475 bp region of *hina*. Only the positions of polymorphisms are indicated. Positions are labeled by both the distance away from the A in the start codon (position 1) and by the position in the consensus of the alignment. The haplotype (H) and haplotype group (HG) designations, the number of lines within each haplotype, and haplotype frequencies (f) are included. Positions of nonsynonymous change are indicated (*). Positions fixed within a given subset although polymorphic in the sample as a whole have been grayed.

Figure 4.4.

Multiple protein sequence alignment of *hina*.

Figure 4.5.

Multiple alignment of a 1582 bp region of *hinb-1*. Only the positions of polymorphisms are indicated. Positions are labeled by both the distance away from the A in the start codon (position 1) and by the position in the consensus of the alignment. The haplotype (H) and haplotype group (HG) designations refer to the *hinb* duplication region as a whole. Independent *hinb-1* haplotypes are indicated by Hb1_##. The number of lines within each haplotype and haplotype frequencies (f) are included. Positions of

nonsynonymous change are indicated (*). Positions fixed within a given subset although polymorphic in the sample as a whole have been grayed.

Figure 4.6.

Multiple alignment of a 1671 bp region of *hinb-2*. Only the positions of polymorphisms are indicated. Positions are labeled by both the distance away from the A in the start codon (position 1) and by the position in the consensus of the alignment. The haplotype (H) and haplotype group (HG) designations refer to the *hinb* duplication region as a whole. Independent *hinb-2* haplotypes are indicated by Hb2_##. The number of lines within each haplotype and haplotype frequencies (f) are included. Positions of nonsynonymous change are indicated (*). Positions fixed within a given subset although polymorphic in the sample as a whole have been grayed. The gap after consensus position 3262 indicates the end of the duplicated region.

Figure 4.7.

Multiple protein sequence alignment of *hinb-1*.

Figure 4.8.

Multiple protein sequence alignment of *hinb-2*.

Figure 4.9.

Multiple alignment of a 594 bp region of *PG2* exon 3. Only the positions of polymorphisms are indicated. Positions are labeled by both the distance away from the A in the start codon (position 1) and by the position in the consensus of the alignment. The haplotype (H) and haplotype group (HG) designations, the number of lines within each haplotype, and haplotype frequencies (f) are included. Positions of nonsynonymous change are indicated (*). Positions fixed within a given subset although polymorphic in the sample as a whole have been grayed.

Figure 4.10.

Multiple protein sequence alignment of *PG2*.

Figure 4.11.

Plots of the levels of diversity for each gene region with respect to their physical location along the sequenced region of barley. Results for all three sample sets are shown: cultivated (blue), landrace (pink), and wild (yellow). Observations exemplify the utility

of partitioning sequence data to acquire a more comprehensive understanding of the evolutionary history of the individual genes and the genomic region as a whole.

Figure 5.1.
Plots of LD (r^2) as a function of distance (bp) between informative ($f > 0.1$) polymorphic sites in four different gene regions in the cultivated sample.

Figure 5.2.
Plots of LD (r^2) as a function of distance (bp) between informative ($f > 0.1$) polymorphic sites in four different gene regions in the landrace sample.

Figure 5.3.
Plots of LD (r^2) as a function of distance (bp) between informative ($f > 0.1$) polymorphic sites in four different gene regions in the wild sample.

Figure 5.4.
Representation of the intergenic space between the different genes located in the sequence surrounding the *Ha* locus. Median LD values across these regions for the cultivated sample are indicated. Coding sequence is represented by rainbow boxes and arrows designate gene orientation. Repetitive sequence is coded similar to the legend in Figure 3.1.

Figure 5.5.
Plots of LD (r^2) as a function of distance (kb) for the cultivated, landrace, and wild samples.

Figure 5.6.
Plots of the median association value for each group of pairwise comparisons against the corresponding median distance. Groups are as follows: within gene comparisons (1-4 kb), comparisons between markers within *hina* and *GSP* (28-32 kb), comparisons between markers within *hinb* and *hina* (77-83 kb), comparisons between markers within *GSP* and *PG2* (98-101 kb), comparisons between markers within *hinb* and *GSP* (107-113 kb), comparisons between markers within *hina* and *PG2* (128-131 kb), and comparisons between markers within *hinb* and *PG2* (207-212 kb).

LIST OF ABBREVIATIONS

ABI	Applied Biosystems Incorporated
AFLP	Amplified Fragment Length Polymorphism
am	Ante Meridian
APS	Ammonium Persulfate
ATP	Adenosine Triphosphate
BAC	Bacterial Artificial Chromosome
bp	base pairs
BSA	Bovine Serum Albumin
cc	cubic centimeter
cDNA	copy Deoxyribonucleic Acid
CHS	Chalcone Synthase
CIAP	calf intestine alkaline phosphatase
CIMMYT	International Maize and Wheat Improvement Center
cm	centimeter
cM	centimorgan
DH	double haploid
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleotide triphosphate
EDTA	Ethylene-Diamine-Tetra-Acetic-Acid
EST	Expressed Sequence Tags
ExoSAP-IT	exonuclease-shrimp alkaline phosphatase
f	frequency
g	gram
g	gravitational force
GlcNAc	N-acetylglucosaminyltransferase
GSP	Grain Softness Protein
Ha	Hardness locus
HEPES	N-2-Hydroxyethylpiperazine-N'-2-ethanesulfonic acid
hina	hordoindoline-a
hinb1	hordoindoline-b1
hinb2	hordoindoline-b2
HKA test	Hudson, Kreitman, and Aguadé test
ICARDA	International Center for Agricultural Research in the Dry Areas
IR	inverted repeat
ITPG	isopropyl-beta-D-thiogalactopyranoside
kb	kilobase
kg	kilogram

KOAc	potassium acetate
kV	kilovolt
L	liter
LB	Luria-Bertani broth
LD	linkage disequilibrium
LINE	long interspersed nuclear element
LOD	logarithmic odds ratio
LP-PCR	long primer polymerase chain reaction
LTR	long terminal repeat
M	molar
mg	milligram
min	minute
MIR	mini-inverted repeat
MITE	Mini-Inverted Repetitive Elements
mL	milliliter
mm	millimeter
mRNA	messenger Ribonucleic Acid
NCBI	National Center of Biotechnology Information
ng	nanogram
°C	degree Celsius
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
PG1	putative gene 1
PG2	putative gene 2
pH	negative logarithm of hydrogen ion concentration
PSI	pounds per square inch
QTL	Quantitative Trait Locus
r^2	measure of association
RFLP	Restriction Fragment Length Polymorphism
RiceGAAS	Rice Genome Automated Annotation System
RNA	Ribonucleic Acid
rpm	revolutions per minute
S	segregating sites
SCRI	Scottish Crop Research Institute
SDW	sterile distilled water
sec	second
SINE	Short Interspersed Nuclear Element
SNP	Single Nucleotide Polymorphism
SOC	Salt Optimized Broth + Carbon
SSC	sodium chloride-sodium citrate

SSPE	sodium chloride-sodium phosphate-EDTA
SSR	Simple Sequence Repeat
TAE	Tris-Acetate-EDTA
TAIR	The Arabidopsis Information Resource
TBE	Tris-Boric acid- EDTA
TDR	terminal direct repeat
TE	Tris-EDTA
TEMED	N,N,N',N'-Tetramethylethylenediamine
TEN	Tris-EDTA-sodium chloride
TIR	terminal inverted repeat
TRIM	Terminal-repeat Retrotransposons in Miniature
tRNA	transfer Ribonucleic Acid
TSD	target site duplication
U	unit
UDP	Uridine Diphosphate
UGT	Uridine Diphosphate Glycosyltransferases
USA	United States of America
UV	ultraviolet
V	volt
VAMP	Vesicle Associated Membrane Protein
v/v	volume per volume
X-Gal	5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside
YAC	Yeast Artificial Chromosome
α^{32}P-dCTP	radioactively labeled Cytidine Triphosphate
θ_w	Watterson's estimator
η	Total number of mutations
µF	microFaraday
µL	microliter
µM	micromolar
µC	microCurie
π	nucleotide diversity
Ω	ohm

ABSTRACT

As more genomes are fully or partially sequenced, attention is being devoted to studying the patterns of intra-specific sequence variation. In human genetics detailed knowledge of linkage disequilibrium (LD) is considered essential for effective population based, high-resolution, gene mapping and cloning. Similar opportunities should exist in plants but analysis is complicated by differences in breeding system and population history which are major contributors to LD. Studies of LD in plants have been largely restricted to maize (outbreeding species; Remington *et al.*, 2001; Tenaillon *et al.*, 2001; Ching *et al.*, 2002; Palaisa *et al.*, 2003) and *Arabidopsis* (inbreeding model organism; Nordborg *et al.*, 2002; Tian *et al.*, 2002; Shepard *et al.*, 2003). This thesis assesses the feasibility of LD mapping for fine-scale association studies in an inbreeding crop species, specifically *Hordeum vulgare*.

A comprehensive comparative analysis of the genome organization of the focal region and the corresponding region in rice, *Oryza sativa*, reveals a complex evolutionary history post-speciation. LD extends across the region to at least 212 kb in barley cultivars but is rapidly eroded in related inbreeding ancestral populations. An undulating pattern of LD was observed with several regions of notable LD increase at intermediate distances. An investigation into the levels of sequence polymorphism and divergence of several genes within the region reveals heterogenous patterns of diversity indicating contrasting gene histories consistent with this observation. These results suggest that haplotype based sequence analysis in multiple populations will provided new opportunities to adjust the resolution of association studies in inbreeding crop species.

DECLARATION

The following thesis is based on the results of investigations conducted by myself, except where otherwise noted in the text, and is of my own composition. The information contained within has not previously been presented in part or in whole for the award of any other degree or qualification. Two manuscripts have been drafted and submitted for publication:

1. Caldwell, K.S., Langridge, P., and Powell, W. Comparative sequence analysis of the region harboring the hardness locus (*Ha*) in barley and its colinear region in rice.
2. Caldwell, K.S., Russell, J.R., Langridge, P., and Powell, W. The extent and magnitude of linkage disequilibrium across a quality trait region in barley.

The barley BAC sequence information (Chapter 3) has been submitted to Genbank under accession number AY643842-AY643844 and the gene haplotype sequences (Chapter 4) under accession numbers AY643845-AY644336.

Upon its acceptance, a copy of this thesis shall be made available for loan and photocopying at the University Library.

Katherine S. Caldwell

March 14, 2005

ACKNOWLEDGEMENTS

It is a surreal and yet amazing feeling to be writing these final words at the end of this incredible journey. I would like to take this moment to thank all the people who have helped me along the way. I am forever grateful to my research supervisor, Wayne Powell, for believing in me, giving me the initiative to reach my goal, and extending to me not only his expertise but also his family, home, and friendship. His enthusiasm and love of science is truly inspiring and his constant encouragement guided me through even the toughest of times. I would like to thank my academic supervisor, Peter Langridge, for his insight and assistance in appreciating the bigger picturing. I am also thankful to Joanne Russell for her aid in navigating the daily trials of laboratory process and scientific frustration.

I am grateful to Andy Flavell, Amar Kumar, Maura Lyons, Craig Simpson, Roger Ellis, Bill Thomas, and Deborah Charlesworth for their intellectual input and guidance on various aspects of my project. I am also thankful for Dave Marshall, Linda Cardel, and Paul Shaw for bioinformatics support. I appreciate the kindness and statistical help of Jim McNichol and the vast sequencing support of Clare McQuade. To the members of Genome Dynamics and the Langridge lab, I am grateful for making the lab more than just a workplace. I am particularly thankful for Mary Woodhead and Ingo Hein for their support, encouragement, laughter, and friendship. To Philip Smith, I am indebted for his grammatical review of the manuscript. If it weren't for him, all the 'closet' homologs would surely have been outed. I would also like to thank Dianne Beharrie for assisting me through three separate Visa applications and a manuscript submission process I wouldn't wish on anyone.

Most importantly, I am forever grateful for the love and support of my husband, Dave, who said goodbye to his family and friends to give me the opportunity to reach my goals.

He has been both my toughest critic and my greatest advocate and not a single step of the journey would have been complete without him.

CHAPTER 1: GENERAL INTRODUCTION

1.1 Introduction

This thesis describes the first study of linkage disequilibrium (LD) in an inbreeding crop species, specifically *Hordeum vulgare* (von Bothmer *et al.*, 1995). In order to establish a scientific context for the importance and relevance of this study, the following literature review will cover the current issues concerning plant breeding and how the application of new approaches to association genetics, specifically LD, could aid in the future progress of the industry. Because this research is founded upon the analysis of local levels of LD in the region surrounding the hardness locus in barley, an overview of the current knowledge concerning the genes within this region and their agronomic importance is provided. In addition, general issues relating to the structure and sequence composition of large genome cereal crops will also be covered.

1.2 Barley: the Crop

Barley is a highly self-fertilizing diploid (2n = 2x = 14) member of the grass family (Poaceae/Gramineae; (Brown *et al.*, 1978). Traditionally wild and cultivated forms of barley have been designated *Hordeum vulgare* ssp. *spontaneum* (or *Hordeum spontaneum*) and *Hordeum vulgare* spp. *vulgare* respectively, although biologically the two are extremely similar. Hybrids occur naturally between the two species yielding fully fertile progeny with normal segregation and chromosomal pairing. One significant difference, however, is that during maturation the brittle rachis of wild barley allows the grain to fall and scatter. In contrast, the cultivated barley spike remains tough throughout maturity, thus, retaining the grain for harvest. This mutational event represents a key domestication step for barley and other cereals (Salamini *et al.*, 2002).

The use of barley as a staple food source can be dated as far back as 17,000 BC based on archaeological remains (Kislev *et al.*, 1992). However, the fermentation of barley grain for alcohol production probably did not arise until 3,000 to 5,000 BC (Edney, 1996).

Today human consumption of barley has declined compared to other cereal crops, namely wheat, maize, and rice. Regardless, its robust adaptive abilities have maintained it as the major cereal crop in regions under high abiotic stress, such as drought (Spain), temperature extremes (Middle East), high salinity (Syria & Iraq), and short vegetation period (Norway, Finland, & Estonia). In 2003/2004 ~60 million hectares will be grown worldwide producing ~135 million tons of grain (http://www.fas.usda.gov/psd/psdselection.asp). Although ~85% of the crop is distributed as animal feed, the highest commercial profits are obtained from malt barley (Fischbeck, 2001). In 2001, the world market of malt barley was almost 5.5 million tons generating 1.5 billion US dollars in revenue (http://apps.fao.org/).

1.3 Barley: the Experimental Organism

In addition to being the fourth largest cereal crop in the world, barley has also proven to be an important experimental organism. One significant on-going experiment, known as the barley Composite Cross II (CCII), involved the generation of 378 F1 hybrid families from all possible pairwise crosses involving 28 global elite barley cultivars (reviewed in Allard, 1999). Each family has been maintained since 1928 without conscious selection and grown annually in different environmental habitats and under different agricultural regimes. Phenotypic information for both Mendelian traits, such as two-row versus six-row and smooth versus rough awn, and quantitative traits, including flowering time and plant height, were carefully recorded in early generations. This approach was complemented by isozyme and genetic marker diversity data in later generations as new technologies emerged. The culminated data has offered valuable information on the interaction of existing diversity, environmental constraints, and agricultural practices and the individual effects they have on the development of multilocus systems.

Population genetic assays in barley have provided new insights into the effects evolutionary forces have on the patterns of nucleotide diversity in inbreeding species.

The direct sequence analysis of nine different genetic loci in *Hordeum spontaneum* has identified a heterogeneous pattern of diversity indicating contrasting gene histories despite the low levels of effective recombination associated with high levels of self-fertilization (Cummings *et al.*, 1998; Lin *et al.*, 2001; Lin *et al.*, 2002; Morrell *et al.*, 2003). Furthermore, estimates of low levels of heterozygosity within wild Israeli populations suggest that frequent introgression from cultivated barley cannot account for the level of diversity seen between and within wild barley populations (Brown *et al.*, 1978). This finding is of particular importance because it indicates that despite close ancestry, the majority of the variation found within wild populations is distinct from that within cultivated germplasm and could be a potential resource for novel variation in breeding programs.

1.4 Technological Advances in Genomic Resources

Over the past decade, substantial effort has gone into the development of different technologies for the generation of genomic resources (Table 1.1). These tools have proven to be valuable in influencing the direction of barley breeding and research. The establishment of expressed sequence tag (EST) databases has paved the way for gene discovery and differential expression analysis across various tissues and growth stages (http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=4513&lvl=3&nucl=1&keep=1&srchmode=1&unlock). In addition, the correlation of phenotypic data with genetic linkage analysis has advanced the understanding of the number, organization, location, and contribution of individual loci controlling complex phenotypes including agronomic, quality, and disease resistance traits (Hayes *et al.*, 1993; Han *et al.*, 1995; Graner, 1996; Ullrich *et al.*, 1997; Mather *et al.*, 1997; Han *et al.*, 1997a; Han *et al.*, 1997b; Hackett *et al.*, 2001; Kleinhofs *et al.*, 2002). The generation of yeast artificial chromosome (YAC) and bacterial artificial chromosome (BAC) libraries have allowed the sequencing of large contiguous genomic fragments providing insight into genome organization and composition (Kleine *et al.*, 1993a; Panstruga *et al.*, 1998; Shirasu *et al.*,

Table 1.1. List of recent genomic resources, their methodology, and application.

Genomic Resources	Methodology	Application	References
EST Libraries			
Unnormalized libraries Normalized libraries Subtraction libraries	Isolation of cDNA from specific tissues for sequencing	Gene discovery by direct sequencing Estimation of expression levels Identification of different alleles	(Patanjali *et al.*, 1991) (Hedrick *et al.*, 1984)
Large Insert Libraries			
Yeast Artificial Chromosomes (YACs) Bacterial Artificial Chromosomes (BACs)	Use of vectors capable of retaining large inserts of foreign DNA >250 kb in length	Full genome sequencing Genome structure and organization Positional Cloning	(Kleine *et al.*, 1993b) (Shizuya *et al.*, 1992b) (Woo *et al.*, 1994b) (Zhang *et al.*, 1996)
Full Genome Sequencing			
Shotgun approach Large insert library sub-cloning	Methods for sequence and assembly of complete genomes	Gene discovery by sequence similarity Genome structure and organization Genome composition	(Venter *et al.*, 2001) (Lin *et al.*, 1999) (Goff *et al.*, 2002)
Framework Genetic Linkage Maps			
based on SSRs, AFLPs, RFLPs, etc.	Determination of the genetic distance (cM) between markers/genes.	Marker/Gene order Gene discovery by microcolinearity Quantitative Trait Loci (QTL) mapping	(Botstein *et al.*, 1980) (Qi *et al.*, 1998) (Melis *et al.*, 1993)
Physical Maps			
Large insert library fingerprinting Gametocidal substitution lines Deletion lines Radiation Hybrid HAPPY mapping	Determination of the physical distance (nucleotides) between markers/genes.	Genome structure and organization High-resolution mapping Positional cloning	(Marra *et al.*, 1997) (Werner *et al.*, 1992) (Goss *et al.*, 1975) (Barrett, 1992) (Dear *et al.*, 1993) (Waugh *et al.*, 2002) (Kunzel et al. 2000)
Gene Targeted Reverse Genetics			
RNAi Viral Induced Gene Silencing (VIGS)	Homology dependent post-transcriptional gene silencing	Targeted functional validation	(Depicker *et al.*, 1997) (Anandalakshmi *et al.*, 1998) (Ruiz *et al.*, 1998)
Structured Mutant Populations			
Over expression constructs T-DNA knockouts T-DNA activation tagging Transposition tagging Chemical Mutagenesis	Forward & reverse genetics	Functional validation Gene discovery Discovery of novel phenotypes	(May *et al.*, 2003) (Hirochika, 2001) (Weigel *et al.*, 2000) (Parinov *et al.*, 2000) (Enoki *et al.*, 1999) (McCallum *et al.*, 2000)

2000; Yu *et al.*, 2000; Dubcovsky *et al.*, 2001; Rostoks *et al.*, 2002; Wei *et al.*, 2002; Yan *et al.*, 2002; Gu *et al.*, 2003). Furthermore, the generation of genetic and physical maps provided a platform for comparative and integrative mapping, thus, establishing a link between genetic and physical distance for unraveling the nature of regional and global recombination (Pedersen *et al.*, 1995; DeScenzo *et al.*, 1996; Lahaye *et al.*, 1998; Druka *et al.*, 2000). The above technologies also enable interspecies comparative approaches for extrapolating information on gene structure, function, location, and history (Cummings *et*

al., 1998; Han *et al.*, 1998; Dubcovsky *et al.*, 2001; Brunner *et al.*, 2003). Currently several structured mutant populations based on the Targeting Induced Local Lesions in Genomes (TILLING) methodology and transposition tagging based on Ac/Ds are under construction for forward and reverse genetics approaches (http://www.intl-pag.org/11/abstracts/P5c_P397_XI.html; (Koprek *et al.*, 2000). The specific utilization of certain technologies will be discussed in greater detail in later chapters.

1.5 Plant Breeding Approaches for Self-Pollinated Plants and the Significance for Barley Improvement

Ancient farmers relied upon the utilization of wild barley variants with specific qualities more desirable for agriculture than those of other natural variants. These landraces were planted mainly for subsistence farming and although some degree of selection probably occurred it is likely that minimal effort was put into the development of superior varieties. In 1865, the publication of Mendel's experiments provided a first glance at the basic principles of genetic transmission (Mendel, 1865). A better understanding of basic genetic principles and phenotypic selection established the necessary foundation for the implementation of crop improvement strategies. Likewise, plant breeding programs provided a platform for further research into understanding the underlying genetic principles.

Despite the fact that ~85% of barley worldwide crops is sold to the animal feed industry (Fischbeck, 2001), until the 1960s little effort was put into the development of variants with superior nutrient and digestibility qualities. Feed barley was limited to varieties that did not meet malting standards; therefore, targeted breeding programs could provide substantial improvement in these areas. Furthermore, although human consumption of barley has reduced in favor of other cereal crops, such as maize, rice, and wheat, barley's superior adaptive nature has maintained its status as a staple food in regions of high abiotic stress, including the Middle East and lower Nordic Regions. Further advances in

abiotic tolerance as well as nutrient utilization properties would be highly beneficial for farmers and end users throughout these regions. Commercially, malting barley for distilling and brewing has the greatest economic potential. Economic interest in this area drives the major portion of barley advancement programs to develop grain with superior malting quality, for example protein content, diastatic power, and malt extract. Barley breeding programs are, therefore, integral for the development of new varieties that contribute to the ultimate agronomic and end product success of the crop while remaining complementary to the specific environmental conditions, agricultural practices, and production requirements. Successful breeding is ultimately dependent upon the generation of novel, superior genotypes, the accurate association of genetic diversity and phenotypic variation for reliable selection regimes, and an understanding of the environmental components contributing to phenotype.

Several advances have been made in recent years to accelerate improvement of agronomic and quality traits. From 1946 to 1980, an approximate 2-fold increase in average barley yields was obtained in the UK (Silvey, 1986). Although a certain percentage of this gain is attributable to advances in agricultural technology and xenobiotics, approximately 50% can be accredited to the development and introduction of novel genotypes (Riggs *et al.*, 1981). In addition, both low-phytic acid varieties with increased levels of inorganic phosphorus and high-lysine lines have been developed for animal feed improvement (Seko *et al.*, 1981; Raboy *et al.*, 2001). High phosphorus levels are desirable for animal nutrients; however, a considerable portion of barley phosphorus content is found in phytic acid which is unavailable to non-ruminant animals. Likewise, crops with balanced amino acid composition and high protein content are desirable for animal nutrition. A considerable amount of research has also been conducted for the improvement of malting quality varieties spanning a wide range of traits involved in hormone control, enzyme production, starch and protein accumulation, and the transport of soluble material (reviewed in Swanston *et al.*, 1999).

1.5.1 Pedigree or Progeny Breeding

The "classical pedigree" breeding method entails an initial cross between two elite pure-line parents to establish a hybrid F_1 population. The progeny of the selfed F_1 hybrids (F_2 generation) are planted with adequate spacing to minimize the possibility of rare outcrossing events. Those plants demonstrating desirable/superior phenotypes are selected for selfing to generate F_3 families. Each family is subjected to several subsequent rounds of selection and selfing of individual superior plants. By maintaining complete inbreeding, heterozygosity levels are expected to decrease by ½ with each successive generation. Therefore, by the F_7 generation, the depletion of within family variation through selfing coupled with increased between family variation through selection provide a more reliable platform for the initiation of single plant selection based on rigorous yield and quality testing. Although the "classical pedigree" method is still widely employed by many commercial breeders, it has several distinct disadvantages. Evaluation of prospective worth is predominantly focused on the early stages when heterozygosity levels are high. Consequently, the inheritance of novel allelic combinations forming favorable multi-locus genotypes is low. In order to ensure that a sufficient number of "superior" families survive to the more rigorous testing stages, several thousand F_2 plants may be required for a single cross or else smaller numbers of multiple crosses are maintained. This culminates in an extremely labor intensive program that requires meticulous record keeping and ample growing space.

In an attempt to alleviate some of the drawbacks of the "classical pedigree" method, the "bulk population" method postpones family evaluation until the F_5 or F_6 generation where a higher level of homozygosity can be achieved. Early generations are subjected to negative selection where the elimination of inferior plants allows a greater proportion of potentially superior plants to proceed to future generations. Without the need to distinguish between different families, these generations can be handled in bulk reducing

labor and quantity of record keeping. The "single-seed descent" method eliminates all selection from early generations of the breeding program relying on the concept that accurate quality testing cannot be reliably achieved until plants are allowed to reach near-homozygosity and all multi-locus systems are completely established. Instead, one seed from each plant of the previous generation is automatically carried forward to the next generation (until F_5 or F_6) significantly increasing the number of overall novel homozygous allelic combinations proceeding to the quality and yield testing stages. Furthermore, with the elimination of selection and the need for only a single seed from each individual, optimal growing conditions are no longer necessary. Through the early stages, plants can be grown under glasshouse conditions enabling production of 3 to 4 generations per year, thus reducing time constraints. Although the single-seed method allows a broader and more random selection of novel homozygous lines to proceed to future generations, it can be argued that the selection employed at the earlier stages of the other two methods could still provide a superior overall sample, provided that selection was effective.

1.5.2 Backcrossing

Hybridization between two genetically diverse parents is capable of producing enormous genetic complexity. In barley, a single allelic difference per chromosome between parental lines can result in 2^7 or 128 different homozygous genotypes. Therefore, crosses between genetically similar parents are preferred to reduce the number of progeny required in subsequent generations and maximize the probability of identifying superior hybrids. Furthermore, the continued use of proven elite breeding parents reduces the risk of destabilizing well established favorable linkage blocks. However, many desirable phenotypes, such as disease resistance and abiotic stress tolerance, are not commonly found within the elite parental lines of modern breeding programs. Introgression of these traits is often accomplished by the incorporation of wild germplasm that, through natural selection, has adapted to more extreme and diverse environments. Unfortunately, with the

exception of those regions harboring the genes controlling the desired phenotype, the majority of the genetic background of the introgressed wild line is typically substandard compared to that of its parental pair. The ultimate goal is to incorporate only the integral loci of the donor parent into the elite genetic background of the recipient parent. To accomplish this end, the progeny of successive generations are subjected to a series of backcrossing to the recurrent parent followed by selection for the presence of the desired donor phenotype.

Backcrossing schemes have several advantages over traditional hybridization methods. The incorporation of multiple desirable phenotypes into one elite genetic background can be done in stages by utilizing the improved backcross-derived progeny from the first trait as the recurrent parental line in crosses for additional traits. This is exemplified by the ICARDA/CIMMYT initiative that successfully incorporated several disease resistance genes into high yielding varieties (Vivar, 2000). Furthermore, optimal growing conditions are not required providing the environment allows sufficient expression of the phenotype under selection. Most importantly, all unfavorable donor alleles, regardless of heritability, will automatically be resolved with little attention from the breeder. Although backcrossing does limit progress to predictable change, the maintenance of a stable genetic background is an invaluable tool in combating the reluctance of farmers to switch to a new variety by providing assurance that minimal re-optimization is necessary to incorporate the new line into current farming practice.

1.5.3 Double Haploid Lines

Double haploid technology enables the duplication of the genetic components within a haploid cell to create a completely homozygous diploid individual (Pickering *et al.*, 1992). The production of double haploids can be incorporated into any stage of a breeding program ensuring homozygosity at all loci. This results in increased confidence in selection efficiency as dominant effects caused by heterozygosity are eliminated. This

also allows added control over the number of recombination events in order to conserve previously existing desirable linkage groups. Furthermore, double haploidy ensures pure-line varieties for marketing. Double haploid lines have been particularly successful for genetic mapping because they guarantee uniformity allowing inbreds to be evaluated in different seasons and environments and providing a resource that can be shared throughout the research community (Langridge *et al.*, 2003).

1.5.4 Marker Assisted Selection

Many agronomic and quality traits, such as yield, starch composition, and various aspects of malting quality, require substantial quantities of grain for accurate phenotypic scoring. Often there is not enough grain available for testing at the early stages of breeding programs limiting selection strategies to obvious visible phenotypes. In addition, several testing assays are destructive to the grain and consequently prevent further evaluation and require additional seed for propagation of subsequent generations. Furthermore, only phenotypic data can be accurately assessed using traditional evaluation procedures. Replication plots from several diverse environments are often necessary to assure detection of significant environmental interactions. This process multiplies workload and acquisition cost by a set factor (number replicates) ultimately reducing the number of plants that can realistically be brought forward to the testing stages within manageable limits.

Occasionally a visible phenotype for one gene can act as a diagnostic marker for a less easily scored phenotype of another closely linked gene. Several successful morphological markers have been identified in barley utilizing these tight correlations (Davis *et al.*, 1997), for example awn texture and yield (Everson *et al.*, 1955), hairy leaf sheath (*Hs*) and vernalization (*Sh1*; Sogaard *et al.*, 1987), and glossy-sheath (*sg4*), orange lemma (*o*) and root rot (Kutcher *et al.*, 1996; Davis *et al.*, 1997). Unfortunately, the use of morphological markers is limited to genes of simple inheritance as control by complex

multi-locus systems could confuse results leading to inaccurate selection. When possible, breeders employ the use of molecular markers to aid in the selection of superior lines. A summary of the most commonly used marker systems can be seen in Table 1.2. Appropriate markers for trait selection are determined through the correlation of phenotypic information with the segregation pattern of molecular markers in diverse mapping populations. Regions demonstrating significant marker/trait associations for complex traits are designated quantitative trait loci (QTL) and are believed to harbor the necessary components for phenotypic expression. By determining the percentage of phenotypic variation attributable to a given QTL, a statistical degree of confidence can be placed upon probability that selection of a marker defining the QTL will also result in the selection of the trait of interest.

Table 1.2. List of the most prominent molecular marker systems.

Marker System	Description of Method	Type	References
Restriction Fragment Length Polymorphisms (RFLPs)	Differentiation in the presence or absence of restriction endonuclease cleavage sites or large insertion/deletions by hybridization with cDNA probes	codominant	(Botstein *et al.*, 1980)
Random Amplified Polymorphic DNA (RAPDs)	Differentiation in the arbitrary amplification fingerprint of genomic DNA using randomly generated primers	dominant	(Welsh *et al.*, 1990) (Williams *et al.*, 1990)
Amplified Fragment Length Polymorphisms (AFLPs)	Differentiation in the amplification fingerprint of genomic DNA based on specific restriction digests	dominant	(Vos *et al.*, 1995b)
Simple Sequence Repeats (SSRs)	Detection of length polymorphism caused by differentiation in the number of 1-5 bp repeated nucleotides	codominant	(Tautz *et al.*, 1984)
Single Nucleotide Polymorphisms (SNPs)	Detection of single nucleotide polymorphisms	codominant	Reviewed in: (Rafalski, 2002a) (Rafalski, 2002b)

Throughout the 1990s, the emergence and availability of several genome wide meiotic maps in different barley crosses created the platform for the first robust association

studies in barley (Graner *et al.*, 1991; Heun *et al.*, 1991; Kleinhofs *et al.*, 1993; Langridge *et al.*, 1995; Sherman *et al.*, 1995; Becker *et al.*, 1995; Waugh *et al.*, 1997; Qi *et al.*, 1998). Furthermore, the establishment of fully automated, high-throughput genotyping platforms allowed new opportunities for reliable identification of genetic diversity within plant breeding material. Genetic markers, such as Restriction Fragment Length Polymorphisms (RFLPs; Botstein *et al.*, 1980) and Simple Sequence Repeats (SSRs; Tautz *et al.*, 1984) enabled scoring of both dominant and recessive alleles. These marker systems as well as Amplified Fragment Length Polymorphisms (AFLPs; Vos *et al.*, 1995a) further increased confidence in regions harboring closely linked desirable and adverse traits and expanded evaluation programs to include genes with otherwise difficult to discern phenotypes. These advantages of marker assisted selection have the additive effect of greatly reducing time constraints allowing breeders to get new varieties to market up to 2 and 7 years faster for dominant and recessive traits respectively.

To date, many published marker/trait associations in barley have focused on disease resistance and other phenotypes for which selection is already reasonably efficient and cost-effective (reviewed in Thomas, 2002). Consequently, commercial breeders still use conventional methods as the high cost of establishing high-throughput genotyping platforms outweighs the small increase in efficiency for such traits. More encouraging examples in the literature are those for traits, such as beta-amylase activity (*Bmyl*; Erkkila, 1999), (1-3)-beta-glucanase (*Glb*; Li *et al.*, 1996), limit dextrinase (Li *et al.*, 1999), high-amylose starch (*Amo1*; Schondelmaier *et al.*, 1992), and resistance to yellow mosaic virus complex (Graner *et al.*, 1993; Weyen *et al.*, 1996; Bauer *et al.*, 1997; Saeki *et al.*, 1999; Graner *et al.*, 1999), whose phenotypes are otherwise difficult to assess. As more sequence data and higher resolution association techniques become available, marker-assisted selection will inevitably become a more compelling option.

1.6 Association Genetics

The study of traits controlled by simple Mendelian inheritance aided scientists in acquiring a strong foundation for exploiting the basic genetic principles for the advancement of scientific research. However, Mendelian genetics does not encompass the true extent of genetic complexity as many phenotypes are controlled by intricate multi-locus systems. Traditional approaches for identifying the number and location of the individual components contributing to complex traits involve the integration of phenotypic data and genetic linkage analysis using meiotic maps as the structure for uncovering QTLs. Linkage analysis involves the correlation of two loci or markers as a consequence of close proximity along a chromosome. These studies are, therefore, limited to the genetic diversity segregating within families and often do not provide resolution below the megabase level.

More recently, scientists have started to investigate alternate approaches in an attempt to improve the resolution of these studies. In contrast to linkage analysis, association genetic approaches evaluate the correlation between two loci or markers as a result of shared population history. These approaches include the genetic diversity across families from seemingly unrelated individuals within the natural population of a species. Therefore, association mapping effectively increases the number of recombination events to include all occurrences within the history of the sample. This presents a distinct advantage over traditional methods by potentially improving resolution from the megabase to the kilobase scale.

1.6.1 Quantitative Trait Loci (QTL) Mapping

The resolution of QTLs through linkage analysis is limited by the number and distribution of mapped genetic markers across regions controlling complex traits. For this reason, significant effort has been devoted to developing techniques to increase the resolution of previously identified QTLs. Substitution mapping utilizes a segregating population of

backcross-derived isogenic lines (recombinant chromosome substitution lines) differing only by short chromosomal segments within the QTLs of interest (Paterson *et al.*, 1990). Phenotypic effects shared by overlapping segments are attributed to the related chromosomal fragment while effects unique to a given segment are attributed to the unique region. This technique can be used to either individually investigate one QTL at a time (Han *et al.*, 1997a) or simultaneously examine all QTL for a given trait (Paterson *et al.*, 1990; Ramsay *et al.*, 1996). Treating QTL on an individual basis eliminates possible interference from the other segregating QTL. This is particularly useful in the dissection of regions harboring several different overlapping QTLs and could aid in identifying pleiotropic effects (Han *et al.*, 1997a). However, it also negates detection of possible epistatic interactions otherwise revealed by simultaneous approaches.

QTL analysis has aided in locating several genomic regions harboring important agronomic traits including yield, plant height, and time of heading (Hayes *et al.*, 1993; Hackett *et al.*, 2001). In addition, malting quality traits, such as malt extract and diastatic power, along with the enzymatic activities of β-glucanase and α-amylase have also been the focus of QTL analysis (Han *et al.*, 1995; Ullrich *et al.*, 1997; Mather *et al.*, 1997; Han *et al.*, 1997a; Han *et al.*, 1997b). Furthermore, QTL approaches have aided in the mapping and mapped-based cloning of several major genes involved in disease and viral resistance (reviewed in Graner, 1996; Kleinhofs *et al.*, 2002). A summary of mapped barley QTL across several different populations can be found at the following web addresses: http://barleyworld.org/NABGMP/qtlsum.htm and http://greengenes.cit.cornell.edu/WaiteQTL/CxH.html.

Although the development of QTL mapping has allowed considerable advances in the understanding of the number, organization, location, and contribution of individual loci controlling complex phenotypes, several limitations still exist (reviewed in Kearsey *et al.*, 1998). For example, manageable QTL populations are usually not large enough to

account for all the variability observed for a given trait causing small but important components to go undetected. In addition, the complexities of multi-locus systems coupled with unpredictable environmental interactions often generate inconsistencies in QTL location and percent contribution across populations and environments. Furthermore, for crops with relatively large genome size and limited growing seasons per year, appropriate populations for fine-scale mapping can take several years to generate and still only allow resolution at the megabase level. With such a large window for error, the probability of confidently choosing appropriate and reliable markers to aid in selection is extremely poor. Moreover, there is a much higher likelihood that undetected adverse alleles could reside within the same interval. This limited resolution also has negative implications on the feasibility of positional cloning through chromosome walking. In species whose genomes are primarily composed of repetitive sequence, identification of low-copy DNA regions applicable for accurate contig extension poses a considerable challenge. Despite recent advances enabling efficient cloning of large (>150 kb) genomic fragments (http://hbz.tamu.edu/bacindex.html), even under optimal conditions assuming maximal clone length and minimal overlap, a physical map composed of several thousand bacterial artificial chromosomes (BACs) would be necessary to span a single centimorgan of the barley genome (~4 Mb/cM; (Kleinhofs *et al.*, 1993). Clearly, the development of new association methods allowing fine-scale, high resolution mapping are essential to realize the full benefits of marker-assisted selection and positional cloning technology.

1.6.2 Linkage Disequilibrium

Linkage Disequilibrium (LD) is the non-independence of alleles at different loci. At its most basic level, LD is maintained through a delicate balance between mutation and recombination. At the moment of spontaneous generation, all new mutations are in perfect association with their genetic background. Through random genetic drift, these mutations will start to fluctuate in frequency along with their associated haplotypes until

ultimately becoming eliminated or fixed in the population. Over time, as recombination gradually acts to decay older associations, new associations are established with the emergence of additional mutations. This simultaneous interplay between mutation and recombination is one of the factors preventing the complete realization of natural LD decay with distance. However, LD is ultimately a product of several evolutionary and biological factors contributing to the overall population and allelic histories of a species (Table 1.3).

Table 1.3. Factors that influence Linkage Disequilibrium.

Force		Effect on LD	References
Mutation rate	Generates polymorphism within a population and therefore the rate at which new associations are formed	Generates LD	Reviewed in : (Gaut *et al.*, 2003) (Flint-Garcia *et al.*, 2003)
Recombination rate	Determines the rate of association decay	Decrease in LD	(Jeffreys *et al.*, 2001) (Daly *et al.*, 2001) (Reich *et al.*, 2001)
Population substructure & admixture	Increases the level of "spurious" associations through the inter-mixing of individuals from populations of contrasting evolutionary histories	Genome-wide increase in LD	(Pritchard *et al.*, 1999) (Pritchard *et al.*, 2001) (Thornsberry *et al.*, 2001)
Population bottleneck	Causes a sharp decrease in genetic diversity through the rapid reduction in population size	Short term global increase in LD	(Wall *et al.*, 2002) (Dunning *et al.*, 2000) (Hastbacka *et al.*, 1992)
Directional selection	Causes a local reduction in genetic diversity as a result of the increased frequency of advantageous alleles	Short term regional increase in LD	(Przeworski, 2002) (Wall *et al.*, 2002)
Mating system (Inbreeding)	Decreases the effective recombination rate through non-random mating	Genome-wide increase in LD	(Nordborg *et al.*, 1997) (Nordborg, 2000)

The most prominent concern with association studies is the effects of non-random mating, namely population substructure and admixture (Pritchard *et al.*, 1999). As a result of random genetic drift and founder effects, genetically distinct populations will naturally vary in allele-frequencies. Consequently, a predominant genetic trait of one population

will inevitably show strong associations with all high frequency alleles within that population. This results in a high number of "spurious" associations that rapidly complicate the detection of causative alleles. In one documented study, up to 80% of the significant associations detected between the polymorphisms in the maize *dwarf8* (*d8*) gene and flowering time could be eliminated as an effect of population substructure (Thornsberry *et al.*, 2001).

Population bottlenecks and directional selection also perform an important role in the formation of global and local levels of LD respectively. Population bottlenecks are caused by a sharp decrease in population size eliminating a large portion of genetic variability. The recovery phase is marked by an excess of rare alleles as new mutations arise on the backgrounds of surviving ancestral haplotypes. This pattern of genetic diversity is very similar to chromosomal regions subjected to recent directional selection. Although the surplus of rare alleles serves to increase LD, in the absence of continued selection or population size suppression, these effects are usually short-lived (Hastbacka *et al.*, 1992; Dunning *et al.*, 2000; Wall *et al.*, 2002; Przeworski, 2002).

Biological factors, such as mating systems, can also have a profound impact on LD. Simulation studies have demonstrated that in the absence of other mitigating factors, high levels of LD persist to a greater extent in highly selfing species (Nordborg, 2000). This is predominantly a factor of the effective recombination rate. As a result of high levels of homozygosity, a significant proportion of all recombination events occurring within inbreeding systems do not bring about an exchange of genetic variation. In the most extreme comparison, between complete inbreeding and outbreeding systems, a 50-fold reduction of effective recombination can exist (Flint-Garcia *et al.*, 2003). However the presence of other factors, such as inconsistent inbreeding levels and different mutation rates, can have a significant impact. Despite varying levels of inbreeding, the malaria parasite, *Plasmodium falciparum*, demonstrates a significantly higher effective

recombination rate than in some predominantly outbreeding species, such as humans and *Drosophila* (Conway *et al.*, 1999).

As shown above, each evolutionary component contributes to the patterns of diversity and LD seen within and among the different populations of species. As a result, LD analysis can be exploited to aid in discovering the pathways of evolutionary history and reveal the idiosyncrasies of regional recombination. Although the nature of these historical events can sometimes obscure precise identification of causative associations, the development of advanced statistical tools to correct for "spurious" associations greatly improves the employment of LD for gene-mapping purposes (Pritchard *et al.*, 1999; Pritchard *et al.*, 2000). LD mapping has already become a powerful tool in human association studies for pinpointing causative alleles controlling both simple-inherited and complex diseases (Kerem *et al.*, 1989; Corder *et al.*, 1994; Hugot *et al.*, 2001; Ogura *et al.*, 2001b). Furthermore, both local and global human LD studies have provided invaluable insight into LD structure and maintenance in human populations allowing the development of different association strategies for global and high-resolution mapping (reviewed in Goldstein, 2001; Gabriel *et al.*, 2002). These observations could prove to be invaluable in directing parallel investigations into plant systems – a research area currently in its infancy. The current status of both human and plant association studies is discussed in greater detail in Chapter 5.

1.7 Repetitive Sequence

The integration of foreign DNA, namely transposable elements, into a host genome can also have significant evolutionary effects (reviewed in Bennetzen, 2000b). Greater than 70% of some plant genomes are known to be composed of repetitive sequence (SanMiguel *et al.*, 1996; Bennetzen *et al.*, 1997; Wicker *et al.*, 2001). Together with polyploidation, these invasions are the predominant cause of the overwhelming variation in genome size among grass species (Bennetzen *et al.*, 1997; SanMiguel *et al.*, 1998).

The haploid barley genome (5000 Mb) is approximately 2 and 11-fold greater in size than that of maize (2500 Mb) and rice (440Mb) and 1/3 the size of allohexaploid wheat (16000 Mb; Arumuganathan *et al.*, 1991; Shields, 1993). Consequently, in order to fully unravel the complexities of inter- and intra-species variation, detailed knowledge about the nature, location, and orientation of these elements with respect to the regulatory and functional regions of coding sequence is imperative.

1.7.1 Class I Transposable Elements

Class I elements (reviewed in Grandbastien, 1992; Kumar *et al.*, 1999; Bennetzen, 2000b), including both long terminal repeat (LTR) and non-LTR retrotransposons, self-replicate through an RNA intermediate (Figure 1.1). LTR retrotransposons are further classified into Ty1-*copia* and Ty3-*gypsy* elements based on internal sequence homology and gene order. LTRs range from several 100 bp to 5 kb and are flanked by short 2-6 bp inverted repeats (IRs) usually beginning with 5'-TG-3'. LTRs also contain both the primer binding site and termination sequences required for the singular transcription of all gene products necessary for reverse transcription and integration into the host genome. Newly synthesized copies are incorporated through staggered double-stranded cleavage of the target site generating a short 3-6 bp flanking duplication (target site duplication, TSD). With the ability to continually generate and incorporate new copies into the host genome, the solitary action of a single element invasion could potentially have a significant impact on genome size.

Non-LTR retrotransposons consist of three main classes of elements: long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and terminal-repeat retrotransposons in miniature (TRIM; Figure 1.1). Sequence diversity studies suggest that LINEs are the oldest retrotransposons in eukaryotic genomes. Unlike their LTR bearing counterparts, these elements do not contain the integrase gene but instead

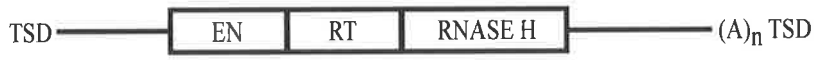
LTR Retrotransposons

Ty1-*copia* elements

Ty3-*gypsy* elements

non LTR Retrotransposons

LINE



SINE

TRIM

Class II Transposons

MITE

Figure 1.1. Structure of the different classes of repetitive elements. Encoded proteins i.e. protease (PR), integrase (INT), reverse transcriptase (RT), endonuclease (EN), and transposase (TRANS) and other key features i.e. target site duplication (TSD), inverted repeat (IR), long terminal repeats (LTR), primer binding site (PBS), polypurine tract (PPT), terminal direct repeat (TDR), terminal inverted repeat (TIR), and mini-inverted repeat (MIR) are labelled.

encode a protein with endonucleolytic activity (EN) for incorporation into the host genome. Although LINEs are bordered by TSDs, the lack of LTRs makes them significantly harder to recognize. The high conservation of reverse transcriptase and presence of a polyA tract just upstream of the 3' TSD help to define the 3' terminal end. However, low sequence homology of the upstream GAG (group-specific antigen) genes and frequent truncation of the 5' portion of these elements make full length copies difficult to identify.

All SINEs are believed to have been derived from RNA polymerase III products (tRNA) and maintain the A- and B-box promoter recognition motifs for reverse transcription. These elements are less than 500 bp in length, terminate in polyA sequence, and are flanked by 7-24 bp TSDs. In contrast to other retrotransposons, SINEs do not encode for any of the proteins needed for self-replication and integration and are, therefore, believed to rely upon the translation products of other retrotransposons to perform these functions. Although, the exact mechanism of replication and transposition are still unclear, the presence of several 5' truncated copies flanked by perfect TSDs suggest that the integral components are located within the terminal 3' region (Myouga *et al.*, 2001).

TRIM elements (Witte *et al.*, 2001) are similar to LTR-retrotransposons in that they are flanked by short ~5 bp TSDs and contain both a primer binding site and polypurine tract immediately internal to respective 5' and 3' terminal direct repeats (TDRs). The size of these elements (<540 bp) as well as their TDRs (100-200 bp), however, are considerably smaller than those of their predicted predecessors. In addition, similar to SINEs, these elements lack the capacity for self-replication and, therefore, are predicted to utilize the mobility-related proteins of other retrotransposons for reverse transcription and integration.

1.7.2 *Class II transposable elements*

Contrasting retrotransposons, class II elements do not have the capability of replicating through an RNA intermediate. Instead, they contain a single ORF encoding a transposase which is responsible for the excision and reinsertion of the complete element to new locations of the genome (Figure 1.1). This "cut and paste" mechanism hinders element copy number from escalating above the number of independent insertion events. Class II elements are usually several kilobases in length and are flanked by 10 to >200 bp terminal inverted repeats (TIR). Like all other known elements, these transposons are bordered by short TSDs generated during the reintegration into the host genome.

1.7.3 *Mini-Inverted Repetitive Elements (MITEs)*

Frequently <500 bp in length, Mini-Inverted Repetitive Elements (MITEs) are the smallest known transposable elements. MITEs were originally characterized by the presence of flanking mini-inverted repeats (MIRs) allowing them to form stable DNA secondary structures (Figure 1.1). The two largest reported families, Stowaway and Tourist, demonstrate preferential insertion into 2-3 bp TA target sites; however, TSDs up to 9 bps have been reported (Bureau *et al.*, 1992; Bureau *et al.*, 1994a; Charrier *et al.*, 1999).

1.7.4 *Role of Repetitive Sequence in Molecular Evolution and Gene Function*

Initial reports of transposable elements referred to these invasive sequences as parasitic or junk DNA; however, deeper investigation has revealed them to be an invaluable mechanism for molecular evolution (reviewed in Kumar *et al.*, 1999; Kidwell *et al.*, 2000). For example, non-reciprocal recombination between two proximal copies of a full-length element in like orientation can result in either duplication or deletion of the inter-element space (Williamson, 1983; Lim *et al.*, 1994). In a similar fashion, unequal crossing-over between LTRs of the same element can act as a mechanism of reverse genome expansion (genome contraction) resulting in a solo LTR flanked by the original

TSD (Shepherd *et al.*, 1984; SanMiguel *et al.*, 1996; Chen *et al.*, 1998; Vicient *et al.*, 1999). Furthermore, regions of DNA bordered by elements in opposite orientation could be subjected to segmental inversions and elements located on entirely separate chromosomes could lead to translocation events (Williamson, 1983; Lister *et al.*, 1993; Lim *et al.*, 1994).

Some transposable elements have shown preferential insertion into heterochromatic regions of the genome. The high conservation of these positional insertions across related species has led authors to suggest that they may play an important role in centromeric and telomeric structure and function (Pimpinelli *et al.*, 1995; Presting *et al.*, 1998; Miller *et al.*, 1998). In *Drosophila* both the *HeT-A* and *TART* elements have proven to be critical components of telomeric composition and maintenance (Pardue *et al.*, 1997). Likewise, extensive sequence homology of *Tigger2* elements to the highly conserved centromeric-associated protein CENP-B suggests that these elements may perform an evolutionary role in the structural preservation of centromeres in mammalian systems (Kipling *et al.*, 1997).

The acquisition of cellular genes into transposable elements can occur through "inefficient transcription termination" as a result of weak polyadenylation signals (Jin *et al.*, 1994; Bureau *et al.*, 1994b; Takahashi *et al.*, 1999; Elrouby *et al.*, 2001). In addition, the reverse transcriptase and integration activities of retrotransposons can provide a mechanism for the incorporation of mRNA and cDNA copies of host nuclear genes into the genome (Vanin, 1985; Weiner *et al.*, 1986; Drouin *et al.*, 1987). Although rare, these occurrences could provide templates for gene conversion and the evolution of alternate gene function.

Repetitive sequences have been shown to interact directly with functional copies of host genes. Because LTR retrotransposons use a unique method of replication resulting in

identical LTRs, the 3'LTR also contains a functional promoter which can potentially contribute to alternate expression patterns of genes downstream of the insertion site. Even in the absence of promoter activity, the presence of elements within coding or regulatory regions has resulted in changes in tissue specificity and gene regulation (Alleman *et al.*, 1993; White *et al.*, 1994; McDonald *et al.*, 1997; Marillonnet *et al.*, 1997), creation of alternative splice sites (Varagona *et al.*, 1992), and generation of new introns (Giroux *et al.*, 1994). Moreover, the excision of these elements can also have varying effects on gene regulation and function (Kloeckener-Gruissem *et al.*, 1995).

1.8 Grain Texture and Its Importance in Cereal Crops

Grain texture is one of the most important traits directing the end use of cereal crops. This is largely accounted for by different adhesion properties between the starch granule and surrounding protein matrix causing hard textured grain to sustain more damage during the milling process (Barlow *et al.*, 1973; Glenn *et al.*, 1990; Brennan *et al.*, 1996). Damaged grain facilitates water absorption generating superior dough elasticity and loaf volume desired for the bread-making industry (Pomeranz *et al.*, 1984). In contrast, the even particle size and intact granules of soft wheat flour are ideal for products requiring low moisture content, such as cookies, cakes, and pastries (reviewed in Morris *et al.*, 1996). Barley grain texture has been shown to correlate with several aspects of malting quality, including hot water extract (Allison, 1986; Brennan *et al.*, 1996; Thomas *et al.*, 1996), and is one of the top characteristics selected for in both the brewing and distilling industries.

Grain texture is believed to be controlled by a single major locus of simple inheritance (*Ha*) located on the distal end of the short arm of cereal homeologous group 5 chromosomes (Symes, 1965; Mattern *et al.*, 1973; Law *et al.*, 1978). The presence of a 15 kDa protein band, termed friabilin or grain softness protein (*GSP*), isolated from water-washed starch has demonstrated perfect association with the dominant soft textured

phenotype in a global sampling of cultivated hexaploid wheat (*Triticum aestivum*) and several other cereal species (Symes, 1965; Greenwell *et al.*, 1986; Bakhella *et al.*, 1990; Matsoukas *et al.*, 1991; Oda *et al.*, 1992; Morrison *et al.*, 1992; Labuschagne *et al.*, 2000). This correlation provided a useful biochemical marker to assist in cultivar selection and focused the search for individual components of texture control. Characterization of friabilin revealed a complex of polypeptides composed of two major and several minor components: puroindoline-a (*pina*), puroindoline-b (*pinb*), grain softness protein (*GSP-1*), and an α-amylase inhibitor (Oda *et al.*, 1992; Jolly *et al.*, 1993; Morris *et al.*, 1994; Oda, 1994; Rahman *et al.*, 1994; Oda *et al.*, 1997). Both puroindolines, *GSP*, and their barley orthologs have consistently demonstrated tight linkage to the hardness (*Ha*) locus (Rouves *et al.*, 1996; Jolly *et al.*, 1996; Sourdille *et al.*, 1996; Giroux *et al.*, 1997; Beecher *et al.*, 2001).

Haplotype analysis of friabilin's major components, *pina* and *pinb*, revealed perfect correlation with wheat grain texture (Giroux *et al.*, 1997; Giroux *et al.*, 1998; Morris *et al.*, 2001). In each instance, mutational change of the wild type allele caused a phenotypic change from soft to hard texture. This unbreakable association coupled with the lack of recombination between puroindolines and the *Ha* locus indicated direct involvement of puroindolines in determining grain texture in wheat. Recently, this conclusion was strengthened when transgenic rice plants containing one or both of the puroindoline genes were shown to have a soft texture compared to the hard wild type rice grain (Krishnamurthy *et al.*, 2001b). Further transformation experiments involving the complementation of hard textured wheat with the wild type puroindoline-b allele also showed renewal of the dominant soft phenotype (Beecher *et al.*, 2002a).

The mechanism of action of the puroindolines is unknown. However, several proteins speculated to be involved in microbial defense mechanisms i.e. thionins, α-amylase inhibitors, and non-specific lipid transfer proteins (ns-LTPs) share similar characteristics

to puroindolines in that they are basic, cysteine-rich, low in molecular weight, translated as precursors, and involved in membrane-lipid interaction (Blochet *et al.*, 1993; Gautier *et al.*, 1994; Kooijman *et al.*, 1997; Dubreil *et al.*, 1997; Le Guerneve *et al.*, 1998). These similarities have led scientists to suggest that puroindolines may also play a role in plant defense (Blochet *et al.*, 1993). This was confirmed when puroindoline-a and -b alleles demonstrated significant inhibition *in vitro* for 4 out of 5 fungi tested in potato dextrose broth at dosages lower than 100 μg/ml (Dubreil *et al.*, 1998). In addition, rice plants containing puroindoline transgenes exhibited 53.5% and 22.3% increases in resistance to rice blast (*Magnaporthe grisea*) and sheath blight (*Rhizoctonia solani*) respectively (Krishnamurthy *et al.*, 2001a).

1.9 Thesis Objectives

Over the past decade, extensive EST collections and whole genome sequencing technology have largely overcome gene discovery as a rate limiting step in understanding diversity between and within organisms. The new challenge is to utilize the natural patterns of nucleotide diversity to identify causative polymorphisms responsible for phenotypic variation in complex quantitative traits. As mentioned in Section 1.6, despite significant advances, traditional linkage mapping methods still present several limitations, including low resolution and time constraints imposed by the development of appropriate mapping populations. Significant advances in human research have projected LD mapping as an alternative association approach and several successful studies have been documented (Kerem *et al.*, 1989; Corder *et al.*, 1994; Hugot *et al.*, 2001; Ogura *et al.*, 2001b). However, the application of LD mapping in plants is still in its infancy. To date, these studies have largely been restricted to the model inbreeding plant *Arabidopsis thaliana* and the outbreeding crop *Zea mays* (reviewed in Gaut *et al.*, 2003; Flint-Garcia *et al.*, 2003).

The primary objective of this thesis was to use the region surrounding the hardness (*Ha*) locus in barley as a candidate region of commercial interest to assess the feasibility of LD mapping for fine-scale association studies in inbreeding crop species. In order to achieve this goal, several complementary inter-related approaches were involved:

A) Exploiting the BAC libraries to produce a physical contig of the region harboring the hardness locus and determine the local genome organization and gene content.

B) Establishing a better understanding of gene structure and evolutionary history through comparative genomics with rice and *Arabidopsis*.

C) Determining the patterns of nucleotide diversity within coding, 3' flanking, and regulatory regions in genes identified across the physical contig.

D) Examining haplotype diversity and putative signatures of selection in different genepools represented by *H. vulgare* cultivars and landraces and *H. spontaneum* accessions.

E) Determining the extent and magnitude of LD across the physical region and relate empirical estimates of LD to genome composition and the varying evolutionary history of different genepools.

F) Synthesizing the information obtained from these studies to assess the feasibility of association mapping to locate genes controlling complex traits, facilitate map-based cloning, and identify appropriate markers for marker-assisted selection.

CHAPTER 2: MATERIALS AND METHODS

2.1 Polymerase Chain Reaction (PCR)

A reaction mixture of 0.3 µL 3.5U/µL expand HiFi (Roche, Switzerland), 2 µL 10X PCR buffer (supplied with enzyme), 0.4 µL 10 mM dNTPs, 0.5 µL each 10 µM forward and reverse primer, 14.3 µL sterile distilled water (SDW), and 2 µL 50 ng/µL template was prepared. The samples were subjected to 2 minutes 94°C; 40 cycles 30 seconds 94°C, 30 seconds 55°C, 1 minute 72°C; and 10 minutes 72°C in an ABI 9700 thermocycler (PE Applied Biosystems, USA). An aliquot of 2-5 µL of sample was separated on a 1% agarose 1% TBE (Tris-Boric acid- EDTA) gel with ethidium bromide (0.2 ng/mL) for visualization.

2.2 BAC Library Screen

2.2.1 Generation of Hybridization Probe

Primers for the amplification of hordoindoline-a (5'GGTCTGCTTGCTTTGGTAGC3' and 5'AATAGTGCTGGGGATGTTGC3') and –b (5'CTCCTAGCCCTCCTTGCTCT3' and 5'CTCCCATGTTGCACTTTGAG3') were designed from Genbank accessions HVU249929 and HVU249928, respectively using Primer3 software (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). Primers for the amplification of *GSP* (5'CAACATTGACAACATGAAGACC3' and 5'TTTGGCACAACTAACATTGG3') were designed from the Morex BAC122.a5 sequence. Probes were amplified as described in Section 2.1 and cleaned up by QIAquick PCR Purification Kit (Qiagen, Germany) as recommended by the manufacturer. Products were sequenced as described in Section 2.7.1 for verification of correct amplification. Hordoindoline primers were sufficient to ensure amplification of both paralogs.

2.2.2 Oligolabelling of DNA Probes

An aliquot of 2 µL DNA template (20-50 ng), 3 µL random 9mers (0.1 mg/mL), and 3.5 µL SDW were combined in a sterile Eppendorf tube and incubated at 95°C for

3 minutes. After snap-cooling on ice for an additional 3 minutes, 2.5 µL 200 µM dNTPs, 10 µL 2.5X labeling buffer (1 mg/mL BSA; 0.125 M Tris-HCl, pH 8.0; 12.5 mM $MgCl_2$, 12.5 mM dithiothreitol; and 0.5 M HEPES pH 6.6), 3 µL α^{32}P-dCTP (10 µCi/µL), and 1 µL Klenow DNA polymerase (Promega, USA; 0.1 mg/mL) were added. The reaction mixture was homogenized by gentle mixing and incubated at room temperature overnight.

2.2.3 Radioactive Probe Clean-up

A small amount of sterile glass wool was tightly packed to a minimum of 0.1 mL into the barrel of a 1 mL syringe. P-TEN biogel (10 mM Tris-HCl, ph 8.0; 1 mM EDTA, pH 8.0; 0.1M NaCl; 6% biogel) was added to the column for a final volume of 0.8 mL after packing. To ensure even packing, the column was centrifuged twice at 3220 x g (4000 rpm) for 4 minutes in an Eppendorf 5810R (Eppendorf, Germany). An aliquot of 200 µL TEN buffer (10 mM Tris-HCl, ph 8.0; 1 mM EDTA, pH 8.0; 0.1M NaCl) was added to the probe and the entire label reaction was transferred to the column. The purified probe was collected by centrifugation at 4000 rpm for 4 minutes.

2.2.4 Hybridization and Detection

Membranes were pre-wetted with 5X SSC (750 mM sodium chloride, 75 mM sodium citrate, pH 7.0) and inserted into 300 mL hybridization bottles with 30 mL of hybridization solution (0.5 M sodium phosphate, 7% SDS, 1 mM EDTA) and 300 µL of 10 mg/mL denatured salmon sperm DNA. Filters were incubated in a hybridization oven overnight at 65°C. Pre-hybridization solution was discarded and 15 mL fresh hybridization solution and 300 µL denatured salmon sperm DNA were added to the hybridization buffer. The labeled probe was denatured for 5 minutes and snap-cooled on ice prior to addition. Hybridization was incubated overnight at 65°C in a rotating hybridization oven. Membranes were washed in 2X SSC and 1X SSC for 20 minutes at 65°C each. Membranes were dabbed with a paper towel to eliminate residual wash buffer and wrapped in cling film. Filters were exposed to film in a cassette containing an

intensifying screen and stored at -80°C overnight. Film was exposed in a Xograph Imaging Systems Compact X4 automatic developer (Xograph Imaging Systems, UK).

2.2.5 Validation of Positive BACs

Purified BAC DNA was obtained using the Qiagen Large Construct Kit (Qiagen, Germany) as recommended by the manufacturer. A 20 µL reaction mix containing 1 µL *Hin*dIII (Promega, USA; 10 U/µL), 2 µL restriction buffer (supplied with enzyme), and 200-300 ng of template DNA was incubated at 37°C for 3 hours. Digestions were separated on a 1% agarose in 1X TBE gel with ethidium bromide (0.2 ng/mL) at 70 V for 8 hours. Gel was submersed in 0.25 M HCl for 10 min, denaturing solution (1.5 M NaCl, 0.5 M NaOH) for 45 min, and neutralizing solution (3.0 M NaCl, pH 7.0; 0.5 M Tris) for 30 minutes under gentle agitation. A brief rinse with SDW occurred between each step. The blotting stack was assembled in a plastic container from bottom to top in the following fashion: inverted gel tray, layer of 3 mm Whatman paper extending beyond the bottom of the gel tray to the floor of the container, layer of 3 mm Whatman paper slightly larger than the gel, agarose gel with wells facing downwards, Hybond-N nylon membrane same dimensions as gel, 3 layers of 3 mm Whatman paper slightly larger than the gel, stack of absorbent paper towels, and 500-1000 g weight distributed even across the stack. Layers of Whatman paper were previously saturated with 6X SSPE (Sodium chloride-Sodium phosphate-EDTA). A volume of 20X SSPE was added to the container until it reached half way up the sides of the inverted gel tray. Transfer of DNA from the gel to the membrane was allowed to transpire overnight. Stack was disassembled and the membrane was suspended in 6X SSPE for 5 minutes with gentle agitation. Excess solution was removed by blotting membrane on dry Whatman paper. Membrane was subjected to UV cross-linking twice. Generation of hybridization probes was performed as described in Section 2.2.1. Probes were quantified on a UV Spectrometer (Nanodrop Technologies, USA). Labeling of probe, hybridization, post-hybridization washes, signal generation, and detection were performed as recommended by Amersham Pharmacia

Biotech AlkPhos Direct Kit protocol (Amersham, UK). Hybridization and wash temperatures were raised to 65°C.

2.3 Sizing of BACs

A 20 µL reaction mix containing 1 µL *Not*I (Promega, USA; 10 U/µL), 2 µL restriction buffer (supplied with enzyme), and 100-200 ng of template DNA was incubated at 37°C for 3 hours. Digestions were separated overnight on a 1% agarose in 1X TAE (Tris-Acetate-EDTA) gel using a BioRad Chef Mapper (BioRad, USA) with the following parameters: 20 seconds switching time, 120° angle, 6.0 V/cm, and 14°C. Gel was stained with ethidium bromide for visualization of DNA.

2.4 BAC End Sequencing

A sequencing reaction mixture was prepared with 8 µL Big Dye terminator reaction mix (PE Applied Biosystems, USA), 2 µL 10 µM primer, and 10 µL of sample (200-400 ng). The samples were subjected to 2 minutes denaturing at 98°C followed by 100 cycles of 96°C for 30 seconds, 50°C for 20 seconds, and 60°C for 4 minutes in an ABI 9700 thermocycler (PE Applied Biosystems, USA). Aliquots of 50 µL 95% ethanol and 2 µL 3 M NaOAc, pH 4.6 were added to each reaction and the samples were mixed by inversion. After incubation at room temperature for 15 minutes, the plates were centrifuged at 3220 x g (4000 rpm) for 30 minutes in an Eppendorf 5810R centrifuge (Eppendorf, Germany). An aliquot of 1 µL loading buffer (1:4 v/v loading dye to deionised formamide) was added to each sample and the samples were loaded on a 377 ABI automatic sequencer (PE Applied Biosystems, USA).

2.5 Construction of BAC Nebulized Library

2.5.1 Preparation of Culture

A bacterial starter culture was generated by inoculating 10 mL LB (Luria-Bertani broth) medium with one colony from a freshly streaked antibiotic selective plate and incubating

at 37°C for 8 hours with 300 rpm shaking. The 1 mL of starter culture was introduced to 500 mL LB medium with the appropriate antibiotic selection and incubated at 37°C for 16 hours with agitation at 300 rpm. The flask used was able to hold at least four times the volume of the culture in order to assure sufficient aeration and agitation for optimal cell growth.

2.5.2 Maxi-Prep of Culture

The bacteria culture was transferred to centrifugation bottles and spun for 15 minutes at 6000 x g in a Sorvall RC5C centrifuge (Sorvall, USA) using the SLA-1500 rotor. The supernatant was drained and 36 mL of resuspension buffer (50 mM glucose, 25 mM Tris-HCl, and 10 mM EDTA) was added to the pellets. After total resuspension of the bacteria pellet was ensured by vortexing, 4 mL of 40 mg/mL lysozyme were added and the solution was incubated for 10 minutes at room temperature. The suspension was transferred to ice and 80 mL of lysis buffer (0.2 M NaOH, 1% EDTA) was added. The suspension was inverted several times to ensure even mixing and incubated on ice for 10 minutes. An aliquot of 40 mL neutralization buffer (5 M KOAc) was added and the suspension was mixed to homogeneity through inversion. After a 15 minute incubation period on ice, the suspension was centrifuged at 20,000 x g for 20 minutes. The supernatant was decanted from the cell debris through Miracloth™ into a clean sterile centrifuge bottle. Isopropanol was added at 0.6 volumes and the solution was inverted for even mixing. The sample was spun at 20,000 x g and 15°C for 30 minutes. The supernatant was decanted and 100 mL of 70% ethanol was added to the pellet. After centrifuging at 15,000 x g rpm for 10 minutes the ethanol was decanted carefully to avoid dislodging the pellet. The remaining traces of ethanol were aspirated out and the pellet was allowed to sit for several minutes to dry. The dried pellet was resuspended in 8 mL of TE (10 mM Tris-HCl, pH 8.0; 1 mM EDTA, pH 8.0).

2.5.3 Cesium Chloride Extraction

A 100 µL aliquot of ethidium bromide from a 10 mg/mL stock and 1 gram of cesium chloride were added for every milliliter of sample. The sample was transferred to suitable tubes for high speed centrifugation. A 1 g/mL solution of cesium chloride was used to fill the remaining volume of the tubes. The samples were centrifuged at 56,000 rpm overnight in a Beckman ultracentrifuge (Beckman Coulter, Germany) using the VTI65 rotor. Two distinct bands corresponding to nicked circular or linear DNA (top band) and closed circular plasmid DNA (bottom band) were observed. The needle of a syringe was inserted into the top of the tube for air intake. Each of the two bands was then extracted with separate needles and stored in Eppendorf tubes. One volume isopropanol/20X SSPE was added. The sample was mixed, allowed to settle, and the top layer was removed and discarded. This step was repeated at least 4 more times until all the ethidium bromide was extracted. The sample was then transferred to dialysis tubing and placed in one liter 0.1X TE (pH 8.0) at 4°C. The TE buffer was changed once and continually stirred overnight. The sample was transferred to a clean sterile Eppendorf tube, precipitated (2.5 volumes of ethanol, 0.1 volumes 3 M NaOAc, and 1 µL/mL glycerol) at -20°C for 30 minutes, and spun at 16,000 x g for 25 minutes in an Eppendorf 5415C tabletop centrifuge (Eppendorf, Germany). The supernatant was decanted, the pellet was rinsed with 1 mL of 70% ethanol, and centrifuged at 16,000 x g for 10 minutes. The supernatant was drained and the pellet was resuspended in 50 µL TE (pH 8.0) after being allowed to dry. Quantification was performed on a UV Spectrometer (Nanodrop Technologies, USA). The maxi-prep and cesium chloride extraction can be bypassed by using the Qiagen Large Construct Kit (Qiagen, Germany) as recommended by the manufacturer.

2.5.4 Digestion

The integrity of the sample was tested through *Not*I and *Hin*dIII digestion. A solution containing 5 µL of sample, 1 µL enzyme (Promega, USA; 10 U/µL), 2 µL 10X reaction buffer (supplied with enzyme), and 12 µL SDW was prepared and incubated at 37°C for

2 hours. The sample was separated on a 1% agarose in 1X TBE gel with ethidium bromide (0.2 ng/mL) for visualization.

2.5.5 Nebulization

Between 3 and 5 μg of sample DNA was added to 1 mL of a 30% glycerol solution. The solution was transferred into a Sidestream nebulizer and incubated on ice for 20 minutes. The nebulizer was then attached to a nitrogen gas chamber and the sample was subjected to 10 PSI for 15 seconds. An aliquot of 20 μL of sample was separated through electrophoresis on a 1% agarose in 1X TBE gel with ethidium bromide (0.2 ng/mL) to ensure that the majority of the resulting fragments were between 1 and 5 kilobases. After dividing the sample into two equal parts for precipitation in 1.5 mL Eppendorf tubes, 40 μL 5 M NaOAc and 1 mL 100% ethanol were added to each tube and the samples were incubated at 20°C for 30 minutes. The sample was pelleted through centrifugation at 16,000 x g for 25 minutes in an Eppendorf 5415C tabletop centrifuge (Eppendorf, Germany). The resulting pellets were washed twice with 100% ethanol; each was followed by a 10 minute spin at 16,000 x g. The sample was then allowed to dry and resuspended in 20 μL SDW. The two halves of the sample were recombined for the next stage of the protocol.

2.5.6 Blunt Ending

A polishing solution containing 40 μL sample, 5 μL 2 mM dNTPs, 3 μL *Pfu* DNA polymerase (Promega, USA; 3 U/μL), and 5 μL 10X polishing buffer (supplied with enzyme) was prepared. The solution was incubated at 72°C for 30 minutes.

2.5.7 Size Fractionation

The sample was loaded in one large well of a 1% low melting point agarose in 1X TBE gel with ethidium bromide (0.2 ng/mL) and resolved at low voltage to ensure good separation. The desired size fragments were cut out of the gel and transferred to 15 mL

Falcon tubes. The DNA was extracted from the gel using QIAquick Gel Extraction Kit (Qiagen, Germany) following the manufacturer's protocol. The sample concentration was quantified on a UV Spectrometer (Nanodrop Technologies, USA).

2.5.8 Preparation of Linear Plasmid DNA

Circular plasmid pUC18 DNA (10 µg) was digested with 6 µL *Sma*I (Promega, USA; 10 U/µL) at 30°C for a minimum of one hour. Linear and circular plasmids were compared side by side on a 1% agarose in 1X TBE gel with ethidium bromide (0.2 ng/mL) to ensure complete digestion. After the digestion was purified using a QIAquick PCR Purification Kit (Qiagen, Germany) following the manufacturer's protocol, the sample was resuspended in 90 µL SDW and 10 µL 10X dephosphorylation buffer (Promega, USA) was added. One unit of calf intestine alkaline phosphatase (CIAP; Promega, USA) was added for every 2 pmoles of plasmid DNA termini and the reaction was incubated at 37°C. After 15 minutes a second aliquot of CIAP was added and the reaction temperature was raised to 55°C for 45 minutes. The sample was purified using a QIAquick PCR Purification Kit (Qiagen, Germany) and resuspended in 50 µL SDW. Vector concentration was determined on a UV Spectrometer (Nanodrop Technologies, USA).

2.5.9 Ligation

The ligation solution was prepared by combining 7 µL of sample (100-200 ng), 0.5 µL pUC18 cloning vector (Invitrogen, USA; 50 ng), 1 µL T4 DNA ligase (Promega, USA; 3 U/µL), 1 µL 10X ligation buffer (supplied with enzyme), and 0.5 µL SDW in a 1.5 mL Eppendorf tube. The solution was incubated at 4°C overnight.

2.5.10 Transformation

An aliquot of 1 µL ligation mix was pipetted into 20 µL DH10b electrocompetent cells (Invitrogen, USA). The sample was transferred to a 0.1 cm electroporation cuvette and

electroporated at 200 Ω, 25 μF, and 1.8 kV. The sample was suspended in 1 mL SOC medium (2% w/v bactotryptone, 0.5% w/v yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, and 20 mM glucose) and incubated at 37°C with 300 rpm shaking for an hour. Glycerol was added to the sample so that the final concentration was 15%. Aliquots of 20 and 50 μL were plated out on LB-ampicillin (100 μg/mL) plates with ITPG and X-Gal and incubated at 37°C for 16 hours in order to calculate the titer of the library. The remaining sample was stored at -80°C.

2.6 Library Handling

2.6.1 Library Plating and Picking

An aliquot of the nebulized BAC library was plated out on ampicillin (100 μg/mL) selective plates with IPTG (100 μg/mL) and X-Gal (20 μg/mL). The plates were incubated at 37°C overnight. The white colonies were picked into 384 well culture plates with 60 μL LB-ampicillin (100 μg/mL) freezing media per well and incubated at 37°C overnight.

2.6.2 Culture Preparation

2.6.2.1 Method A: Basic Alkaline Lysis

Deep well blocks containing 1 mL LB-ampicillin (100 μg/mL) were inoculated with 1 μL of culture and incubated at 37°C overnight with 300 rpm shaking. The inoculation blocks were centrifuged at 2576 x g (4000 rpm) for 10 minutes in a Sigma 4-15C centrifuge (Sigma, Germany) and the supernatant was drained from the pellets. An aliquot of 200 μL resuspension buffer (40 U/mL RNAsc T1; 40 μg/mL RNAse A; 10 mM EDTA, pH 8.0; 50 mM Tris-HCl, pH 7.6) was added to the samples and vortexed to homogeneity. After the addition of 200 μL of lysis buffer (20% SDS, 2 M NaOH), the sample was mixed through inversion and allowed to stand for 5 minutes. The solution was then neutralized with 200 μL 3M NaOAc (pH 4.8), inverted to mix, and incubated at -20°C for 30 minutes. The sample was centrifuged at 2576 x g for 45 minutes at 4°C.

The 200 µL of supernatant was transferred to a clean deep well block and 500 µL of 100% ethanol was added to each well. The block was sealed, inverted several times to mix, and incubated at -20°C for 30 minutes. The ethanol was decanted and the deep well block was spun inverted at 161 x g (1000 rpm) for 1 minute to remove excess alcohol. The pellets were resuspended in 30 µL of SDW. In order to quantify and check the integrity of the DNA and to ensure the plasmids contain inserts, the samples were digested with *Bam*H1 and *Eco*R1 (Promega, USA). A solution containing 2 µL of sample, 1 µL 10X reaction buffer (supplied with enzyme), 0.5 µL of each enzyme (10 U/µL), and 6 µL SDW was prepared and incubated at 37°C for 2 hours. The digestions were analyzed on 1% agarose 1% TBE gels with ethidium bromide (0.2 ng/mL).

2.6.2.2 Method B: Millipore Multiscreen Plasmid Preparation

Deep well blocks containing 1 mL 2X LB-ampicillin (100 µg/mL) were inoculated with 5 µL of culture and incubated at 37°C overnight with 300 rpm shaking. The inoculation blocks were centrifuged at 3220 x g (4000 rpm) for 10 minutes in an Eppendorf 5810R centrifuge (Eppendorf, Germany) and the supernatant was drained from the pellets. An aliquot of 80 µL pre-chilled resuspension buffer (30 mM glucose; 15 mM Tris-HCl, pH 8.0; 30 mM EDTA, pH 8.0; and 60 µg/mL RNAse A) was added to the samples and vortexed to ensure complete resuspension. After addition of 80 µL fresh lysis buffer (0.2 M NaOH, 1% SDS), samples were vortexed for 1 minute and then incubated at room temperature for an additional 2 minutes. The samples were then neutralized with 80 µL neutralization buffer (3.6 M potassium, 6 M acetate) and vortexed for 1 minute. The entire sample was transferred to a Multiscreen MANANLY clearing plate (Millipore, USA) and subjected to 10" Hg pressure in a vacuum manifold. The cleared lysate was dispensed into a MAFBNOB binding plate (Millipore, USA) already containing 160 µL binding solution (8 M guanidine hydrochloride) located at the bottom of the vacuum manifold. The samples were mixed by pipetting up and down 3 times and the

MAFBNOB plate was transferred to the top of the vacuum manifold. Full vacuum pressure was applied to the samples for 1 minute. The samples were washed twice with 200 µL 20% ethanol each time applying full vacuum pressure. Vacuum was applied at full pressure for an additional 3 minutes to remove all residual ethanol. The binding plate was secured to a sterile microtiter plate utilizing Millipore alignment frames (MACF09604). Membranes were dried by centrifugation at 1811 x g (3000 rpm) for 10 minutes and incubated at room temperature for 10 minutes. Plasmid DNA was resuspended in 100 µL SDW and eluted by centrifugation at 1811 x g (3000 rpm) for 10 minutes. In order to quantify and check the integrity of the DNA and to ensure the plasmids contain inserts, the samples were digested with *Not*I (Promega, USA). A solution containing 2 µL of sample, 1 µL 10X reaction buffer (supplied with enzyme), 0.5 µL enzyme (10 U/µL), and 6.5 µL SDW was prepared and incubated at 37°C for 2 hours. The digestions were analyzed on 1% agarose in 1X TBE gels with ethidium bromide (0.2 ng/mL).

2.6.3 Pouring of Polyacrylamide Gels for 377 ABI Sequencer

A gel mix of 18 g urea, 5.2 mL acrylamide/bis-acrylamide, 5.0 µL 10X TBE, and 25 µL SDW was prepared and stirred for 10 minutes with a small amount of Amberlite™. The gel mix was filtered and allowed to degas for 30 seconds before adding 250 µL 10% APS and 35 µL TEMED. The gel was mixed to homogeneity and poured along the bottom edge of the back plate and between the 0.2 mm spacers that line the sides of the plate. The top plate was aligned with the back plate at the bottom edge and slowly lowered to ensure no bubbles would be created in the gel between the two plates. After the top plate was lowered fully onto the bottom plate, "bulldog" clips were used to securely hold them together flush and the comb was inserted tooth side outwards into the well at the top of the plate. The gel was allowed to stand for 2 hours until complete polymerization had occurred. The plate was washed of all residual acrylamide and the comb was removed from between the plates and reinserted tooth side inward.

2.7 Sample Sequencing

2.7.1 Method A: Quarter Reactions and NaOAc Precipitation

A sequencing reaction mixture was prepared with 2 µL Big Dye terminator reaction mix (PE Applied Biosystems, USA), 0.25 µL 10 µM primer, 0.75 µL SDW, and 2 µL of sample (100-200 ng). The samples were subjected to 40 cycles of 96°C for 10 seconds, 50°C for 5 seconds, and 60°C for 4 minutes in an ABI 9700 thermocycler (PE Applied Biosystems, USA). Sequences were precipitated using NaOAc method described in Section 2.4. An aliquot of 1 µL loading buffer (1:4 v/v loading dye to deionised formamide) was added to each sample before loading on an ABI 377 automatic sequencer (PE Applied Biosystems, USA).

2.7.2 Method B: Eighth Reactions and GENETIX genCLEAN Plates

A sequencing reaction mixture was prepared with 1 µL Big Dye terminator (PE Applied Biosystems, USA), 1.0 µL 10 µM primer, 4.25 µL SDW, 0.8 µL sequencing buffer (400 mM Tris-HCl, pH 9.0; 10 mM $MgCl_2$), and 3 µL of sample (50-100 ng). Reactions were assembled using a Biomek 2000 (Beckman Coulter, Germany). The samples were subjected the same cycle conditions mentioned in Section 2.7.1. Samples were prepared for loading using genCLEAN plates (GENETIX, UK) as recommended by the manufacturer and loaded directly onto an ABI 3700 automatic sequencer (PE Applied Biosystems, USA).

2.8 BAC Nebulized Library Assembly

Preassembly and assembly analysis of the sequencing reads were performed by using Phred Version 0.020425.c and Phrap Version 0.990329 software (Ewing *et al.*, 1998a; Ewing *et al.*, 1998b). The combined information was viewed and edited through consed Version 12.0 software (Gordon *et al.*, 1998). Gaps were closed and weak consensus regions strengthened by either direct sequencing of subclones using nested primers or

sequencing PCR amplicons spanning the region between contig ends. PCR was performed as described in Section 2.1. Extension time was increased to 2-4 minutes for amplification of larger products.

2.9 Sequence Characterization

Preliminary characterization of the complete full-length sequence of the region containing the *Ha* locus was preformed using standard nucleotide-nucleotide (BLASTN) and nucleotide-protein (BLASTX) BLAST (Altschul *et al.*, 1997) searches against the non-redundant database (dbnr) at the National Center of Biotechnology Information (NCBI, http://ncbi.nlm.nih.gov/BLAST/) and the Triticeae Repeat Sequence Database (TREP, http://wheat/pw.usda.gov/ggpages/ITMI/Repeats/balstrepeats3.html; Wicker *et al.*, 2002). Inverted and direct repeats of previously uncharacterized elements were detected through Bestfit analysis using WebANGIS (http://www.angis.org.au/WebANGIS/WebFM). SINEs were detected by scanning the genomic sequence for similarity to the conserved *Arabidopsis* A (TRKYNNARNGG) and B (RGTTCRANHYY) boxes spaced 25 to 50 bp apart (Myouga *et al.*, 2001). Initial gene prediction analysis was performed using the Rice Genome Automated Analysis System (RiceGAAS, http://ricegaas.dna.affrc.go.jp/; Sakata *et al.*, 2002) which couples the integration of several programs for the prediction of open reading frames (GENSCAN, RiceHMM, FGENESH, MZEF) with homology search analysis programs (BLAST, HMMER, ProfileScan, MOTIF). Expression of putative genes was determined using BLASTN analysis against the EST database (dbest) at the NCBI. Exon:intron splice junctions were determined by genomic alignment with Triticeae ESTs. Where barley and wheat ESTs were insufficient to cover the entire length of the predicted protein, ESTs from other grass species, e.g. rice and sorghum, were utilized. Splice junctions were confirmed by the presence of the conserved GT and AG intron boarders and a minimum of 5 of the 9 (5' CAG:GTAAGT 3') and 3 of the 5 (5' GCAG:G 3') consensus nucleotides for the respective exon:intron and intron:exon splice sites in plants. Putative functions and conserved protein domains were determined using

BLASTP analysis against the non-redundant (nr) and swissprot databases at NCBI. Identification of colinear and homologous *Arabidopsis* and rice sequences were performed at the *Arabidopsis* Information Resource (TAIR, http://www.arabidopsis.org/Blast/), The Institute for Genomic Research (TIGR, http://tigrblast.tigr.org/euk-blast/index.cgi?project=osa1), and GRAMENE websites (http://www.gramene.org/). The Dotter program (Sonnhammer *et al.*, 1995; word length 25, similarity 80) was used to identify conserved regions of sequence homology between the barley BAC contig and the rice colinear sequence (Genbank Accession AL928743).

2.10 Plant Material

Plant material was grown from seed in 50 cc soil volume and transplanted after 2 weeks into 6 inch diameter round pots (2000 cc soil volume). Soil mixture was a composition of peat (1200 L), sand (100L), calcium limestone (2.5 kg), magnesium limestone (2.5 kg), sincrostart base fertilizer (1.5 kg; Sinclair Ltd, UK), ficote 70 controlled release fertilizer (3.0 kg; Fisons Ltd, UK), celcote water retaining gel (1.5 kg; Hortichem Ltd, UK), and intercept systemic insecticide (0.39 kg; Levington, UK). Plants were sustained in SCRI glasshouses under the following conditions: 16 hour day length (8:00 am – 12:00 am) at 20°C and 8 hours night length (1:00 am – 7:00 am) at 15°C. No solar shading was applied. Germplasm was selected with the help of Bill Thomas based on unpublished extensive SSR diversity studies. A record of germplasm used can be seen in Table 2.1.

Table 2.1. Accessions and geographical origin of germplasm sampled and genotyped. When available, phenotypic information i.e. spring vs. winter, two- vs. six-row, and malting vs. feed is provided.

Species	Accession	Origin	Season	Morphology	Commercial Use
H. vulgare (cv)	Abed Binder		spring	two-rowed	
H. vulgare (cv)	Alexis	Germany	spring	two-rowed	malting
H. vulgare (cv)	Aramir	Netherlands	spring	two-rowed	malting
H. vulgare (cv)	B83		spring		
H. vulgare (cv)	Bavaria	Germany	spring	two-rowed	
H. vulgare (cv)	Bernice	France	spring	two-rowed	
H. vulgare (cv)	Borwina		winter	six-rowed	
H. vulgare (cv)	Carlsberg	Denmark	spring	two-rowed	

Table 2.1 cont.

H. vulgare (cv)	Carlsberg II		spring	two-rowed	
H. vulgare (cv)	Casino		spring	two-rowed	
H. vulgare (cv)	Chime (AB)		spring	two-rowed	malting
H. vulgare (cv)	Clipper	Australia	spring	two-rowed	malting
H. vulgare (cv)	Criewener 403		spring	two-rowed	
H. vulgare (cv)	Derkado	Germany	spring	two-rowed	malting
H. vulgare (cv)	Europa		spring	two-rowed	
H. vulgare (cv)	Fanfare		winter		malting
H. vulgare (cv)	Franka		winter	six-rowed	
H. vulgare (cv)	Friedrichswerther Berg		winter	six-rowed	
H. vulgare (cv)	Galleon	Australia	spring	two-rowed	feed
H. vulgare (cv)	Georgie	Great Britain	spring	two-rowed	
H. vulgare (cv)	Golden Promise	Great Britain	spring	two-rowed	malting
H. vulgare (cv)	Gotlands	Sweden	spring	two-rowed	
H. vulgare (cv)	Gull	Sweden	spring	two-rowed	
H. vulgare (cv)	Haisa	Germany	spring	two-rowed	
H. vulgare (cv)	Hana		spring	two-rowed	malting
H. vulgare (cv)	Harrington	Canada	spring	two-rowed	malting
H. vulgare (cv)	Hatif de Grignon		winter	six-rowed	
H. vulgare (cv)	Heils Franken		spring	two-rowed	
H. vulgare (cv)	Ingrid	Sweden	spring	two-rowed	
H. vulgare (cv)	Isaria	Germany	spring	two-rowed	
H. vulgare (cv)	Kenia	Sweden	spring	two-rowed	feed
H. vulgare (cv)	Kneifel	Germany	spring	two-rowed	
H. vulgare (cv)	Lina		spring	two-rowed	
H. vulgare (cv)	Livet	Great Britain	spring	two-rowed	
H. vulgare (cv)	Maja	Sweden	spring	two-rowed	
H. vulgare (cv)	Marinka		winter	two-rowed	
H. vulgare (cv)	Maris Otter	Great Britain	winter	two-rowed	malting
H. vulgare (cv)	Melanie		winter	two-rowed	malting
H. vulgare (cv)	Monte Cristo	India	spring	six-rowed	
H. vulgare (cv)	Morex		spring	six-rowed	malting
H. vulgare (cv)	Natasha	France	spring	two-rowed	
H. vulgare (cv)	Olli		spring	six-rowed	
H. vulgare (cv)	Opal	Sweden	winter	two-rowed	feed
H. vulgare (cv)	Optic	Great Britain	spring	two-rowed	malting
H. vulgare (cv)	Pallas		spring	two-rowed	
H. vulgare (cv)	Plaisant		winter	six-rowed	malting
H. vulgare (cv)	Plumage Archer		spring	two-rowed	malting
H. vulgare (cv)	Prisma	Netherlands	spring	two-rowed	malting
H. vulgare (cv)	Proctor	Great Britain	spring	two-rowed	
H. vulgare (cv)	Puffin		winter	two-rowed	malting
H. vulgare (cv)	Ragusa		winter	six-rowed	
H. vulgare (cv)	Regina		winter	two-rowed	malting
H. vulgare (cv)	Rika	Sweden	spring	two-rowed	
H. vulgare (cv)	Romanze		winter	two-rowed	
H. vulgare (cv)	Scarlett		spring	two-rowed	malting
H. vulgare (cv)	Scots Bere		spring	six-rowed	
H. vulgare (cv)	Sewa		spring	two-rowed	
H. vulgare (cv)	Sonja		winter	two-rowed	
H. vulgare (cv)	Spratt Archer		spring	two-rowed	malting
H. vulgare (cv)	Static		spring	two-rowed	feed
H. vulgare (cv)	Sultan	Netherlands	spring	two-rowed	
H. vulgare (cv)	Svalof Svanhals		spring	two-rowed	
H. vulgare (cv)	Tankard	Great Britain	spring	two-rowed	malting
H. vulgare (cv)	Tern	Germany	spring	two-rowed	

Table 2.1 cont.

H. vulgare (cv)	Triumph		spring	two-rowed	
H. vulgare (cv)	Trumph		spring	two-rowed	malting
H. vulgare (cv)	Tschermaks 2 row		winter	two-rowed	
H. vulgare (cv)	Tyra		spring	two-rowed	
H. vulgare (cv)	Union		spring	two-rowed	
H. vulgare (cv)	Vada	Netherlands	spring	two-rowed	
H. vulgare (cv)	Valticky		spring	two-rowed	
H. vulgare (cv)	Vogelsanger Gold		winter	six-rowed	
H. vulgare (cv)	Volla	Germany	spring	two-rowed	
H. vulgare (cv)	Wisa	Germany	spring	two-rowed	
H. vulgare (lr)	Cyrrhus				
H. vulgare (lr)	NJSS101	Syria SW			
H. vulgare (lr)	NJSS111	Syria SW			
H. vulgare (lr)	NJSS121	Syria SW			
H. vulgare (lr)	NJSS141	Syria SW			
H. vulgare (lr)	WS231	Syria NW			
H. vulgare (lr)	WS241	Syria NW			
H. vulgare (lr)	WS281	Syria NW			
H. vulgare (lr)	CS301	Syria Cent.			
H. vulgare (lr)	SJ31	Jordan			
H. vulgare (lr)	SJ41	Jordan			
H. vulgare (lr)	NES421	Syria NE			
H. vulgare (lr)	SJ51	Jordan			
H. vulgare (lr)	SJ81	Jordan			
H. vulgare (lr)	SJ91	Jordan			
H. spontaneum	180044/HS5700	Afghanistan			
H. spontaneum	180046/HS5701	Iraq			
H. spontaneum	180049/HS5702	Iraq			
H. spontaneum	180052/HS5704	Iran			
H. spontaneum	181277/HS5738	Greece			
H. spontaneum	181498/HS5770	Uzbekistan			
H. spontaneum	180994/HS5835	Israel			
H. spontaneum	181243/HS5856	Pakistan			
H. spontaneum	181436/HS5865	Jordan			
H. spontaneum	181164	Iran, Station 10			
H. spontaneum	181170	Iran, Station 10			
H. spontaneum	181174	Iran, Station 10			
H. spontaneum	181267	Turkey, Gaziantep, Kilis			
H. spontaneum	181319	Iran			
H. spontaneum	181549	Syria, Sweida, 3 Km before Um Dobeib, coming from Dheqqa			
H. spontaneum	181679	Turkey, Gaziantep, 2Km from Al Baeyli North on the way to Gaziantep			
H. spontaneum	220664	Afghanistan, Herat, Herat			
H. spontaneum	245739	Turkey, Urfa, Ceylanpinar			
H. spontaneum	2691	Iran, FAO NO. 3897			
H. spontaneum	284738	Israel, NetivLamid Hei			
H. spontaneum	296791	Israel, Ashqilon National bank			
H. spontaneum	296860	Israel, Poriyya			
H. spontaneum	296874	Israel, Sha'ar Ephraim Jct			
H. spontaneum	296912	Israel, Misgav'Am			
H. spontaneum	391131	Israel, 'En Yorqeam, Northern Negev Lowlands, site 1			
H. spontaneum	391132	Israel, 'En Yorqeam, Northern Negev Lowlands, site 2			
H. spontaneum	391133	Israel, 'En Yorqeam, Northern Negev Lowlands, site 3			
H. spontaneum	391134	Israel, 'En Yorqeam, Northern Negev Lowlands, site 4			
H. spontaneum	391135	Israel, 'En Yorqeam, Northern Negev Lowlands, site 5			
H. spontaneum	39864	Sibirien, by MPIZColgne			
H. spontaneum	39870	Sibirien, by MPIZ Colgne			
H. spontaneum	466446	Israel, Rosh Pinna			
H. spontaneum	466470	Israel, Gadot			
H. spontaneum	HIS 34	Turkey, Dyarbakir, 38 KmWest			

2.11 DNA Extraction

Approximately 100-200 mg of leaf tissue was added to 500 µL of urea extraction buffer (7 M Urea; 0.35 M NaCl; 0.05 M Tris-HCl, ph 8.0; 0.02 M EDTA; 1% Sarkosyl) in 2.0 mL screw cap Eppendorf tubes. This tissue was homogenized with 3 mm glass beads in a Bio 101 Fast Prep bead mill (Q-biogene, USA) two times for 30 seconds at speed 6.5. The samples were quick spun for 30 seconds at 16,000 x g in an Eppendorf 5415C tabletop centrifuge (Eppendorf, Germany) and then incubated at 37°C for 30 minutes in a water bath. An equal volume of phenol:chloroform:isoamyl alcohol (25:24:1 v/v) was added and the sample was vortexed to homogenization. The samples were spun for 25 minutes at 16,000 x g to ensure good phase separation. The top phase was pipetted into a new clean 1.5 mL Eppendorf tube containing 2 µL of RNAse A (10 mg/mL) and incubated at room temperature for 20 minutes. A precipitation mixture of 1/10 volume of 3M sodium acetate (pH 4.8), equal volume of 100% isopropanol, and 1 µL of glycogen (20mg/mL) was added, mixed well, and allowed to incubate for 30 more minutes at -20°C. Samples were centrifuged at 16,000 x g for 25 minutes. After the supernatant was decanted, the pellet was rinsed with 1 mL of 70% ethanol and centrifuged again at 16,000 x g for 10 minutes. The supernatant was drained and the pellet was resuspended in 50 µL TE after being allowed to dry. Quantification was performed on a UV Spectrometer (Nanodrop Technologies, USA). Alternatively, DNeasy Plant Mini Kits (Qiagen, Germany) were employed as recommended by the manufacturer.

2.12 Primer Design and Amplification

Primers were designed to amplify the 5' flanking, coding, and 3' untranslated regions of the candidate grain texture genes and the majority of Exon 3 of PG2 (Putative Gene 2) using Primer3 software (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). A record of the primers used can be seen in Table 2.2. Within each gene region, the consecutive primer pairs were designed to overlap the flanking amplified region by a minimum of 50 bp in order to generate a contiguous

analyzed region. In some instances primer specification was not ideal for the amplification of particular haplotypes. These products were obtained using the forward primer of one primer pair and the reverse primer of the adjacent primer pair. Amplification was performed as described in Section 2.1.

Table 2.2. Summary of primers used for amplification of *GSP*, *hina*, *hinb*, and *PG2* gene regions.

Gene	Region	Direction	Sequence
GSP-1	5' flanking	forward	TGGTGAGCCATTGTTTATCG
GSP-1	5' flanking	reverse	CGAGAGCAAGGAAGGATAGG
GSP-1	5'flanking/gene	forward	CGTCTCAACAACTTGCGAAC
GSP-1	5'flanking/gene	reverse	CACGCATCGATCCATAACAT
GSP-1	coding	forward	CAACATTGACAACATGAAGACC
GSP-1	coding	reverse	TTTGGCACAACTAACATTGG
GSP-1	3' flanking	forward	TGTGACCAATAAATATGACAATGG
GSP-1	3' flanking	reverse	GCTTGCTGGAACTGACATCC
hina	5' flanking	forward	ACACGGAGAGAGAAGGGTCA
hina	5' flanking	reverse	CCTGCAGGAATCTAGCTTTG
hina	coding	forward	GGTCTGCTTGCTTTGGTAGC
hina	coding	reverse	AATAGTGCTGGGGATGTTGC
hina	3' flanking	forward	AACGTGATCTCGGTGGTTTC
hina	3' flanking	reverse	GGGTTCCAATTTCCGAACA
hinb-1	5' flanking	forward	TTTTGCAGAATACAAGGTAACATCA
hinb-1	5' flanking	reverse	GTTGAGAACCGCCTCCTCCG
hinb-1	coding	forward	CCACCAACACCAAAACAACG
hinb-1	coding	reverse	CGAGGGGAAACCCTTTCTCG
hinb-1	3' flanking	forward	TTCCCCTTGTATATTGGCCC
hinb-1	3' flanking	reverse	GGGTTTATTAGGACAAAGAGAG
hinb-2	5' flanking	forward	AGCCAAACCACCGCTTAACACC
hinb-2	5' flanking	reverse	TGAGAACCACCTCCTCCACC
hinb-2	coding	forward	CTTCACCAACACCAAATAAACA
hinb-2	coding	reverse	TGAGGGGAAACCATTTTTTG
hinb-2	3' flanking	forward	TCCGCTTGTATATTGGTCTGTG
hinb-2	3' flanking	reverse	GCGTGGTATGCCCTGTTG
PG2	Exon3	forward	GCAAGCTAGTCTGGGACACC
PG2	Exon3	reverse	CATCAATCAAGGCATGAACG

2.13 Amplicon Purification and Sequencing

2.13.1 Method A: Exonuclease Clean-up and Half Sequencing Reactions

An aliquot of 2 µL ExoSAP-IT (United States Biochemical Corporation, USA) was added to 2-4 µL of PCR product. The sample was subjected to 37°C and 80°C for 15 minutes each. Aliquots of 2 µL Big Dye terminator reaction mix (PE Applied Biosystems, USA),

0.25 µL 10 µM primer, 1.75-3.75 µL SDW were added to the sample and cycle sequencing and precipitation was performed as described in Section 2.7.1. An aliquot of 1 µL loading buffer was added to each sample before loading on an ABI 377 automatic sequencer (PE Applied Biosystems, USA).

2.13.2 Method B: genPURE Purification and Eighth Sequencing Reactions

genPURE 96 well plates (GENETIX, UK) were used as recommended by the manufacturer and eighth sequencing reactions were performed as described in Section 2.7.2. Samples were purified using genCLEAN plates (GENETIX, UK) as recommended by the manufacturer and loaded directly on an ABI3700 automatic sequencer (PE Applied Biosystems, USA).

2.14 Sequence Alignment and Nucleotide Analysis

Amplification products were aligned using Sequencher Version 4.1.4 (Gene Codes, USA). Estimates of nucleotide polymorphism (Watterson's estimate, θ_w; nucleotide diversity, π), neutrality test (Tajima's D, McDonald Kreitman, HKA), recombination (Hudson and Kaplan), and linkage disequilibrium (r^2) and their statistical significance were calculated using DnaSP Version 3.53 (Lewontin, 1964; Lewontin *et al.*, 1964; Watterson, 1975; Tajima, 1983; Nei, 1987; Hudson *et al.*, 1987; Tajima, 1989; Nei *et al.*, 1990; McDonald *et al.*, 1991; Tajima, 1993; Rozas *et al.*, 1999). Plots of all informative pairwise comparisons relative to physical distance were generated in Microsoft Excel. The median association value for each set of pairwise comparisons between sites located within two different gene regions was calculated and plotted against the corresponding median distance for each set using GenStat for Windows (2002. Release 6.2. Sixth Edition. VSN International ltd). Programming for GenStat calculations was kindly performed by Jim McNicol.

2.15 Mapping

GSP, *hina*, and *hinb-1* were mapped by single pass sequencing of 96 double haploid (DH) lines derived from the F_1 hybrid of the barley Steptoe x Morex mapping population (Kleinhofs *et al.*, 1993) using primers designed to the coding regions of the individual genes (see Section 2.12). Mapping was done with JoinMap Version 2.0 using the Steptoe x Morex population segregation data (http://wheat.pw.usda.gov/index.shtml) with a LOD score of 5.0 (Stam, 1993; Stam, 1995).

CHAPTER 3: SEQUENCE AND ANALYSIS OF THE REGION HARBORING THE *Ha* LOCUS IN BARLEY AND THE EXPLOITATION OF COMPARATIVE GENOMICS WITH THE COLINEAR RICE REGION

3.1 Introduction

Gene content among plant species appears to be relatively consistent ranging from 25,000-43,000 genes despite a 600-fold difference in genome size among angiosperm species (Bennett *et al.*, 1982; Bennett *et al.*, 1995; Miklos *et al.*, 1996; Bennett *et al.*, 1997). In the Gramineae, the allohexaploid genome of bread wheat (16000 Mb) is approximately 3, 6, and 35 times larger than the barley (5000 Mb), maize (2500 Mb), and rice (440 Mb) genomes, respectively (Arumuganathan *et al.*, 1991; Shields, 1993). Comparative mapping studies have shown that despite substantial variation in genome size and chromosome number, grass species have maintained significant conservation of marker order (colinearity) and have sustained a minimal number of large chromosomal rearrangements since their divergence 50-80 million years ago (Wolfe *et al.*, 1989; Crepet *et al.*, 1991; Ahn *et al.*, 1993; Moore *et al.*, 1995; Clark *et al.*, 1995; Devos *et al.*, 1997; Gale *et al.*, 1998; Keller *et al.*, 2000). The high degree of observed colinearity, coupled with the basic assumption that the essential components for accurate growth and development are relatively conserved within plant systems, established the use of model organisms with small genome sizes, specifically *Arabidopsis thaliana* and rice, as an integral tool for comparative genomics studies.

Another significant advance in the study of comparative genomics was the development of low replication bacterial plasmids capable of retaining inserts of foreign DNA greater than 250 kb in length, Bacterial Artificial Chromosomes (BACs). This technology had several advantages over its yeast counterpart (YACs) in that it improved transformation efficiency, decreased the number of chimeric clones, increased the ease and speed of isolation of the insert DNA, and provided greater stability during liquid handling

(Shizuya *et al.*, 1992a; Woo *et al.*, 1994a; Zhang *et al.*, 1996). The utilization of BACs, as well as PACs (P1 Artificial Chromosomes), enabled the full length genomic sequencing of several organisms including humans, *Arabidopsis*, and rice (Lin *et al.*, 1999; Mayer *et al.*, 1999; Theologis *et al.*, 2000; Salanoubat *et al.*, 2000; Tabata *et al.*, 2000; Yu *et al.*, 2002; Goff *et al.*, 2002; Sasaki *et al.*, 2002; Hillier *et al.*, 2003; Wu *et al.*, 2004) and the generation of physical maps at a significantly higher resolution than other physical mapping strategies, such as chromosome deletion, radiation hybrid, use of gametocidal chromosomes, and HAPPY mapping (Werner *et al.*, 1992; Shizuya *et al.*, 1992a; Dear *et al.*, 1993; Waugh *et al.*, 2002; Wardrop *et al.*, 2002). Furthermore, after the genomic region harboring genes controlling a particular trait of interest is identified on a genetic map, flanking markers can be used to establish a local contig of BACs spanning the locus. This approach, together with comparative analysis of colinear regions in model organisms, can aid in the map-based cloning of genes otherwise difficult to obtain in species of large genome size. For example, the colinear genomic regions of both rice and sorghum played an integral role in the positional cloning of the *Triticum monococcum* vernalization gene *VRN1* (Yan *et al.*, 2003).

Despite the apparent conservation of gene order and content on a full genome scale, at the local level various small chromosomal rearrangements, such as segmental inversions, translocations, insertions, and deletions, have been reported to disrupt the degree of microcolinearity (reviewed in Bennetzen, 2000a; Feuillet *et al.*, 2002; Bennetzen *et al.*, 2002). One mechanism able to explain numerous small chromosomal arrangements across an entire genome is the occurrence of an ancient polyploidization event and the subsequent loss of individual duplicated segments. Evidence of such events has been documented in several plant species, including maize, *Brassica*, cotton, and soybean (Helentjaris *et al.*, 1988; Lagercrantz, 1998; Wendel, 2000).

Eleven large contiguous barley genomic sequences have been reported in the literature to date comprising 1.35 megabases of sequence (Panstruga *et al.*, 1998; Shirasu *et al.*, 2000; Dubcovsky *et al.*, 2001; Rostoks *et al.*, 2002; Wei *et al.*, 2002; Yan *et al.*, 2002; Gu *et al.*, 2003). Coupled with the reports of large contiguous sequence from wheat and maize, these studies have allowed invaluable insight into the genome organization of large genome crop species (SanMiguel *et al.*, 1996; Wicker *et al.*, 2001). Although several studies have described the levels of microcolinearity between Triticeae species and rice (Kilian *et al.*, 1997; Han *et al.*, 1998; Han *et al.*, 1999; Druka *et al.*, 2000; Li *et al.*, 2002), only two previously reported studies have compared large orthologous regions from rice and barley at the sequence level (Dubcovsky *et al.*, 2001; Brunner *et al.*, 2003). Such comparisons are vital for determining the extent to which comparative genomics approaches with the fully sequenced rice genome will be applicable for association mapping and positional cloning strategies.

This chapter presents the use of BACs for the construction of a physical map of the region harboring the *Ha* locus in barley. Full characterization allowed a detailed representation of the patterns of genome organization within the region. Comparison with the colinear region in rice suggests that comparative genomics can be an invaluable resource for the identification of gene location and determination of gene structure. However, the large number of small chromosomal rearrangements could cause serious complications in the application of comparative genomics in association mapping and positional cloning.

3.2 Results

3.2.1 Generation of a Physical Map of the Region Harboring the Hardness (Ha) Locus in Barley

A set of 14 BACs (*H. vulgare* cv. Morex ; Yu *et al.*, 2000) identified through positive hybridization with a wheat *GSP-1* cDNA clone was obtained from Professor Andris Kleinhof's laboratory at Washington State University (http://barleygenomics.wsu.edu/

db3/db3.html). These BACs were fingerprinted in Professor Michele Morgante's lab at DuPont Agriculture and Nutrition, and BAC122.a5 was selected for sub-clone library construction and full length sequencing (see Sections 2.5-2.7). In order to extend the physical region to include the genetically linked hordoindolines genes (Rouves *et al.*, 1996; Beecher *et al.*, 2001), orthologous wheat sequences (puroindolines; Genbank accessions AJ249929 and AJ249928) were used as templates for the generation of gene specific probes for additional BAC library screens (see Section 2.2). Size determination and BAC end sequencing were employed to identify BACs that would allow minimal overlap and ensure maximum coverage (see Sections 2.3 and 2.4). BACs 519.k7 and 799.c8 were selected for further experimentation. Sequences from the subclones libraries of the three contiguous BACs (6912 clones with an average length 600 quality basepairs) were assembled and characterized to generate a complete physical map of the region containing the hardness (*Ha*) locus at approximately 14-fold coverage (see Section 2.8 and 2.9). Two problematic regions prevented the realization of one complete continual sequence. The first difficult region was composed of a ~340 bp AT-rich tandemly repeated segment located within the truncated Caspar_AY643842_1 transposon at the extreme 5' region of the contig (BAC517.k9; Figure 3.1A). PCR amplification confirmed the sub-contig assembly and the estimated gap length indicated that 3 to 4 copies of the tandem duplication are missing from the sequence. The second problematic region also involves an AT-rich tandem duplication (42 bp) located ~3 kb downstream of the chalcone synthase (*HvCHS*) gene (BAC799.c8; Figure 3.1A). *GSP*, *hina*, and *hinb-1* were mapped to the distal end of the short arm of chromosome 5H (see Section 2.15) confirming the previously reported map location (Rouves *et al.*, 1996; Beecher *et al.*, 2001).

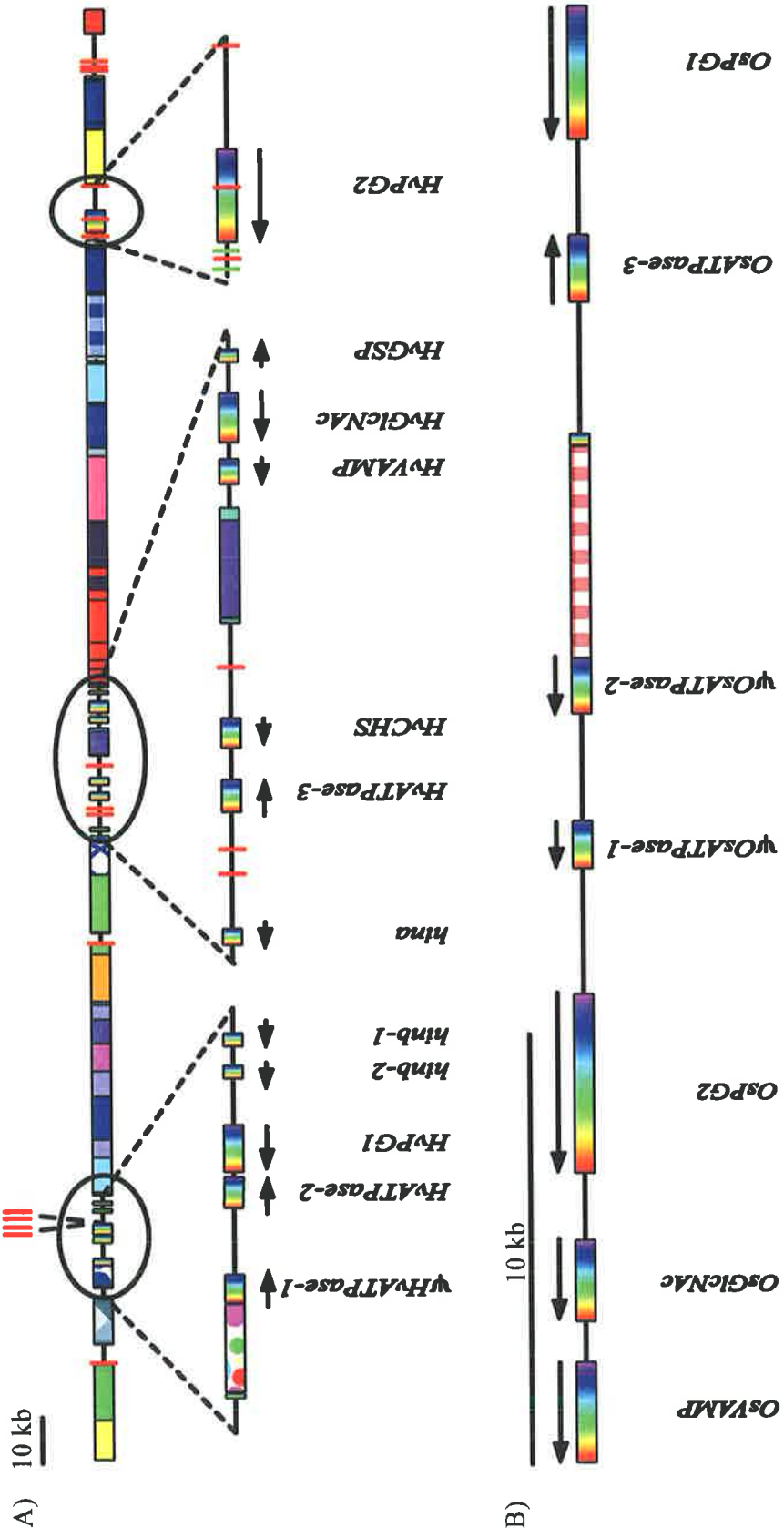


Figure 3.1. A linear representation of the gene content and organization of the A) region containing the barley *Ha* locus and its B) collinear rice region. Coding sequence is represented by rainbox boxes and arrows designate gene orientation. $tRNA^{ARG}$ are represented by green vertical lines. Repetitive sequence is coded similar to the elements in Figure 3.2.

3.2.2 *Characterization and Organization of the Barley Genomic Region*

3.2.2.1 *Gene Density*

The gene density of the region was determined through the integration of several different gene prediction applications and homology to characterized genes and expressed sequence tags (ESTs) in the public databases (see Section 2.9). In total, twelve putative protein-coding and two duplicated $tRNA^{ARG}$ genes (Figure 3.1A) were identified within the 300 kb sequenced region. All exon:intron splice junctions contained the conserved GT and AG intron borders and a minimum of 5 of the 9 (5' CAG:GTAAGT 3') and 3 of the 5 (5' GCAG:G 3') consensus nucleotides for the respective exon:intron and intron:exon splice sites in plants with one exception. The border between exon 1 and intron 1 of the putative synaptobrevin (*HvVAMP*) only contained 4 of the 9 exon:intron consensus nucleotides. However, both the presence of the mandatory GT intron boarder and splice agreement with more than one EST support this as a functional splice site.

Three of the four candidate grain texture genes, namely *hinb-1*, *hinb-2*, and *hina*, were found in the same orientation. However, *HvGSP* was in the opposite orientation (Figure 3.1A). Sequence homology at the protein level (>46%) suggests that all four are members of the same gene family and, therefore, may be the result of duplications of a single ancestral gene. Based on nucleotide sequence homology, the original duplication resulted in *HvGSP* and one of the hordoindoline genes. Subsequent duplications generated templates for the gradual divergence of *hina* and *hinb* and an additional *hinb* copy.

Three of the putative genes belong to the AAA+ (ATPase Associated Activities) superfamily characterized by one or two conserved domains (AAA modules) responsible for ATP binding (Patel *et al.*, 1998). This family of genes is ubiquitous to all kingdoms of life and is involved in numerous cellular activities including membrane fusion, proteolysis, and DNA replication (Ogura *et al.*, 2001a). All three ATPases are located

within the same 37 kb gene cluster (Figure 3.1A). *HvATPase-2* and *HvATPase-3* code for 518 and 516 aa proteins which are 84% and 80% identical at the nucleotide and protein level, respectively. The *ψHvATPase-1* pseudogene has maintained 84% and 91% nucleotide homology to *HvATPase-2* and *HvATPase-3*, respectively, despite the insertion of the HORPIA-2_AY643843 retrotransposon and several insertion and deletion events causing shifts in the reading frame (Figure 3.2). Remnants of an additional ATPase gene (*ψHvATPase-4*) were detected immediately downstream to *HvATPase-3* demonstrating 81% homology to *HvATPase-3*. This copy has been severely truncated by a deletion of over 1 kb from the internal portion of the coding sequence. Evidence of yet another degenerate ATPase (*ψHvATPase-5*) gene exists in the region flanked by *ψHvATPase-1* and *HvATPase-2*. A stretch of ~500 bp demonstrates 88% homology to the immediate 5' flanking sequence of *HvATPase-2*. This precedes a shorter segment with 88% homology to the later portion of the gene. Although the full length ATPase genes have maintained considerable identity across the entire coding region, very little homology was detected among the flanking sequence to suggest the duplication history of this gene family cluster. Based on coding sequence homology alone, the original duplication probably resulted in *HvATPase-2* and one of the other two full-length copies with a second duplication generating the third copy. Additional duplications of both *HvATPase-2* and *HvATPase-3* resulted in *ψHvATPase-4* and *ψHvATPase-5*, respectively. Genomic sequences of other barley lines or close barley relatives are needed to discern the exact series of events.

Three out of the five remaining genes showed significant homology to previously described proteins: naringenin-chalcone synthase (*HvCHS*), N-acetylglucosaminyltransferase (*HvGlcNAc*), and synaptobrevin (*HvVAMP*), a vesicle associated membrane protein. *CHS* is a member of the chalcone synthase gene family. Chalcone is an integral starting component for phenylpropanoid biosynthetic pathways involved in various cellular functions including flower pigmentation (anthocyanin) and

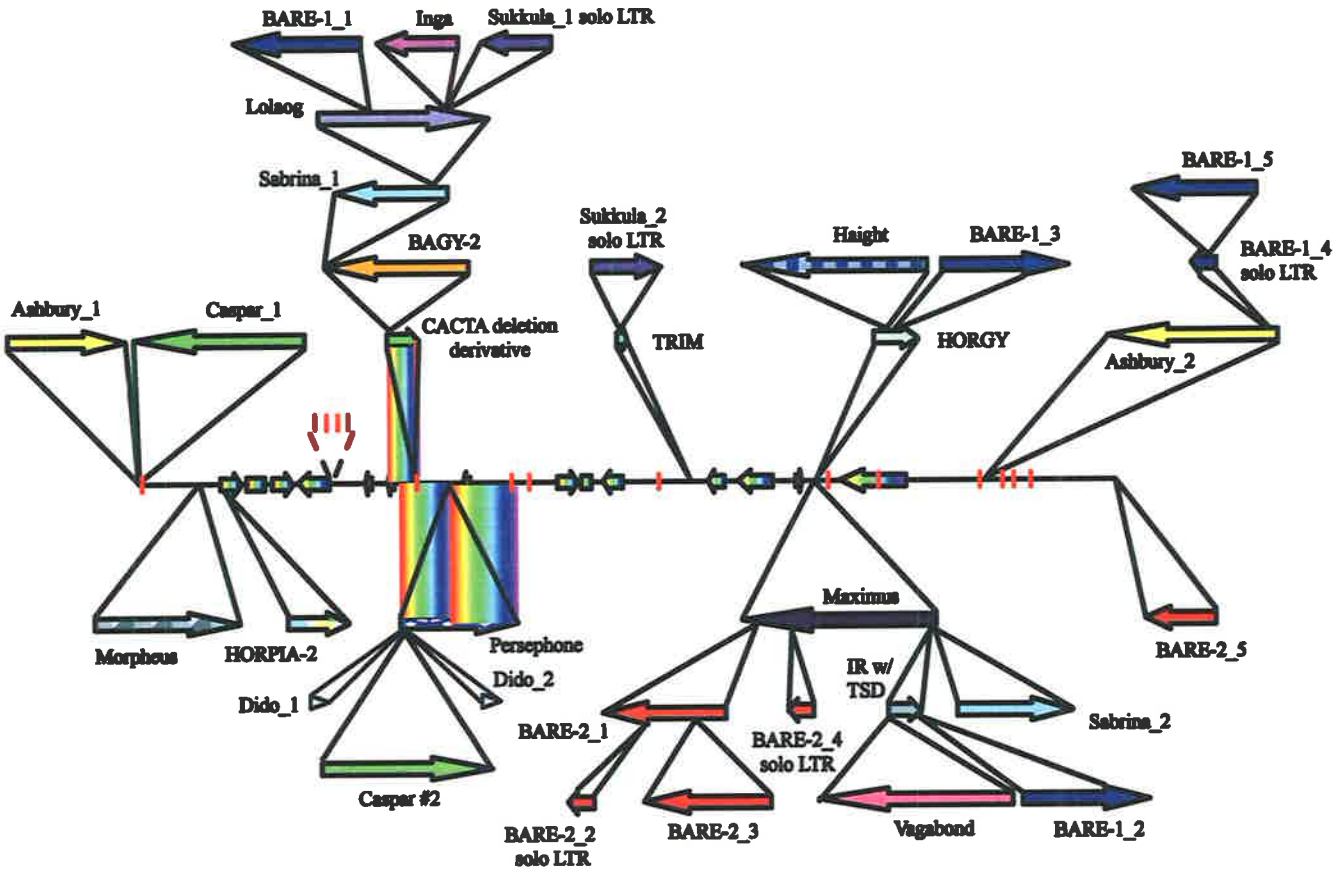


Figure 3.2. Stacked representation of the genome organization of the region containing the *Ha* locus in barley. Arrows directly on the "base" sequence represent putative genes; designation can be seen in Figure 3.1. Arrows above and below the "base" sequence represent the position, orientation, and order of insertion of various transposable elements. Vertical red bars illustrate MITES.

microbial defense (phytoalexins; Dixon *et al.*, 1995a; Dixon *et al.*, 1995b; Dixon *et al.*, 1996; Shirley, 1996). Synaptobrevin is involved in a complex of SNARE (soluble-*N*-ethylmaleimide-sensitive factor attachment protein receptors) proteins that control the regulation of vesicle docking and fusion during transport (Trimble *et al.*, 1988; Baurnert *et al.*, 1989; Sollner *et al.*, 1993; Weber *et al.*, 1998; Chen *et al.*, 2001). *GlcNAc* is a member of the large enzymatic superfamily of uridine diphosphate (UDP) glycosyltransferases (UGTs). UGTs regulate the transfer of sugar molecules (glycosyl residues) between different chemical R-groups (aglycones); thus, indirectly regulating the biochemical properties of aglycones. Aglycones are secondary metabolites involved in abiotic stress and defense responses, hormones, and foreign chemical substances (xenobiotics such as pesticides and herbicides; Li *et al.*, 2001; Ross *et al.*, 2001). The two additional genes are putative genes whose functions have yet to be determined. These will be referred to as putative gene 1 (*HvPG1*) and putative gene two (*HvPG2*) throughout the remainder of this thesis. Although EST homology is low (pLog > E-6) for both genes and limited to members of the grass family, *HvPG2* demonstrates significant protein homology (pLog ≥ E-44) to several predicted proteins from mammalian species, including *Rattus norvegicus*, *Homo sapiens*, and *Mus musculus* (GI accession numbers 34867764, 13376072, and 21313472 respectively).

3.2.2.2 Copia-like retrotransposons

Over 75% of the contiguous barley sequence was composed of repetitive elements. The overwhelming majority of retrotransposon insertions were part of the BARE family (Figure 3.2 & Table 3.1; Manninen *et al.*, 1993). In total, four full length and one solo LTR and two full length and two solo LTRs were identified of BARE-1 and BARE-2 respectively. Only partial sequence of an additional BARE-2 element (BARE-2_AY643844_5) was obtained as a result of its location at the exact 3' end of the contig. Although premature stop-codons and frameshift mutations indicate most to be non-functional copies, all full length BARE elements demonstrated greater than 97% sequence

homology across the ~1800 bp LTRs and BARE-1_AY643844_3 maintains an intact open-reading frame suggesting that this family of elements may have been active in the recent barley genome. This element could potentially be used as a marker in transposition studies to determine conclusively if BARE-1 is currently active.

Table 3.1. Summary of the transposable elements found within the 300 kb barley sequence.

Name	Element Type	Element Subgroup	Size (bp)	TSD	Reference Sequence
Ashbury_AY643842_1	LTR retrotransposon	Ty3/gypsy	8278	N/A	novel
Ashbury_AY643844_2	LTR retrotransposon	Ty3/gypsy	12131	GTGAG	novel
BAGY-2_AY643843	LTR retrotransposon	Ty3/gypsy	10260	CTAAA	TREP206; AF254799
BARE-1_AY643843_1	LTR retrotransposon	Ty1/copia	8917	GTTGA	TREP725; AF227791
BARE-1_AY643844_2	LTR retrotransposon	Ty1/copia	8932	GCGTG	TREP725; AF227791
BARE-1_AY643844_3	LTR retrotransposon	Ty1/copia	8957	CATGT	TREP725; AF227791
BARE-1_AY643844_5	LTR retrotransposon	Ty1/copia	8503	CAAGA	TREP725; AF227791
BARE-1_AY643844_4 solo LTR	LTR retrotransposon	Ty1/copia	1818	GGAAG	TREP725; AF227791
BARE-2_AY643844_1	LTR retrotransposon	Ty1/copia	9203	ACACC	AJ279072
BARE-2_AY643844_2	LTR retrotransposon	Ty1/copia	8619	GTGAC/G	AJ279072
BARE-2_AY643844_5	LTR retrotransposon	Ty1/copia	5021	N/A	AJ279072
BARE-2_AY643844_3 solo LTR	LTR retrotransposon	Ty1/copia	1807	GTTAC	AJ279072
BARE-2_AY643844_4 solo LTR	LTR retrotransposon	Ty1/copia	1813	AT/GGCT	AJ279072
CACTA_AY643843	transposon	CACTA	2140	TAT	novel
Caspar_AY643842_1	transposon	CACTA	7646	N/A	TREP788
Caspar_AY643844_2	transposon	CACTA	12085	TTA	TREP788
Dido_AY643843_1	non-LTR retrotransposon	SINE	256	N/A	novel
Dido_AY643843_2	non-LTR retrotransposon	SINE	256	N/A	novel
Haight_AY643844	LTR retrotransposon	Ty3/gypsy	13050	CCCGC	novel
HORGY_AY643844	LTR retrotransposon	Ty3/gypsy	3077	TCCTC	TREP728; AF427791
HORPIA-2_AY643843	LTR retrotransposon	Ty1/copia	4285	CGCGC	TREP730; AF427791
Inga_AY643843	LTR retrotransposon	Ty1/copia	5650	N/A	TREP704; AF474982
IR with TSD	unclassified	N/A	2244	ATAGG	novel
Lolaog_AY643843	LTR retrotransposon	Ty3/gypsy	10698	GCATA	AY268139
Maximus_AY643844	LTR retrotransposon	Ty1/copia	13775	CCAAC	novel
Morpheus_AY643843	non-LTR retrotransposon	LINE	7966	ATGCCG	novel
Persphone_AY643843	non-LTR retrotransposon	LINE	7889	ATGTCTGCCCAACGG	novel
Sabrina_AY643843_1	LTR retrotransposon	Ty3/gypsy	8000	GTCAT	TREP710; AF474071
Sabrina_AY643844_2	LTR retrotransposon	Ty3/gypsy	8183	GAC/ACC	TREP710; AF474071
Sukkula_AY643843_1 solo LTR	LTR retrotransposon	Ty3/gypsy	4961	CAAGC/CG	TREP715; AF474072
Sukkula_AY643843_2 solo LTR	LTR retrotransposon	Ty3/gypsy	4844	ACTGG	TREP715; AF474072
TRIM_AY643843	non-LTR retrotransposon	TRIM	725	GCCGG	AY164585
Vagabond_AY643844	LTR retrotransposon	Ty3/gypsy	13918	GGTCAA	novel*

* similarity to TREP253; AF459639 was to internal region only

The novel *copia*-like element, Maximus_AY643844, was classified by high identity to the entire length of the Inga polyprotein (PTREP3; Table 3.1). The full length element is 13,775 bp in length as defined by perfect 5 bp TSD flanking 97% identical ~1400 bp LTRs. This element has been interrupted by eight separate repetitive element insertion events, four direct and four indirect, indicating one of the oldest insertion events in the region (Figure 3.2). An interesting feature of Maximus_AY643844 is the potential coding capacity in the internal region downstream of the polyprotein that demonstrates weak homology to a cell surface glycoprotein-1 precursor (Swissprot accession number Q06852). This is the same region where the envelope protein is usually found in retroviruses. Maximus_AY643844 has inserted less than 1 kb downstream of the *HvGSP* stop codon. An additional copy of this element, identified by homology to the first 149 bp of the 5' LTR, has inserted 566 bp downstream of *Hv1PST* (GenBank accession number AF004950) suggesting the tendency to insert within the genic space of the barley genome.

Although no insertion events have occurred in the remaining two *copia*-like elements, both have undergone different mechanisms of change since the time of insertion. A small deletion event has removed ~350 aa from the 3' portion of the polyprotein of HORPIA-2_AY643843 and the entire 3' half of Inga_AY643843 has been truncated leaving only the 5' LTR through to the first 473 aa of the polyprotein.

3.2.2.3 Gypsy-like retrotransposons

Several previously characterized *gypsy*-like elements were found scattered throughout the contig. Four of these were found in a nested retrotransposon cluster in which the Sukkula_AY643843_1 solo LTR, along with BARE-1_AY643843_2 and the truncated Inga_AY643843 mentioned above, were the last of a sequential series of insertions involving the now degenerate CACTA_AY643843 transposon and the BAGY-

2_AY643843, Sabrina_AY643843_1, and Lolaog_AY643843 retrotransposons (Figure 3.2). The Lolaog_AY643843 element has a 2,623 bp insertion located downstream of the polyprotein in the internal region just after the Sukkula_AY643843_1 insertion. A 906 bp segment of this insertion contains 27% identity to a putative *gypsy*-like ORF of RIRE2 (GenBank accession number BAB61182) in the opposite orientation. Although a perfect 5 bp putative TSD flanks the inserted element, no remnants of direct or indirect repeats are present on either side of the coding region (Table 3.1). A second copy of both Sabrina (Sabrina_AY643844_2) and a Sukkula solo LTR (Sukkula_AY643843_2) and the degenerate HORGY_AY643844 element no longer containing any significant homology at the amino acid level were also located within the contig.

Three novel *gypsy*-like elements were detected. Ashbury_AY643844_2 is 12,131 bp in length flanked by perfect 5 bp TSD and 98% identical 558 bp LTRs (Table 3.1). A region of ~5 kb located 1 kb after the 5' LTR including the entire polyprotein showed 73% homology to the Jeli retrotransposon (GenBank accession number AF459088) at the nucleotide level. Although Jeli is also ~12 kb with ~550 bp LTRs, no significant homology was found between the two elements for the LTRs or the entire 6 kb region at the 3'end. Remnants of a *gypsy*-like polyprotein in the opposite orientation were found in the non-coding internal space adjacent to the 3' LTR. This highly degenerate ancient retroelement insertion was probably originally incorporated into an active copy of Ashbury as it is also present in both the 8,278 bp partial copy located at the extreme 5' end of the contig and in the full length copy present in the downstream region of *HvNAS1* gene (GenBank accession number AB023436.1). In addition, potential coding capacity, 54% identical to the 40s ribosomal protein *cyc07* (GenBank accession number AAP80855.1), is located between the polyproteins of Ashbury_AY643844_2 and the inserted element.

Vagabond_AY643844 demonstrates 80% homology at the nucleotide level across the entire coding portion of the internal segment of the Latidu retrotransposon (GenBank accession number AF459639). The ~450 bp LTRs of the later element also show 79% homology to the extreme 3' end of the 4 kb Vagabond_AY643844 LTRs. In addition, two 100 to 200 bp segments of the internal region directly upstream of the Latidu 3' LTR show homology to the terminal region of the Vagabond_AY643844 LTR (bp 2812 to 3251) suggesting these elements have arisen from the same ancestral element. The 2431 to 4017 bp and the 1569 to 2619 bp regions of the Vagabond_AY643844 LTR show 96% and 91% homology to a truncated Latidu element (GenBank accession number AF521177) and the unannotated region between two Sabrina elements (GenBank accession number AF474072), respectively. This particular segment of the truncated AF521177 element, however, demonstrated no significant homology to the full length AF459639 element.

The last novel *gypsy*-like element, Haight_AY643844, shows no significant homology to any previously characterized elements at the nucleotide level. However, a 5 kb internal region demonstrates 53% identity to the predicted Jeli polyprotein (GenBank accession number AF459088-2). Haight_AY643844 is 13,050 bp in length flanked by perfect 5 bp TSD and 509 bp LTRs sharing 97% homology (Table 3.1). Similar to Ashbury_AY643844_2, remnants of a *gypsy*-like polyprotein in the opposite orientation were found in the non-coding internal space adjacent to the 3' LTR.

3.2.2.4 non-LTR retrotransposons

Two novel Long Interspersed Nuclear Elements (LINEs) were identified. Persephone_AY643843 was characterized by two putative ORFs containing 32% and 69% identity at the protein level to the 605 and 1,381 aa polyproteins from a previously identified LINE, Karin (GenBank accession number AF459088). This element is 7,889 bp in length, contains an indicative poly-A tail, and is flanked by perfect 15 bp TSD

(Table 3.1). Morpheus_AY643843 could be identified by 35% identity across 1146 aa of the Karin Pol (polymerase) polyprotein domain (GenBank accession number AY146587). A second putative ORF located in the Gag (group-specific antigen) polyprotein region demonstrates 32% identity at the protein level to four different hypothetical rice proteins (GenBank accession numbers AAM08894.1, AAP53401.1, AAM08645.1, and AAN31795.1). This 7,966 bp element is flanked by perfect 6 bp TSD and also contains an indicative poly-A tail (Table 3.1).

Two copies of a putative Short Interspersed Nuclear Element (SINE), Dido_AY643843_1 & 2, were found inserted in tandem just inside the 5' TSD of the Persephone_AY643843 LINE element. Dido includes both the A box (TRKYNNARNGG) and B box (RGTTCRANHYY) consensus sequences for *Arabidopsis* SINE elements (Myouga *et al.*, 2001). Neither copy has retained obvious TSD. The presence of additional copies maintaining the TSD is needed to confirm these as SINEs.

Despite expectations that the recently discovered Terminal-repeat Retrotransposons in Miniature (TRIM) are ubiquitous and frequent in plant genomes (Witte *et al.*, 2001), only one such element was present in this contig (TRIM_AY643843). The 725 bp element was classified by 264 bp terminal direct repeats (TDRs) of 92% homology and a primer binding site just downstream of the 5' TDR which is complementary to the methionine tRNA (Table 3.1).

3.2.2.5 Class II Transposable Elements

The only class II transposon elements located within the region were two Caspar insertions, one full length (Caspar_AY643843_2) and the other truncated at the 3' end (Caspar_AY643842_1). This truncation event is most likely caused by the insertion of the partial Ashbury_AY643842_1 element with the remaining portions of both elements just 5' of the end of the contig (Figure 3.2). However, without the ability to find precise

TSD, it is impossible to say for certain whether this is the case. One non-autonomous deletion derivative of a CACTA element was also present in this region. The 8 bp CACTAGTG MIR and 3 bp TSD designated it as a member of the recently discovered Caspar family (Wicker *et al.*, 2003a).

3.2.2.6 Mini Inverted Transposable Elements (MITEs)

In total 15 different MITE insertions were found composing less than 1% of the total genomic region. The majority of these were members of the Stowaway and Tourist families contributing seven and four respective copies. One full-length and one partial copy of the XI element were located in this region. This element, previously described as a potential novel element (Brunner *et al.*, 2003), demonstrates high homology to intron five of an *Aegilops tauschii* isoamylase gene (GenBank accession number AF548379). The isoamylase copy maintains 36 out of 41 bp imperfect MIRs suggesting that this element originated in the Triticeae as a MITE. However, only two of the six copies located within the *Hordeum vulgare* contig 211252 (Genbank accession number AF521177) remain as intact full-length copies and both sets of MIRs have degenerated to less than 75% identity suggesting that additional mechanisms such as non-reciprocal recombination could account for the high accumulation of this element in this region. This is further supported by lack of intact TSDs and the tandem nature of several copies. The presence of all known copies near or within $(TA)_n$ microsatellites suggests a strong insertion bias.

3.2.3 Characterization of the Colinear Region in Rice

To facilitate a comparative study between rice and barley, all repetitive elements were removed from the barley genomic sequence and flanking segments were merged at the site of target duplication. The resulting 69 kb barley sequence was used as a template for additional searches of the "nr" and "dbest" databases at the National Center of Biotechnology Information (NCBI). Several regions of considerable homology were

identified across a 34 kb unannotated segment of rice chromosome 12 (Genbank accession numbers AL928743 and AL732378). All seven conserved regions corresponded to the genic space of the barley contig and no significant sequence similarity longer than 25 bp was found beyond the coding regions of the genes. This was confirmed by comparison of the complete barley sequence to AL928743 using the Dotter computer program (Sonnhammer *et al.*, 1995).

Similar to the barley region, rice also contains three ATPase gene copies (Figure 3.1B). However, a greater degree of sequence homology exists among paralogs within species than between homologs of the different species. This indicates that gene duplication occurred independently post speciation (Figure 3.3). *OsATPase-3* is the only functional rice copy encoding a 524 aa protein with 68% and 72% identity to *ψHvATPase-1* and *HvATPase-3*. *OsATPase-3* maintains a minimum of 81% nucleotide homology to both rice paralogs and was probably a product of the original duplication event. *ψOsATPase-1* contains a premature stop codon resulting in the truncation of the C-terminal end of the protein. Although *ψOsATPase-1* maintains 94% homology to the first two-thirds of *ψOsATPase-2*, no significant homology is observed after the truncation suggesting that either *ψOsATPase-1* resulted from a partial gene duplication event or the terminal end has been subsequently deleted. *ψOsATPase-2* has been interrupted by the insertion of a 5 kb Ty1/*copia* element between codons 57 and 58.

A TBLASTN comparison using the GSP protein identified a small stretch of 120 bp in the colinear rice sequence with high similarity (64%, E=0.55) to the C-terminal end of the protein. This putative unannotated rice protein was previously identified through a similar comparison using the monococcum GSP gene and further analysis revealed the presence of both a putative TATA-box and polyadenylation signal (Chantret *et al.*, 2004). To determine the most closely related sequence to the barley grain texture genes in the rice genome, BLASTP and TBLASTN comparisons to the annotated rice proteins and the

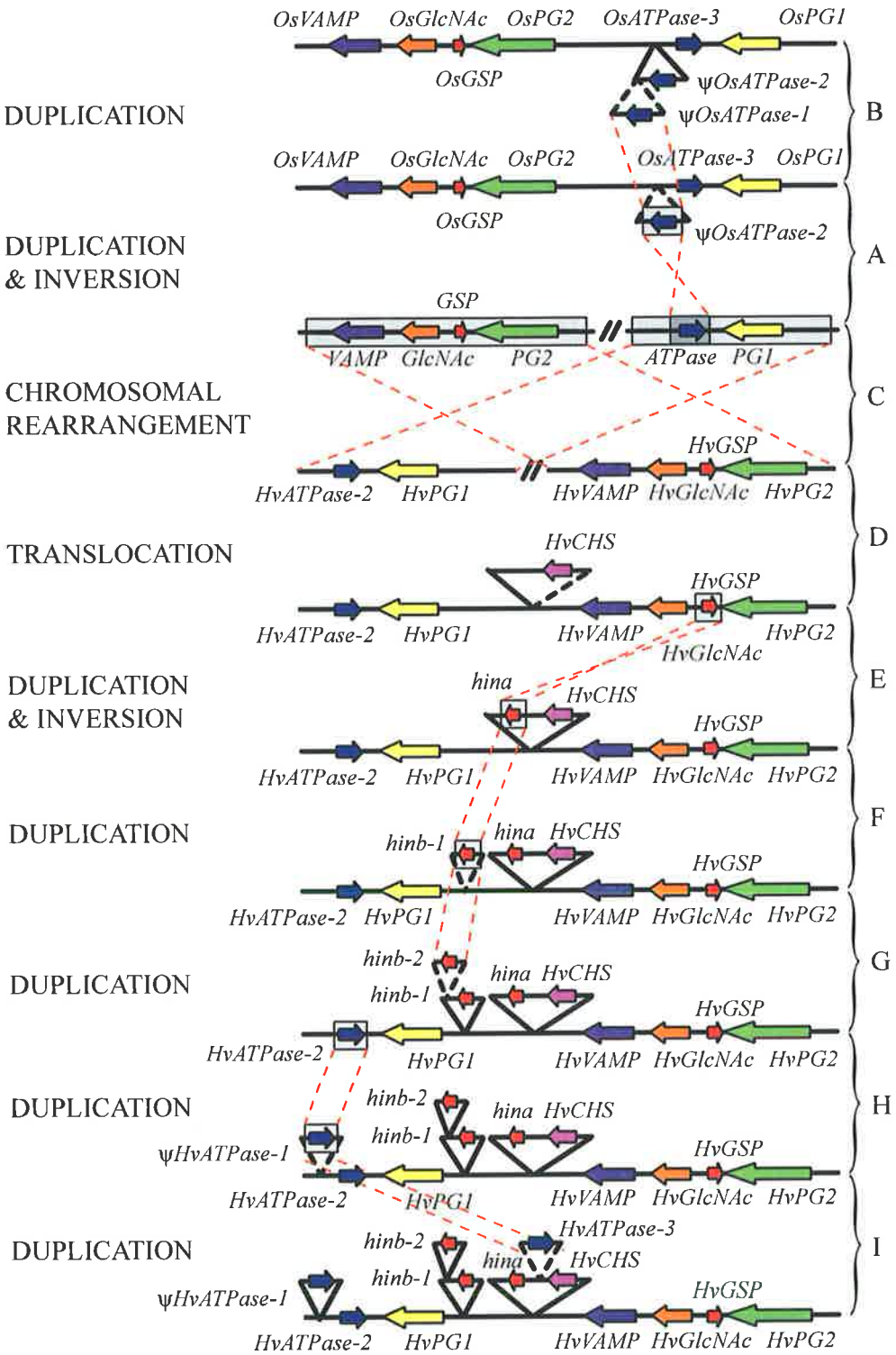


Figure 3.3. A visual representation of one possible evolutionary scheme between the rice and barley collinear sequences. Evolutionary events move upwards towards present day rice (A-B) and downwards towards present day barley (C-I) from the presumed last common ancestor. C) An intra-chromosomal rearrangement results in the repositioning of two conserved gene clusters. D) Translocation involves the relocation of *CHS* from a separate chromosomal location. E-G) Subsequent duplications and a gene inversion generate the individual grain texture genes. A-B & H-I) Independent gene duplications and inversions generate numerous copies of ATPase in both species. The two severely degenerate copies of barley ATPase are not present in this scheme.

rice genomic sequence, respectively, were preformed. The highest protein similarity found was to a family of rice prolamin genes (54 to 51% similarities, 2E-14 to 0.0025). This similarity is not surprising as puroindolines have previously been classified in the prolamin superfamily, albeit a different class than the prolamins themselves, characterized by the conserved number and spacing of cysteine residues (Shewry *et al.*, 2002). Although a higher E value was obtained in comparison with the prolamin genes than the unannotated protein described above, several lines of evidence exist that suggest prolamins are not orthologous to the grain texture gene ancestor. Similarity to GSP did not extend across the entire protein and was predominantly restricted to the conserved cysteine backbone. Furthermore, prolamins show a higher similarity to other barley ESTs (59% similarity, 2e-26) that extends beyond the cysteine and glutamine residues.

Homologs of four of the five remaining barley genes were located within the colinear region of rice. However, the orientation and organization of these genes is not entirely conserved between the two grass species (Figure 3.1). A chromosomal rearrangement has reversed the positions of two gene clusters (GC1: ATPase and *PG1* and GC2: VAMP, *GlcNAc*, *GSP*, and *PG2*) while maintaining gene order and orientation within clusters. Although a *CHS* homolog is not present within the colinear rice region, a homolog with 91% identity at the nucleotide level exists on rice chromosome 7 (GI number 34395291). This is suggestive of a past translocation event involving either the relocation of *CHS* from chromosome 12 to chromosome 7 in rice or of *CHS* from another region of the barley genome to the region surrounding the *Ha* locus (Figure 3.3).

3.2.4 Determination of Gene Structure

To help confirm the gene structure of the barley and rice orthologs, a BLASTP search was done at The *Arabidopsis* Information Resource (TAIR, http://www.arabidopsis.org/Blast/) to ascertain the closest *Arabidopsis* homologs for comparison (Table 3.2). Grain texture homologs could not be detected through BLASTN

Table 3.2. BLASTP comparisons between the predicted barley protein (Hv), the predicted collinear rice protein (Os) or closest homolog, and the closest *Arabidopsis* homolog. BLASTN comparisons between the predicted barley gene and the dbEST database. No significant homologs were found to the grain texture genes in either rice or *Arabidopsis*. N/A – not applicable.

Hv Gene	Size (aa)	Predicted Os		Arabidopsis Gene	BLASTP		EST Accession	BLASTN	
		Score	Expect		Score	Expect		Score	Expect
HvATPase-2	518	N/A	N/A	At5g40010	520	e-147	CD939530, *T. aestivum*	634	0
							BJ257579, *T. aestivum*	698	0
							BJ265958, *T. aestivum*	323	e-85
HvPG1	535	830	0	At1g74780	536	e-152	CA731405, *T. aestivum*	959	0
							CA007346, *H. vulgare*	1235	0
							BU996747, *H. vulgare*	1132	0
							CA005797, *H. vulgare*	825	0
hinb-2	147	N/A	N/A	N/A	N/A	N/A	BE454227, *H. vulgare*	874	0
hinb-1	147	N/A	N/A	N/A	N/A	N/A	BG36753, *H. vulgare*	874	0
hina	149	N/A	N/A	N/A	N/A	N/A	BQ65384, *H. vulgare*	886	0
HvATPase-3	516			At5g40010	514	e-156	BI778940, *H. vulgare*	971	0
							CA684810, *T. aestivum*	753	0
HvCHS	432	691	0	At4g34850	527	e-150	BG343835, *H. vulgare*	1055	0
							CA600207, *T. aestivum*	825	0
							CA502438, *T. aestivum*	323	e-85
HvSNARE	215	362	e-99	At1g04760	309	e-83	CB667109, *O. sativa*	224	e-55
							CA667948, *T. aestivum*	490	e-135
HvGluNAc	425	709	0	At5g39990	559	e-159	BM368259, *H. vulgare*	618	e-174
							BG948458, *S. bicolor*	507	e-140
							CB861600, *H. vulgare*	1152	0
							BU983520, *H. vulgare*	841	0
HvGSP	164	37.4	0.55	N/A	N/A	N/A	BE454072, *H. vulgare*	975	0
HvPG2	723	1076	0	At1g74790	805	0	BU997791, *H. vulgare*	952	0
							BG369772, *H. vulgare*	922	0
							BJ278101, *T. aestivum*	670	0
							CB631610, *O. sativa*	113	e-21
							BU100503, *T. aestivum*	963	0

searches against the entire *Arabidopsis* genome. The highest protein similarities (BLASTP) to any of the grain texture genes was 48% to the C-terminal end of a seed storage protein (At4g27140, score 39.5, E=0.002) and 39% to the N-terminal end of a protease inhibitor/seed storage/lipid transfer protein (At3g42720, score 32.1, E=0.26).

The putative barley ATPases were the only genes within the contig to maintain a higher similarity to the *Arabidopsis* homolog (At5g40010, 75% similarity, Table 3.2) than the closest rice homolog (72%). The ATPases in all three species contained a single exon.

HvPG1 (535 aa) shows 87% and 67% similarity to *OsPG1* (526 aa) and the closest *Arabidopsis* homolog (At1g74780, 533 aa; Table 3.2). All three genes contain 2 exons. However, neither exon is of similar length in any of the three species (Figure 3.4A). The gene structure in barley and rice was confirmed by alignment of the genomic sequence with Triticeae and rice ESTs, respectively (Table 3.2). Although the precise function of this gene has yet to be determined in any of the three species, the *Arabidopsis* homolog is annotated as showing similarity to a nodule-specific protein in *Lotus japonicus* (GI number 3329366).

HvCHS (432 aa) showed a high level of similarity to its closest rice (GI number 34395291, 405 aa, 87% similarity) and *Arabidopsis* (At4g3450, 392 aa, 78% similarity) homologs (Table 3.2). The gene structure in barley was confirmed by alignment of the genomic sequence with wheat and barley ESTs (Table 3.2). Both the rice and barley genes contain 2 exons and the *Arabidopsis* gene contains 3 exons (Figure 3.4B). Exon 1 from all three species differs in length by only 9 codons. The main difference between exon 2 from *Arabidopsis* and rice is the presence of a $(GC)_9$ microsatellite just before the stop codon in rice. Interestingly, the resulting translated amino acids are moderately

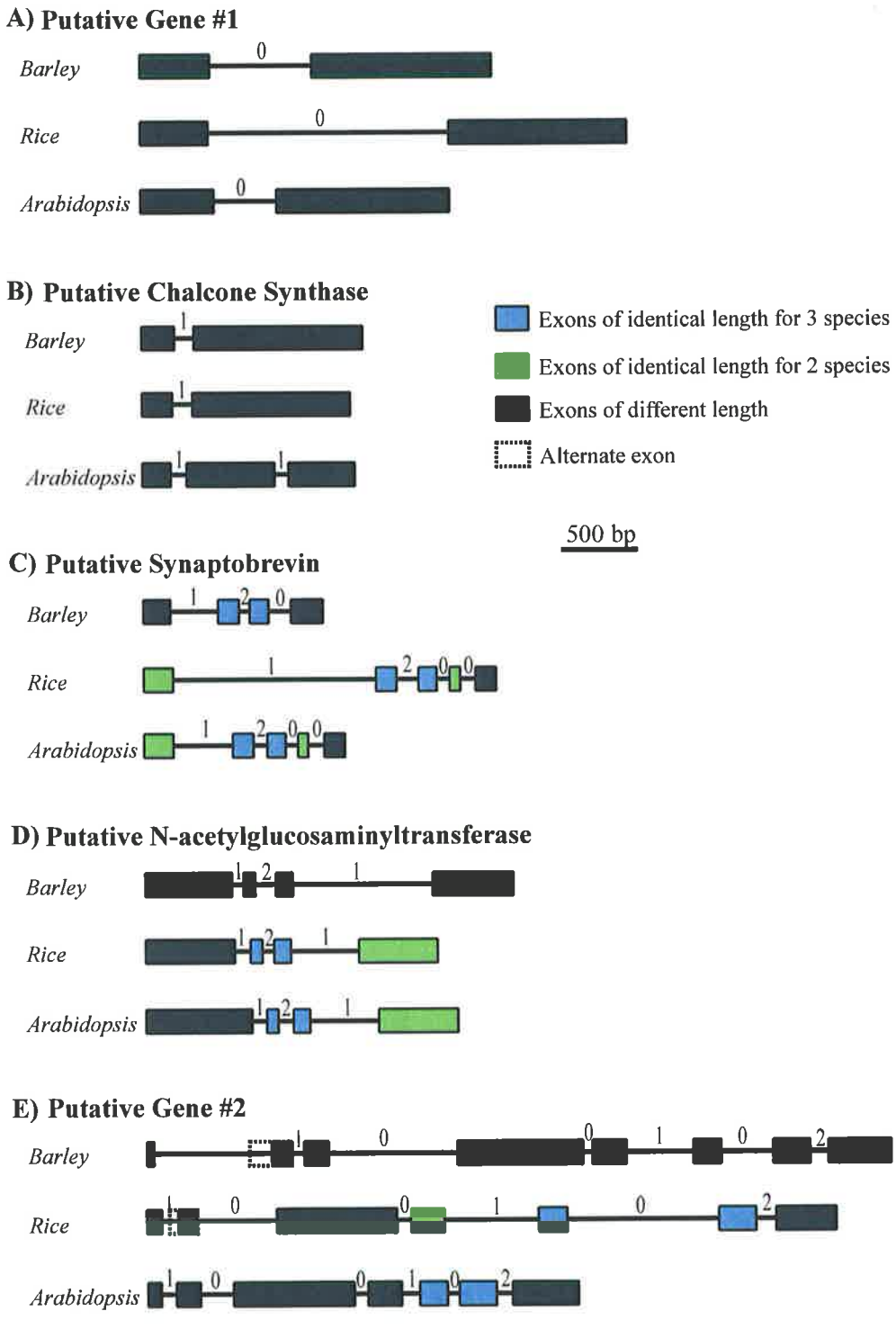


Figure 3.4. Structure of the genes located within the barley contig, the colinear rice region, and their closest *Arabidopsis* homologs (also see Table 3.1). Intron phase is indicated by the number above each intron.

conserved in the barley protein although the microsatellite structure is no longer present. Exon 2 of *HvCHS* contains an additional stretch of 23 codons not found in either of the other two species.

A high level of similarity exists between the *HvVAMP* protein (215 aa) and both the *OsVAMP* protein (219 aa, 90% similarity) and closest *Arabidopsis* homolog (At1g04760, 220 aa, 84% similarity, Table 3.2). The gene structure in barley and rice was confirmed by alignment of the genomic sequence with wheat and rice ESTs (Table 3.2). However, two rice ESTs (Genbank accession numbers CB667109 and CB667110) showed that the second intron of the rice transcript was not being spliced. Both the 5' and 3' splice sites maintain reasonable homology to the 5'CAG:GTAAGT3' and 5'GCAG:G3' plant consensus sites and the UA content of the intron is within the expected range. However, intron 2 does not contain a strong branchpoint consensus and this could reduce splice efficiency (Simpson *et al.*, 2002). In addition, both ESTs were from 3 week old leaf tissue that had been inoculated with rice blast 24 hours before harvest. It is, therefore, possible that improper splicing is either tissue specific or somehow induced by infection. However, this is purely speculative without the availability of ESTs from other tissue types. The elimination of this splice event introduces a premature stop codon immediately after the predicted splice site (codon 111). If the splice site is preserved, *OsVAMP* would maintain identical exon length and structure to the entire *Arabidopsis* gene with the exception of one less codon in the fourth exon (represented in Figure 3.4C). The length of exons 2 and 3 is also conserved in *HvVAMP*. However, intron 4 has been removed in comparison to the coding sequences of the other species

HvGlcNAc and *OsGlcNAc* are almost identical in length (425 v 426 aa, respectively) and demonstrate a high degree of similarity (87%, Table 3.2). The slightly larger *Arabidopsis* homolog (At5g39990, 447 aa) is 76% similar to both the barley and rice proteins. The gene structure in barley and rice was confirmed by alignment of the genomic sequence

with barley, sorghum, and rice ESTs (Table 3.2). Exons 2 and 3 are identical in length in all three species and exon 4 is identical in length in *Arabidopsis* and rice (Figure 3.4D). In addition, the 1st and 4th exons of *HvGlcNAc* differ in length from those of *HvGlcNAc* by only 3 and 2 codons, respectively. Alternative splicing in *Arabidopsis* to conserve the length of the first exon is highly unlikely as 6 of the 9 consensus bases are absent including the mandatory GT at the precise site of excision.

A high level of similarity exists between the *HvPG2* protein (753 aa) and both the *OsPG2* protein (683 aa, 84% similarity) and closest *Arabidopsis* homolog (At1g74790, 695 aa, 72% similarity, Table 3.2). The gene structure in barley and rice was confirmed by alignment of the genomic sequence with wheat and rice ESTs, respectively (Table 3.2). However, no homologous ESTs were found for the extreme 5' end of either gene. Therefore, two alternate structures for the barley protein were considered. The first, predicted by the Rice Genome Automated Annotation System (RiceGAAS; Sakata *et al.*, 2002), introduced an additional exon and resulted in a 723 aa gene product (Figure 3.4E). The second involved locating the first in-frame start codon upstream of the last confirmed gene region and encoded a 753 aa protein. This second gene structure is represented by a dashed region extending the length of exon 2 in Figure 3.4E. Neither alternative contained any similarity in the extreme 5' region of either *OsPG2* or *Arabidopsis* homologs at the protein or nucleotide level. However, the latter maintains the exon/intron structure of the *Arabidopsis* gene and EST homology extends 56 bp into what would otherwise be the intron region of the first alternative structure. Two alternate structures were also considered for the 5' terminal end of *OsPG2*. The first, predicted by RiceGAAS, maintained the exon/intron structure of the *Arabidopsis* gene and resulted in a 683 aa protein (Figure 3.4E). The second, determined by the first ATG start codon upstream of the last confirmed region with ESTs, eliminated an exon and introduced a premature stop codon 9 aa into the protein. This second gene structure is represented by a dashed region extending the length of exon 2 in Figure 3.4E. Again, no similarity to the

5' terminal end of either barley alternative or the *Arabidopsis* gene was observed at the protein or nucleotide level. Exons 5 and 6 are of identical length in all three species and exon 4 is of identical length in barley and rice. The *Arabidopsis* homolog is annotated as containing similarity to a hedgehog interacting protein from *Mus musculus* (GI number 4868122).

3.3 Discussion

3.3.1 Gene Islands and Intergenic Space

The current gene content of higher plants is estimated to range between 25,000 and 43,000 genes (Miklos *et al.*, 1996). Therefore, an average gene density of 1 gene every 123 to 250 kb would be expected in barley (5300 Mb) assuming even gene distribution. Furthermore, cytogentic studies have previously reported an increase in gene density along the chromosome arms moving away from the centromere towards the telomeres (Gill *et al.*, 1996a; Akhunov *et al.*, 2003). Regardless, despite the location of the hardness locus at the extreme distal end of 5HS, the results reported here suggest a local concentration of genes with approximately 1 gene every 25 kb. This is in concordance with the pattern of genome organization found within other large contiguous regions of barley which demonstrate an average density of 1 gene every 20 kb (1 gene every 12 to 103 kb; Panstruga *et al.*, 1998; Shirasu *et al.*, 2000; Dubcovsky *et al.*, 2001; Rostoks *et al.*, 2002; Wei *et al.*, 2002; Yan *et al.*, 2002; Gu *et al.*, 2003). Moreover, the presence of "gene islands" appears to be wide-spread among several members of the grass family with large genome size, including maize and wheat (SanMiguel *et al.*, 1996; Feuillet *et al.*, 1999; Tikhonov *et al.*, 1999; Wicker *et al.*, 2001). However, not all genes are located within clusters. A span of 96 kb separates *HvPG2* from the nearest upstream gene (*HvGSP*) and a minimum 43 kb gene void exists downstream. In addition, only a single gene was found within the 103 kb barley BAC 745c13 (Rostoks *et al.*, 2002) and on *T. monococcum* BAC111I4 the *RGA-1* gene was isolated from other genes by a minimum of 31 kb (Wicker *et al.*, 2001).

The presence of different transposable elements within the barley contig was the primary contributor to the patterns of genome organization mentioned above and the major factor in generating the vast difference in length between the colinear rice and barley sequences. Although over 75% of the barley sequenced region reported here is composed of repetitive elements, only one element, a 5 kb Ty1/*copia* retrotransposon, was present within the orthologous rice sequence (Figure 3.1B). A third of the repetitive sequence in the barley region consists of the BARE retrotransposon family, with both BARE-1 and BARE-2 contributing equally. This is 3-fold higher than BARE-1 average genome levels estimated in cultivated barley; however, other members of the Hordeae were found to have as much as 40% of their genomes composed of BARE-1 alone (Vicient *et al.*, 1999). Evidence for the breakup of microcolinearity among grass species by nested transposable element insertion has also been reported between the closely related species of sorghum (748 Mb) and maize (2500 Mb; Arumuganathan *et al.*, 1991; Chen *et al.*, 1998; Tikhonov *et al.*, 1999). At the *sh2/a1* locus, with the exception of a single gene duplication in sorghum, gene number and orientation was completely conserved between the two species despite a 3-fold difference in the overall locus lengths of the orthologous sequences (Chen *et al.*, 1998). Furthermore, only 15% of the *adh* locus in sorghum was found to be composed of non-genic sequence compared to over 74% in the orthologous locus in maize (Tikhonov *et al.*, 1999).

It is interesting that the only retrotransposon insertion in the rice sequence occurred within the *ψOsATPase-2* pseudogene. *ψHvATPase-1* was also disrupted by the insertion of a *copia* element of similar length within the same region of the gene. The two elements demonstrate complete lack of nucleotide homology and generate target site footprints of different lengths indicating separate elements within the *copia* family. The presence of several copies of these ATPase genes within the orthologous loci may lead to a level of mutational tolerance. However, given that the insertion of retrotransposons into

coding sequence is rare (SanMiguel *et al.*, 1996), the independent insertion of different elements into the same gene in colinear regions of two different grass species seems extremely coincidental. Especially given that this is the only retroelement insertion within the rice region.

It has been suggested that differences in intron length could also account for a portion of the differences observed in genome size. A greater proportion of rice introns (64%) were longer than their barley counterparts. However, the total length of intron sequence within a given gene favored a similar number of genes in barley (2 genes) as in rice (3 genes, Figure 3.4). When the introns of rice and *Arabidopsis* were compared, all but one rice intron were longer and the total intron length within a gene was always greater in rice. Interestingly this was not the case when comparing the barley and *Arabidopsis* genes despite a considerably larger difference in genome size. Although a greater number of barley introns (69%) were longer than their *Arabidopsis* equivalents, only 3 of the 5 genes gained length (Figure 3.4). In both cases, the longer total intron length in *Arabidopsis* was a result of an extra intron. Although longer intron size within the grass genes suggests either a greater frequency of large insertions or a better retention of such insertions, this may be compensated for by a greater number of smaller introns within *Arabidopsis* genes. Similar comparisons in intron length were reported in barley BAC 635P2 (Dubcovsky *et al.*, 2001). However the positional bias for introns located between codons (phase 0) noted in BAC 635P2 was contrary to the results obtained in this study. These results indicate a bias towards introns positioned within codons (64%) and an additional bias towards phase 1 (located between thc 1st and 2nd codon positions) introns over phase 2 (located between the 2nd and 3rd codon positions) introns. In every case, intron phase was conserved between all three species (Figure 3.4)

3.3.2 *Gene Discovery and Determination of Gene Structure*

Despite the extensive collection of ESTs in the public database, sequences of full length ESTs are still fairly rare. In addition, ESTs for a particular gene are often only represented from a single developmental stage or tissue type and, therefore, may only represent one of many alternative splicing events. The only two available rice ESTs for the synaptobrevin gene indicate failure to splice intron 2 resulting in a putatively severely truncated protein. However, the highly conserved protein similarity and gene structure compared to the barley and *Arabidopsis* proteins indicates that either this gene is still properly spliced in other tissues or under other conditions in rice or the mutations leading to improper splicing have occurred so recently that homology has not yet been degraded. Gene prediction programs, which are reasonably accurate in locating genic regions, often fall short in discerning the intricacies of specific gene structure. The automated gene prediction of *HvPG2* eliminated two entire exons, truncated a third, and generated a false start site. The automated prediction of the *OsPG2* generated an additional exon and introduced a new intron which altered the termination site of the gene; however, it was helpful in discerning the most probable start site in the absence of full length ESTs and with the *Arabidopsis* sequence as a guide. In both these instances, predicted genes from the completely sequenced *Arabidopsis* and rice genomes proved to be an invaluable tool for discerning specific gene structures.

3.3.3 *Microcolinearity and Genome Evolution*

Although some repetitive sequences are remnants of ancient insertion events, the vast majority of transposable element insertions occurred post speciation (SanMiguel *et al.*, 1998). The presence of these elements can often complicate the detection of orthologous loci for comparative genomics studies as critical regions of similarity could be missed within the sea of non-homologous intergenic DNA. The removal of all repetitive elements from the barley sequence generated a more "ancestral" template greatly facilitating the identification of the colinear rice sequence.

A wide variety of small chromosomal rearrangements have occurred between the region containing the *Ha* locus in barley and its colinear rice sequence (Figure 3.3). An inter-chromosomal event concluded in the translocation of the putative chalcone synthase gene. Although at least three copies of ATPase were present within the colinear region in both species, sequence homology revealed a greater conservation among paralogs within the same species than between orthologs of the different species. This indicated a total of six different independent duplications involving one gene inversion post speciation. Three further gene duplications involving a minimum of one inversion also arose from the ancestral grain texture gene in the barley genome. An intra-chromosomal rearrangement resulted in the repositioning of two conserved gene clusters. One of these gene clusters, GC2 (*VAMP*, *GlcNAc*, and *GSP*), has also been conserved in *Triticum monococcum* (Chantret *et al.*, 2004). The high level of conservation in this particular region was further demonstrated by the low level of transposon insertion. No transposable elements were present within GC2 in *T. monococcum* compared to other sequenced contiguous regions of the genome that are composed of 70-80% repetitive elements (Wicker et al., 2001; SanMiguel et al., 2002; Wicker et al., 2003b). Furthermore, the only element insertions within GC2 in the barley region occurred outside of the conserved region with T. monococcum between *GSP* and *PG2*.

Several additional breaks in colinearity existed between the wheat and barley genomes. The rice and wheat sequences contained a putative gene just upstream of GC2 which was not present in the barley sequence (Chantret *et al.*, 2004). Neither genome contained the *CHS* gene located in this position in the barley sequence indicating this translocation event occurred in the barley genome relative to the ancestral grass sequence. Similarly, a putative gene was present in the rice and barley sequences downstream of GC2 which was not found in wheat (Chantret *et al.*, 2004). Therefore, it is probable that the intra-chromosomal rearrangement observed between rice and barley involved the relocation of

the other gene cluster, GC1 (*ATPase* and *PG1*). Furthermore, the puroindoline genes were positioned downstream of GC2 and in the same orientation as GSP in wheat (Chantret *et al.*, 2004), while the hordoindolines were located upstream and in the opposite orientation in barley. All three grain texture genes in wheat and barley demonstrated orthologous relationships indicating that this rearrangement occurred post gene duplication. Extended sequencing of the *T. monococcum* region and addition sequences from related grass species are necessary to discern the exact series of evolutionary events.

A small level of microcolinearity still exists between the two grass species and *Arabidopsis*. The closest homologs to the putative N-acetylglucosaminyltransferase and ATPase are over 14 kb apart on chromosome 5 in reverse orientation separated by one additional gene. In addition, the closest *Arabidopsis* homologs to *PG1* and *PG2* are located only 1 kb apart in similar orientation on chromosome 1. Although the closest homolog to the putative synaptobrevin gene was also located on chromosome 1 it was completely separated from this gene cluster.

Only two other studies have compared large orthologous regions from rice and barley at the sequence level. At the *Xwg644* locus, despite one gene inversion and a single gene duplication in barley as compared to rice, the gene order of all four orthologs was completely conserved (Dubcovsky *et al.*, 2001). A single gene inversion and one gene duplication was also reported at the *Rph7* locus (Brunner *et al.*, 2003). However, inserted within the conserved gene order of four barley gene family members was a segment of 153 kb containing six additional genes not present in the colinear rice region (Brunner *et al.*, 2003). A rice homolog for each additional barley gene was found located elsewhere within the rice genome suggesting the occurrence of at least one post translocation event. The comparison of the region containing the *Ha* locus in barley with its colinear rice sequence represents the most complicated configuration of small chromosomal

rearrangements to be reported between grass species thus far. This may reflect historical evolutionary pressures and/or the telomeric location of these genes in barley. Well conserved colinearity with rice has been frequently reported along proximal regions of the Triticeae chromosomes such as the *Vrn1* (Yan *et al.*, 2003), *Ph1* (Roberts *et al.*, 1999), and *Gpc-B1* (Distelfeld *et al.*, 2004) loci in wheat. However, colinearity has recently been reported to be less conserved at the telomeric regions of the chromosomes among the wheat genomes. Moreover, a breakdown of microcolinearity has repeatedly be shown in comparative studies involving rice and distal regions of the wheat and barley genomes (Kurata *et al.*, 1994) including the *Rpg1* (Kilian *et al.*, 1997), LMW *Glu-A3/SRLK/Lrk10/Tak/Lr10* (Feuillet *et al.*, 1999; Guyot *et al.*, 2004) and *Sh2/X1/X2/A1* (Li *et al.*, 2002) regions in wheat and barley. The results from the *Ha* locus described here, demonstrate that this trend of colinearity breakdown within telomeric chromosomal regions extends beyond the genetic level to the sequence level. However, breaks in colinearity are not limited to distal regions. A comparison of the locations of physically mapped wheat ESTs and the first draft of the rice genomic sequence revealed that within the wheat genome, regardless of chromosomal location, even the most conserved regions of colinearity contain homologous sequences from more than one region of the rice genome (Sorrells *et al.*, 2003; La Rota *et al.*, 2004). These results support the view that grass genomes are more fluid than first anticipated and that structural and functional relationships are complex.

3.3.4 Conclusions

The utilization of the Morex BAC library allowed the construction of a physical map spanning the region harboring the *Ha* locus in barley. Detailed knowledge of the genome content and organization of the barley region enabled a comprehensive study of the level of microcolinearity with respect to the corresponding rice region. The presence of numerous chromosomal rearrangements and conflicting gene content indicate a complex evolutionary history since the divergence of the two species. These results demonstrate

the limitations of model systems for association mapping and positional cloning and; therefore, stress the importance for the continual development of necessary genomic resources for implementing structural and functional approaches directly in the species of interest. Nevertheless, comparative genomics with both the rice and *Arabidopsis* homologs proved to be useful in determining gene structure. The physically defined and genetically characterized region also provided a template for appropriate gene selection and primer design in order to investigate the patterns of sequence diversity across the genomic region and its organization into haplotypes. The next chapter explores the sequence diversity of five different genes i.e. *hinb-1*, *hinb-2*, *hina*, *HvGSP*, and *HvPG2* across different sample sets representing cultivated, landrace, and wild barley genepools.

CHAPTER 4: INVESTIGATION OF THE PATTERNS OF GENETIC DIVERSITY IN THE REGION THAT SPANS THE BARLEY *HA* LOCUS

4.1 Introduction

Nucleotide diversity accounts for the major proportion of phenotypic variation seen between and within organisms. Initial studies of intra- and interspecific diversity were performed through the analysis of isozymes (Hamrick *et al.*, 1990; Hamrick *et al.*, 1997). Although these studies provided a framework for understanding the complexities of genetic diversity, detection of nucleotide substitution was limited to those directly affecting the translated protein and comparisons between species were extremely difficult. The development of several molecular markers systems, such as RFLPs, AFLPs, and SSRs, expanded diversity studies to include regions located outside of the coding sequence and positions of synonymous change ultimately providing more realistic estimates of nucleotide diversity (Powell *et al.*, 1996a; Powell *et al.*, 1996b; Mueller *et al.*, 1999). These methods were also amenable to high throughput automation enabling the evaluation of large sample sizes. However, none of these marker systems provided knowledge of specific nucleotide changes. Therefore, an increasing number of diversity studies now rely on direct sequencing of genomic regions and the comparison of single nucleotide polymorphisms (SNPs; Rafalski, 2002a; Rafalski, 2002b). Direct sequencing ensures the detection of all possible mutations within the region of interest enabling precise estimates of genetic diversity and revealing the underlying organization of individual polymorphisms as defined haplotypes. Furthermore, these haplotypes can be partitioned into gene regions, namely exon, intron, and flanking sequence and sequence differences, including silent, synonymous, and nonsynonymous, providing new opportunities to investigate diversity patterns otherwise masked by regional trends.

The unrestricted resolution obtained through direct sequencing has facilitated the application of population genetic approaches for investigating the degree to which different forces have played in maintaining diversity levels and shaping evolutionary

history. Sequence analyses of gene loci within related species of *Arabidopsis*, *Leavenworthia*, *Caenorhabditis*, and *Lycopersicon* have demonstrated a strong correlation between diversity levels and the degree of inbreeding/self-compatability (Liu *et al.*, 1998; Liu *et al.*, 1999; Savolainen *et al.*, 2000; Baudry *et al.*, 2001). Likewise, a comparison of the levels of nucleotide diversity observed among different loci along chromosome 1 in maize revealed an association between local diversity levels and recombination rates (Tenaillon *et al.*, 2001). Furthermore, studies investigating the level of diversity between different populations/genepools of maize have contributed invaluable insight into the role human intervention played as an added selective pressure during domestication and cultivation (Goloubinoff *et al.*, 1993; Hilton *et al.*, 1998; Eyre-Walker *et al.*, 1998; White *et al.*, 1999).

This chapter presents the use of direct sequence analysis as a tool for unveiling haplotype structure and patterns of nucleotide diversity across the region surrounding the *Ha* locus in barley. Comparison of diversity patterns between cultivated material and related wild germplasm provided an estimate of the diversity level retained during the domestication and cultivation processes, thereby, offering insight into the feasibility of utilizing wild material as a potential source of new genetic diversity for breeding programs. Further comparisons of diversity patterns between different genes, regions of genes, and sequence site types provided a better insight into the processes of evolution which have acted upon the region.

4.2 Results

4.2.1 Diversity within the Region Harboring the Ha locus in Barley

Four independent gene regions were sampled across the 220 kb gene space of the region containing the hardness locus: *hinb-1* and *hinb-2*, *hina*, *GSP*, and exon 3 of *PG2* (see Section 2.12). All four gene regions were amplified across a sample of 123 individuals: 74 cultivated, 15 landraces, and 34 *H. spontaneum* accessions (see Section 2.12). The

aligned sequences resulted in a combined total of 7121 bp, excluding insertions-deletions (indels). The 29 indels ranged from 1 to 17 bp in length and accounted for only 1.5% of the total sequence length. A total of 295 SNPs were found producing an average of one SNP every 24 bp. This was not inclusive of the 176 singleton sequence variants that each occurred in only one line from the sample. The average density of SNPs was increased to one every 15 bp when singleton variants were considered. Of the 295 SNPs, 45 caused amino acid replacements (nonsynonymous), 46 were located within the coding region but had no affect on the translated protein (synonymous), and 250 were either synonymous or located in a non-coding region (silent). Likewise the 176 singletons were defined by 27 nonsynonymous, 24 synonymous, and 149 silent variants. Overall, 2.5% and 3.9% of the nonsynonymous sites, 8.3% and 12.6% of the synonymous sites, and 4.7% and 7.5% of the silent sites were polymorphic with respective exclusion and inclusion of singleton polymorphisms. The percentage of polymorphic nonsynonymous sites was statistically lower than the percentage of polymorphic synonymous or silent sites regardless of the inclusion or exclusion of singleton polymorphisms (Chi-Square Test, $P < 0.00003$).

4.2.2 Patterns of Diversity across the Individual Gene Regions

4.2.2.1 GSP

The sequenced *GSP* gene region totaled 1802 bp including 492 bp of coding sequence and 1310 bp of flanking sequence. Five indels account for 1% (14 bp) of the sequenced region. Together with the indels, 156 SNPs, of which 102 were singletons, defined 32 haplotypes (Figure 4.1). Despite the abundance of rare sequence variants, the exclusion of singletons and indels only reduced the haplotype number to 27. Haplotypes determined by the exclusion of non-parsimony informative sites, such as indels and singletons, will from here on be referred to as haplotype groups.

SNPs appeared to be more or less evenly distributed across the entire gene region (Table 4.1). Likewise, polymorphisms within the coding region were equally spread between

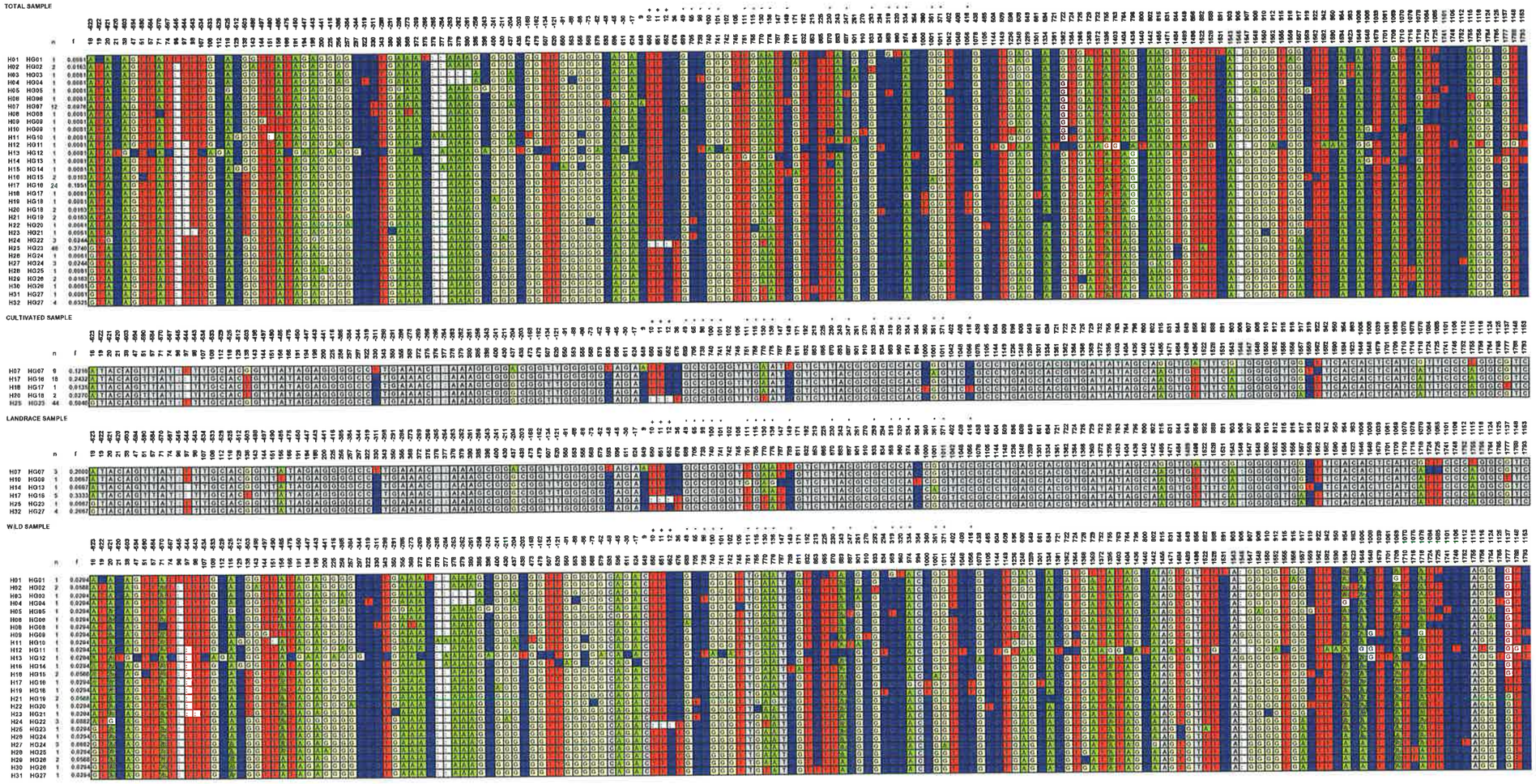


Figure 4.1. Multiple alignment of a 1802 bp region of *GSP*. Only the positions of polymorphisms are indicated. Positions are labeled by both the distance away from the A in the start codon (position 1) and by the position in the consensus of the alignment. The haplotype (H) and haplotype group (HG) designations, the number of lines within each haplotype, and haplotype frequencies (f) are included. Positions of nonsynonymous change are indicated (*). Positions fixed within a given subset although polymorphic in the sample as a whole have been grayed.

Table 4.1. Estimates of nucleotide polymorphism within different germplasm samples across the *GSP* gene region. Number of haplotypes, number of mutations (η), and gene length were calculated after the omission of indels. Length of regions including indels are denoted in parentheses. Tajima's D statistic significance levels are denoted: * P<0.5 and **P<0.1.

Region	Length, bp	Haplotypes	η	θ/bp based on		D
				S_n	π	
***GSP* entire sample (n = 123)**						
Overall	1788 (1802)	31	156	0.01589	0.00479	-2.30491**
Silent	1408.03		136	0.01794	0.00504	
Synonymous	109.03		19	0.03236	0.01038	
Nonsynonymous	379.97		20	0.00977	0.00388	
5' flanking	531 (640)	23	56	0.01619	0.00407	-2.35969**
coding	489 (492)	26	39	0.01443	0.00533	-1.95272*
3' flanking	668 (670)	25	61	0.01668	0.00508	-2.20764**
***GSP* cultivated sample (n = 74)**						
Overall	1788 (1802)	5	21	0.00240	0.00324	1.05816
Silent	1408.03		17	0.00247	0.00331	
Synonymous	109.03		3	0.00564	0.00696	
Nonsynonymous	379.97		4	0.00216	0.00298	
5' flanking	531 (640)	3	5	0.00161	0.00244	1.14820
coding	489 (492)	5	7	0.00294	0.00387	0.77912
3' flanking	668 (670)	5	9	0.00276	0.00354	0.74357
***GSP* landrace sample (n = 15)**						
Overall	1788 (1802)	5	26	0.00446	0.00463	0.16105
Silent	1408.03		21	0.00457	0.00477	
Synonymous	109.03		4	0.01128	0.00995	
Nonsynonymous	379.97		5	0.00405	0.00411	
5' flanking	531 (640)	5	6	0.00290	0.00332	0.51038
coding	489 (492)	5	9	0.00566	0.00541	-0.16295
3' flanking	668 (670)	5	11	0.00506	0.00530	0.18214
***GSP* wild sample (n = 34)**						
Overall	1788 (1802)	26	147	0.02011	0.00754	-2.36210**
Silent	1408.03		128	0.02223	0.00837	
Synonymous	109.03		16	0.03592	0.01641	
Nonsynonymous	379.97		19	0.01223	0.00449	
5' flanking	531 (640)	20	55	0.01970	0.00765	-2.35270**
coding	489 (492)	23	35	0.01700	0.00714	-2.11688*
3' flanking	668 (670)	24	57	0.02050	0.00774	-2.31257**

```
GSP_p1  MKT-FLLSFLALVASTAIAQYAEVPSPAAEVPTGDGFGEWVAMSPSVSGSEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p2  MKTFFLLSFLALVASTAIAQYAEVPSPAAEVPTGDGFGEWVAMSPSVSGSEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p3  MKTFFLLSFLALVASTAIAQYAEVPSPAAEVPTGDGFGEWVAMTPSVSGFEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p4  MKTFFLLSFLALVASTAIAQYAEVPSPAAEVPTGDGFGEWVAMTPSVSGSEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p5  MKTFFLLSFLALVASTAIAQYAEVPSPAAEVPTGDGFGEWVAMTPSVSGSEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p6  MKTFFLLSFLALVASTAIAQYAEVPSPAAEVPTGDGFGEWVAMTPSVSGSEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p7  MKTFFLLSFLALVASTAIAQYAEVPSPAAEVPTGDGLGEWVAMTPGVSGSEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p8  MKTFFLLSFLALVASTAIAQYAEVPSPAAEVPTGDGFGEWVAMTPSVSGSEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p9  MKTFFLLSFLALVASTTIAQYAEVPSPAAEVPTGDGFGEWVAMTPSVSGSEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p10 MKTFFLLSFLALVASTAIAQYAEVPSPAAEVPTGDGFGQWVAMTPSVSGSEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p11 MKTFFLLSFLALVASTAIAQYVEVPSPAAEVPTGDGFGEWVAMTPGVSGSEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p12 MKTFFLLSFLALVASTAIAQYAEVPSPAAEVPTGDGFGEWVAMTPSVSGSEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p13 MKTFFLLSFLALVASTAIAQYAEVPSPAAEVPTGDGFGEWVAMTPSVSGSEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p14 MKTFFLLSFLALVASTAIAQYAEVPSPAAEVPTGDGFGEWVAMTPSVSGSEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p15 MKTFFLLSFLALVASTAIAQYAEVPSPAAEVPTGDGFGEWVAMTPSVSGSEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p16 MKTFFLLSFLALVASTAIAQYAEVPSPAAEVPTGDGFGEWVAMTPSVSGSEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p17 MKTFFLLSFLALVASTAIAQYAEVPSPAAEVPTGDGFGEWVAMTPSVSGSEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p18 MKTFFLLSFLALVASTAIAQYAEVPSPAAEVPMGDGFGEWVAMTPSVSGSEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p19 MKTFFLLSFLALVASTAIAQYAEVPSPAAEVPMTDGFGEWVAMTPSVSGSEQCEQEQPKLNSCSDYVMDRCVTK
GSP_p20 MKTFFLLSFLALVASTAIAQYVEVPSPAAEVPTGDGFGEWVAMTPSVSGSEQCEQEQPKLNSCSDYVMDRCVTK

GSP_p1  DMLLSWVFPRTWGKRSCEEVQDQCCQQLRQTTPDCRCKAIWTSIQGDLSGFMSFQQGLEARTLQTAKSLPSKCN
GSP_p2  DMLLSWVFPRTWGKRSCEEVQDQCCQQLRQTTPDCRCKAIWTSIQGDLSGFMSFQQGLEARTLQTAKSLPSKCN
GSP_p3  DMLLSWVFPRTWGKRSCEEVQDQCCQQLRQTTPDCRCKAIWTSIQGDLSGFMSFQQGLEARTLQTAKSLPSKCN
GSP_p4  DMLLSWVFPRTWGKRSCEEVQDQCCQQLRQTTPDCRCKAIWTSIQGNLSGFMSFQQGLEARTLQTAKSLPSKCN
GSP_p5  DMLLSWVFPRTWGKRSCEEVQDQCCQQLRQTTPDCRCKAIWTSIQGDLSGFMSFQQGLEARTLQMAKSLPSKCN
GSP_p6  DMLLSWVFPRTWGKRSCEEVQDQCCQQLRQTTPDCRCKAIWTSIQGDLSGFMSFQQGLEARTLQTAKSLPSKCN
GSP_p7  DMLLSWVFPRTWGKRSCEEVQDQCCQQLRQTTPDCRCKAIWTSIQGDLSGFMSFQQGLEARTLQTAKSLPSKCN
GSP_p8  DMLLSWVFPRTWGKRSCEEVQDQCCQQLRQTTPDCRCEAIWTSIQGDLSGFMSFQQGLEARTLQMAKSLPSKCN
GSP_p9  DMLLSWVFPRTWGKRSCEEVQDQCCQQLRQTTPDCRCKAIWTSIQGDLSGFMSFQQGLEARTLQTAKSLPSKCN
GSP_p10 DMLLSWVFPRTWGKRSCEEVQDQCCQQLRQTTPDCRCKAIWTSIQGDLSGFMSFQQGLEARTLQTAKSLPSKCN
GSP_p11 DMLLSWVFPRTWGKRSCEEVQDQCCQQLRQTTPDCRCKAIWTSIQGDLSGFMSFQQGLEARTLQTAKSLPSKCN
GSP_p12 DMLLSWVFPRTWGKRSCEEVQDQFCQQLRQTTPDCRCKAIWTSIQGDLSGFMSFQQGLEARTLQTAKSLPSKCN
GSP_p13 DMLLSWVFPRTWGKRSCEEVQDQCCQQLRQTTVDCRCKAIWTSIQGDLSGFMSFQQGLEARTLQTAKSLPSKCN
GSP_p14 DMLLSWVFPRTWGKRSCEEVQDQCCQQLRQTTPDCRCKAIWTSIQGDLSDFMSFQQGLEARTLQTAKSLPSKCN
GSP_p15 DMPLSWVFPRTWGKRSCEEVQDQCCQQLRQTTPDCRCKAIWTSIQGDLSGFMSFQQGLEARTLQTAKSLPSKCN
GSP_p16 DMLLSWVFSRTWGKRSCEEVQDQCCQQLRQTTSDCRCKAIWTSIQGDLSGFMSFQQGLEARTLQTAKSLPSKCN
GSP_p17 DMLLSWVFSRTWGKRSCEEVQDQCCQQLRQTTPDCRCKAIWTSIQGDLSGFMSFQQGLEARTLQTAKSLPSKCN
GSP_p18 DMLLSWVFPRTWGKRSCEEVQDQCCQQLRQTTPDCRCKAIWTSIQGDLSGFMSFQQGLEARTLQTAKSLPSKCN
GSP_p19 DMPLSWVFPRTWGKRSCEEVQDQCCQQLRQTTPDCRCKAIWTSIQGDLSGFMSFQQGLEARTLQTAKSLPSKCN
GSP_p20 DMLLSWVFPRTWGKRSCEEVQDQCCQQLRQTTPDCRCKAIWTSIQGDLSGFMSFQQGLEARTLQTAKSLPSKCN

GSP_p1  IDPKYCNIPITSGYYW*
GSP_p2  IDPKYCNIPITSGYYW*
GSP_p3  IDPKYCNIPITSGYYW*
GSP_p4  IDPKYCNIPITSGYYW*
GSP_p5  IDPKYCNIPITSGYYW*
GSP_p6  IDPKYCNIPITSGYYW*
GSP_p7  IDPKYCNIPITSGYYW*
GSP_p8  IDPKYCNIPITSGYYW*
GSP_p9  IDPKYCNIPITSGYYW*
GSP_p10 IDPKYCNIPITSGYYW*
GSP_p11 IDPKYCNIPITSGYYW*
GSP_p12 IDPKYCNIPITSGYYW*
GSP_p13 IDPKYCNIPITSGYYW*
GSP_p14 IDPKYCNIPITSGYYW*
GSP_p15 IDPKYCNIPITSGYYW*
GSP_p16 IDPKYCNIPITSGYYW*
GSP_p17 IDPKYCNIPITSGYYW*
GSP_p18 IDPKYCNIPITSGYYW*
GSP_p19 IDPKYCNIPITSGYYW*
GSP_p20 IDPKYCNIPITSGYYW*
```

Figure 4.2. Multiple protein sequence alignment of *GSP*.

synonymous (19) and nonsynonymous (20) changes (Table 4.1). Interestingly, all haplotypes existing in the cultivated and landrace samples could be completely defined by replacement substitutions and one 3 bp indel within the coding region. These nonsynonymous substitutions were predominately of intermediate frequency and only two resulted in nonconservative amino acid replacements under basic biochemical classification (polar, nonpolar, aromatic, acidic, basic, and cysteine). The resulting protein variants were also found among the 19 protein variants within the wild material (Figure 4.2). The cultivated and landrace sample sets were marked by a notable decrease in diversity within the 5' flanking region relative to the coding or 3' flanking regions.

4.2.2.2 hina

The sequenced *hina* gene region spanned 450 bp of coding sequence and 1025 bp of flanking sequence to equal 1475 bp total. Although this region had the highest accumulation of indel events (indels = 12), indels only accounted for 3% (47 bp) of the total sequence and no haplotypes were formed by the presence of an insertion-deletion event alone (Figure 4.3). This region also demonstrated the lowest occurrence of singletons (η_s = 15). Nevertheless, these sequence variants increased the number of haplotypes to 28 from the 23 haplotype groups defined by 54 informative SNPs (Figure 4.3). A limited study of eight cultivars previously suggested a high level of diversity in the coding region of *hina* based on haplotype frequency (Beecher *et al.*, 2001). Despite coding sequence contributing only 30% of the total sequence length of the hina genic region in the present study and expectations of higher constraints within coding sequence, the distribution of SNPs was highly biased to the coding region (Table 4.2). This bias was reflected in higher nucleotide diversity values for coding sequence relative to both the 5'and 3' flanking sequence. The diversity level in the coding region was particularly elevated in the wild germplasm where it showed over twice the level of nucleotide diversity compared to flanking regions.

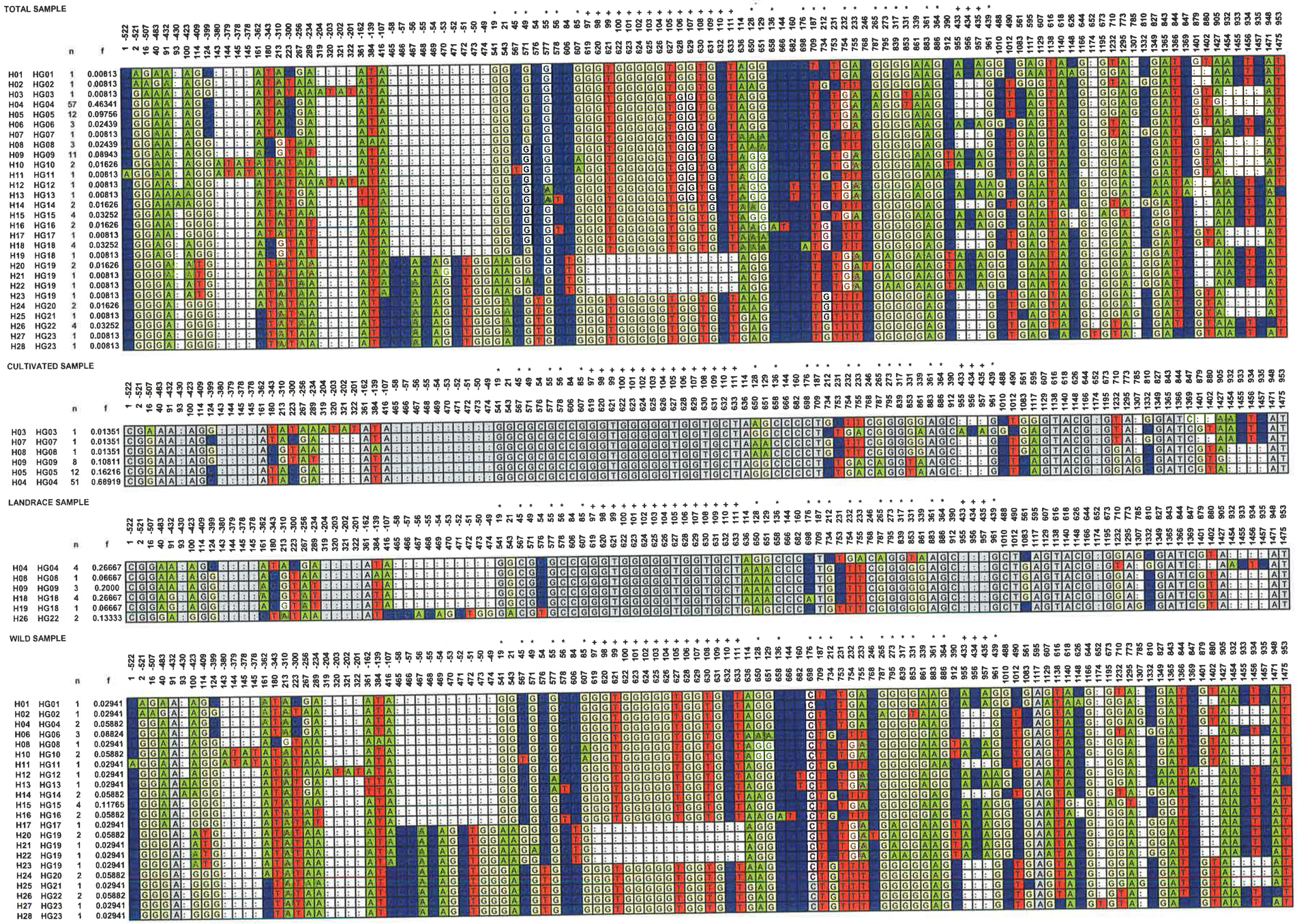


Figure 4.3. Multiple alignment of a 1475 bp region of *hina*. Only the positions of polymorphisms are indicated. Positions are labeled by both the distance away from the A in the start codon (position 1) and by the position in the consensus of the alignment. The haplotype (H) and haplotype group (HG) designations, the number of lines within each haplotype, and haplotype frequencies (f) are included. Positions of nonsynonymous change are indicated (*). Positions fixed within a given subset although polymorphic in the sample as a whole have been grayed.

Table 4.2. Estimates of nucleotide polymorphism within different germplasm samples across the *hina* gene region. Number of haplotypes, number of mutations (η), and gene length were calculated after the omission of indels. Length of regions including indels are denoted in parentheses. Tajima's D statistic significance levels are denoted: * P<0.5 and **P<0.1.

Region	Length, bp	Haplotypes	η	θ/bp based on		D
				S_n	π	
***hina* entire sample (n = 123)**						
Overall	1428 (1475)	28	67	0.00858	0.00710	-0.58649
Silent	1096.24		47	0.00629	0.00796	
Synonymous	100.24		12	0.02223	0.01697	
Nonsynonymous	331.76		20	0.01119	0.00976	
5' flanking	503 (522)	15	17	0.00628	0.00646	0.08289
coding	432 (450)	18	32	0.01375	0.01144	-0.50352
3' flanking	493 (503)	14	18	0.00678	0.00394	-1.16253
***hina* cultivated sample (n = 74)**						
Overall	1428 (1475)	6	24	0.00341	0.00297	-0.39626
Silent	1096.24		18	0.00335	0.00256	
Synonymous	100.24		3	0.00586	0.00656	
Nonsynonymous	331.76		6	0.00360	0.00428	
5' flanking	503 (522)	4	8	0.00326	0.00277	-0.38398
coding	432 (450)	4	9	0.00413	0.00481	0.43210
3' flanking	493 (503)	6	7	0.00290	0.00151	-1.18448
***hina* landrace sample (n = 15)**						
Overall	1428 (1475)	6	28	0.00596	0.00695	0.69458
Silent	1096.24		21	0.00585	0.00670	
Synonymous	100.24		6	0.01751	0.02079	
Nonsynonymous	331.76		7	0.00630	0.00775	
5' flanking	503 (522)	5	12	0.00734	0.00879	0.76905
coding	432 (450)	4	13	0.00894	0.01082	0.82615
3' flanking	493 (503)	3	3	0.00186	0.00157	-0.45679
***hina* wild sample (n = 34)**						
Overall	1428 (1475)	22	63	0.01062	0.01000	-0.27063
Silent	1096.24		44	0.00981	0.00918	
Synonymous	100.24		12	0.02912	0.02759	
Nonsynonymous	331.76		19	0.01403	0.01270	
5' flanking	503 (522)	12	16	0.00778	0.00712	-0.28270
coding	432 (450)	17	31	0.01698	0.01618	-0.27691
3' flanking	493 (503)	12	16	0.00794	0.00752	-0.1725

```
hina_p1  MKAFFLVGLLALVASAAFAQYGEVVGSYEGGAGGGGAQQCPLETKLDSCRNYLLDRCTTMKDFPVTWRWWRW
hina_p2  MKAFFLVGLLALVASAAFAQYGEVVGSYEGGAGGGGAQQCPLGTKLDSCRNYLLDRCTTMKDFPVTWRWWTW
hina_p3  MKAFFLVGLLALVASAAFAQYGEVVGSYEGGAGGGGAQQCPLGTKLDSCRNYLLDRCTTMKDFPVTWRWWTW
hina_p4  MKAFFLVGLLALVASAAFAQYGEVVGSYEGGAGGGGAQQCPLETKLDSCRNYLLDRCTKMKDFPVTWRWWRW
hina_p5  MKAFFLVGLLALVASAAFAQYGEVVGSYEGGAGGGGAQQCPLGTKLDSCRNYLLDRCTTMKDFPVTWRWWTW
hina_p6  MKAFFLVGLLALVASAAFAQYGEVVGSYEGGAGGGGAQQCPLGTKLDSCRNYLLDRCTTMKDLPVTWRWWTW
hina_p7  MKAFFLVGLLALVASAAFAQYGEVVGSYEGGAGGGGAQQCPLGTKLDSCRNYLLDRCTTMKDFPVTWRWWRW
hina_p8  MKAFFLVGLLALVASAAFAQYGEVVGSYKGGAGGGGAQQCPLGTKLDSCRNYLLDRCTTMKDFPVTWRWWTW
hina_p9  MKAFFLVGLLALVASAAFAQYGEVVGSYKGGAGGGGAQQCPLGTKLDSCRNYLLDRCTTMKDFPVTWRWWTW
hina_p10 MKAFFLVGLLALVASAAFTQYGEVVGSYEGGAGGGGAQQCPLGTKLDSCRNYLLDRCTTMKDFPVTWRWWTW
hina_p11 MKAFFLVGLLALVASAAFMQYGEVVGSYEGGAGGGGAQQCPLGTKLDSCRNYLLDRCTTMKDFPVTWRWWTW
hina_p12 MKAFFLVGLLALVASAAFVQYGEVVGSYEGGAGGGGAQQCPLETKIDSCRNYLLDRCTTMKDFPVTWRWWRW
hina_p13 MKAFFLIGLLALVARAAFAQYGEVVGSYEGGA-----QQCPLGTKLDSCRNYLLDRCTTMKDFPVTWRWWTW
hina_p14 MKAFFLIGLLALVARATFAQYGEVVGSYEGGA-----QQCPLGTKLDSCRNYLLDRCTTMKDFPVTWRWWTW
hina_p15 MKAFFLIGLLALVARAAFAQYGEVVGSYEGGA-----QQCPLGTKLDSCRNYLLDRCTTMKDFPVTWRWWTW

hina_p1  WKGGCLELLHDCCSQLGQMPPQCRCNIIQGSIQRDLGGVFGFQRDRTVKVIQAAKNLPPRCNQGPACNIPST
hina_p2  WKGGCEELLHDCCSQLSQMPPQCRCNIIQGSIQRDLGGFFGFQRDRTVKVIQAAKNLPPRCNQGPACNIPST
hina_p3  WKGGCEELLHDCCSQLGQMPPQCRCNIIQGSIQRDLGGVFGFQRDRTVKVIQAAKNLPPRCNQGPACNIPST
hina_p4  WKGGCLELLHDCCSQLGQMPPQCRCNIIQGSIQRDLGGVFGFQRDRTVKVIQAAKNLPPRCNQGPACNIPST
hina_p5  WKGGCEELLHDCCSQLGQMPPQCRCNIIQGSIQRDLGGVFGFQRDRTVKVIQAAKNLPPRCNQGPACNIPST
hina_p6  WKGGCEELLHDCCSQLGQMPPQCRCNIIQGSIQRDLGGVFGFQRDRTVKVIQAAKNLPPRCNQGPACNIPST
hina_p7  WKGGCLELLHDCCSQLGQMPPQCRCNIIQGSIQRDLGGVFGFQRDRTVKMIQAAKNLPPRCNQGPACNIPST
hina_p8  WKGGCEELLHDCCSQLGQMPPQCRCNIIQGSIQRDLGGVFGFQRDRTVKVIQAAKNLPPRCNQGPACNIPST
hina_p9  WKGGCEELLHDCCSQLGQMPPQCRCNIIQGSIQRDLGGVFGFQRDRTVKVIQAAKNLPPRCNQGPACNIPST
hina_p10 WKGGCEELLHDCCSQLGQMPPQCRCNIIQGSIQRDLGGVFGFQRDRTVEVIQAAKNLPPRCNQGPACNIPST
hina_p11 WKGGCEELLHDCCSQLGQMPPQCRCNIIQGSIQRDLGGVFGFQRDRTVEVIQAAKNLPPRCNQGPACNIPST
hina_p12 WKGGCLELLHDCCSQLGQMPPQCRCNIIQGSIQRDLGGVFGFQRDRTVKVIQAAKNLPPRCNQGPACNIPST
hina_p13 WKGGCEELLHDCCSQLGQIPPQCRCNIIQGSIQRDLGGVFGFQRDRTVKVIQAAKNLPPRCNQGPACNIPST
hina_p14 WKGGCEELLHDCCSQLGQIPPQCRCNIIQGSIQRDLGGVFGFQRDRTVKVIQAAKNLPPRCNQGPACNIPST
hina_p15 WKGGCEELLHDCCSQLGQIPPQCRCNIIQGSIQHDLGGVFGFQRDRTVKVIQAAKNLPPRCNQGPACNIPST

hina_p1  -TGYYW*
hina_p2  -TGYYW*
hina_p3  TTGYYW*
hina_p4  -TGYYW*
hina_p5  -TGYYW*
hina_p6  TTGYYW*
hina_p7  -TGYYW*
hina_p8  TTGYYW*
hina_p9  TTGYYW*
hina_p10 TTSYYW*
hina_p11 TTSYYW*
hina_p12 TTGYYW*
hina_p13 TTGYYW*
hina_p14 TTGYYW*
hina_p15 TTGYYW*
```

Figure 4.4. Multiple protein sequence alignment of *hina*.

Of the changes within the coding sequence, a greater proportion resulted in replacement substitutions than synonymous change. The majority of the replacement substitutions was found in low frequency ($f < 0.05$) and was predominantly restricted to the wild germplasm. However, 5 replacement substitutions were present at intermediate frequencies ($0.28 < f < 0.47$) and were represented in all three germplasm subsets. It is of interest that all but one of these caused a nonconservative replacement. In the cultivated sample, these 5 nonsynonymous changes defined 2 major protein variants with intermediate frequencies in contrast to 14 existing variants in the wild sample: one with an intermediate frequency of 0.353 and the rest with frequencies below 0.09 (Figure 4.4). The first cultivated protein variant (*hina_p1*, $f = 0.135$), composed of nucleotide haplotypes 3, 8, and 9, corresponded to the protein variants with the highest frequencies in both the landrace ($f = 0.667$) and wild ($f = 0.353$) samples. The second cultivated protein variant (*hina_p2*, $f = 0.851$), composed of nucleotide haplotypes 4 and 5, was also present in the landrace and wild material but at much lower frequencies ($f = 0.267$ and 0.0588, respectively). One cultivated line, Monte Cristo, contained the first three replacement substitutions from *hina_p2*, and the second two replacement substitutions from *hina_p1,* and an additional threonine amino acid. This variant was found at a frequency of 0.088 in the wild sample and was probably the product of at least one recombination event. The extra threonine residue was also present in 9 additional protein variants from the wild sample and three of these variants were further differentiated by the removal of a 5 amino acid stretch of residues.

4.2.2.3 hinb-1 and hinb-2

The contiguous sequenced region encompassing the hordoindoline-b duplication event was split into two separate contigs. Each contig was restricted to only the region of duplication in order to make more direct comparisons between the evolutionary histories of the paralogs. The sequenced gene regions of *hinb-1* and *hinb-2* were 1582 and 1671 bp, respectively. Although each contained 441 bp of coding sequence, several fixed

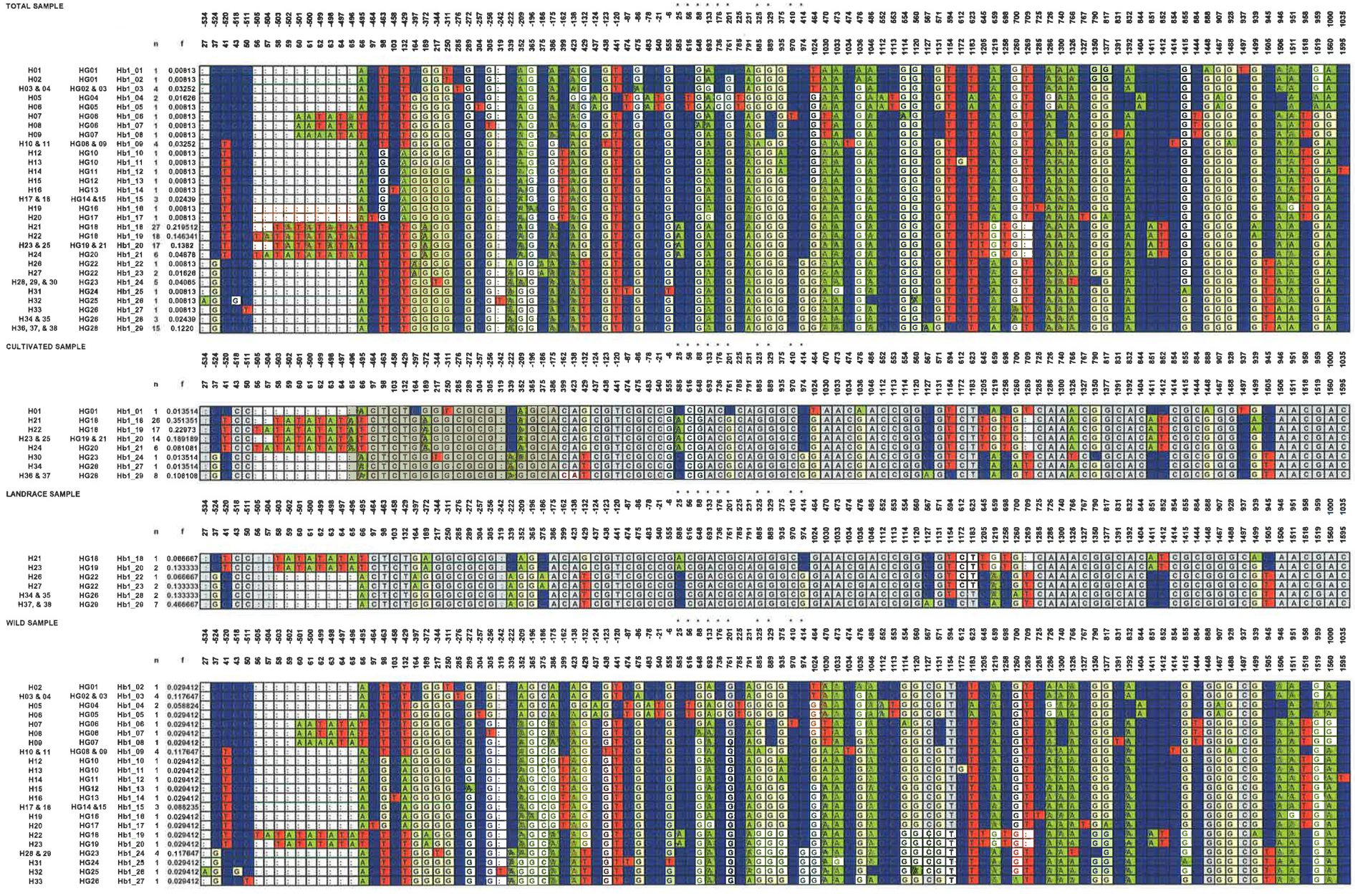


Figure 4.5. Multiple alignment of a 1582 bp region of *hinb-1*. Only the positions of polymorphisms are indicated. Positions are labeled by both the distance away from the A in the start codon (position 1) and by the position in the consensus of the alignment. The haplotype (H) and haplotype group (HG) designations refer to the *hinb* duplication region as a whole. Independent *hinb-1* haplotypes are indicated by Hb1_##. The number of lines within each haplotype and haplotype frequencies (f) are included. Positions of nonsynonymous change are indicated (*). Positions fixed within a given subset although polymorphic in the sample as a whole have been grayed.

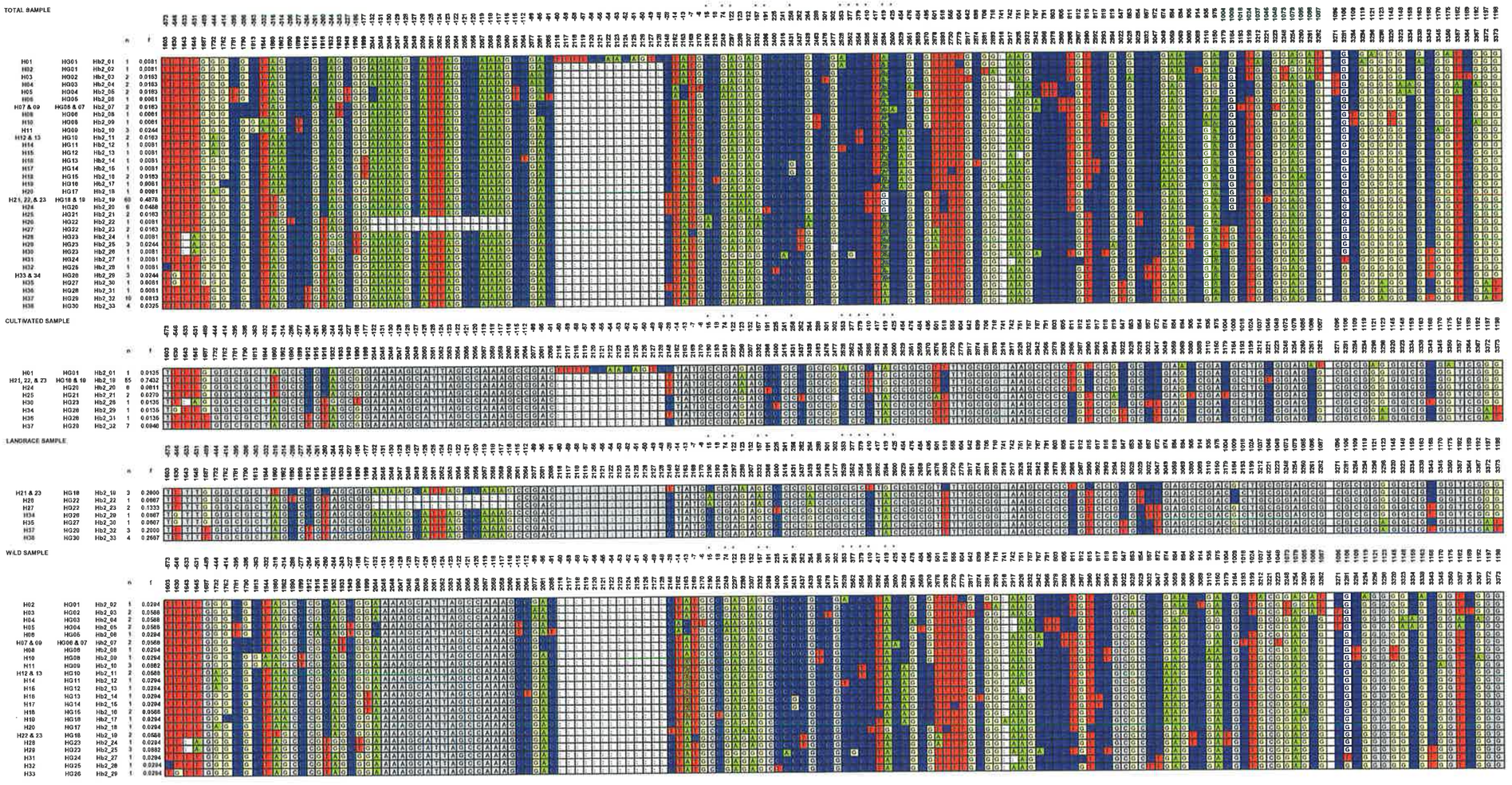


Figure 4.6. Multiple alignment of a 1671 bp region of *hinb-2*. Only the positions of polymorphisms are indicated. Positions are labeled by both the distance away from the A in the start codon (position 1) and by the position in the consensus of the alignment. The haplotype (H) and haplotype group (HG) designations refer to the *hinb* duplication region as a whole. Independent *hinb-2* haplotypes are indicated by Hb2_##. The number of lines within each haplotype and haplotype frequencies (f) are included. Positions of nonsynonymous change are indicated (*). Positions fixed within a given subset although polymorphic in the sample as a whole have been grayed. The gap after consensus positions 3262 indicates the end of the duplicated region.

indels have altered the length of the flanking sequences to 1141 bp and 1230 bp for *hinb-1* and *hinb-2*, respectively. A total of 98 mutations including 66 informative SNPs, 28 singletons, and 4 indels defined 24 haplotype groups containing 29 individual haplotypes across the *hinb-1* gene region (Figure 4.5). Similarly, a total of 104 mutations including 78 informative SNPs, 21 singletons, and 6 indels defined 28 haplotype groups containing 33 individual haplotypes across the *hinb-2* gene region (Figure 4.6).

Both hordoindoline-b gene copies appeared to have acquired similar levels of nucleotide substitution within the flanking sequences since duplication (Tables 4.3 & 4.4). However, selective constraints appeared to be operating to a higher degree in the *hinb-1* coding region as it has accumulated approximately half the number of sequence variants relative to *hinb-2*. Furthermore, these constraints seemed to be preferentially acting on synonymous sites. Nonsynonymous to synonymous change appeared at a ratio of 3 to 1 in the wild germplasm and no synonymous changes were present within either the cultivated or landrace samples despite 3 and 2 respective replacement substitutions. Similar to observations in the *hina* gene region, the majority of the nonsynonymous changes appeared at low frequencies ($f < 0.05$) and was predominately restricted to the wild sample. However, two replacement substitutions within each gene were found at intermediate frequencies between 0.24 and 0.45 and were represented in all three germplasm subsamples. All four of these nonsynonymous changes were conservative under basic biochemical classifications. The two intermediate frequency replacement substitutions in both the *hinb-1* and *hinb-2* coding regions each defined two major protein variants in the cultivated sample (Figures 4.7 & 4.8). The first *hinb-1* protein variant (*hinb-1_p1*), composed of nucleotide haplotypes 30, 34, 36, and 37, was perfectly associated with the first *hinb-2* protein variant (*hinb-2_p1*, $f = 0.135$) although one line, Monte Cristo, contained a frameshift mutation caused by a single bp deletion within the *hinb-2* coding sequence. The perfect correlation of these two protein variants extended to the landrace sample ($f = 0.600$) and was disrupted in the wild sample with 1 of the 7 lines

containing *hinb-1_p1* containing two additional replacement substitutions to *hinb-2_p1* and 13 additional lines containing *hinb-2_p1* that did not also have *hinb-1_p1*. The

Table 4.3. Estimates of nucleotide polymorphism within different germplasm samples across the *hinb-1* gene region. Number of haplotypes, number of mutations (η), and gene length were calculated after the omission of indels. Length of regions including indels are denoted in parentheses. Tajima's D statistic significance levels are denoted: * P<0.5 and **P<0.1.

Region	Length, bp	Haplotypes	η	θ/bp based on		D
				S_n	π	
***hinb-1* entire sample (n = 123)**						
Overall	1569 (1582)	26	93	0.01089	0.00710	-1.14123
Silent	1229.79		83	0.01253	0.00809	
Synonymous	101.79		3	0.00547	0.00141	
Nonsynonymous	339.21		10	0.00547	0.00353	
5' flanking	434 (547)	19	34	0.01145	0.00827	-0.90179
coding	441	11	13	0.00547	0.00304	-1.16680
3' flanking	593 (594)	21	46	0.01440	0.00908	-1.14355
***hinb-1* cultivated sample (n = 74)**						
Overall	1569 (1582)	6	28	0.00379	0.00312	-0.55869
Silent	1229.79		26	0.00434	0.00355	
Synonymous	101.79		0	0.00000	0.00000	
Nonsynonymous	339.21		3	0.00181	0.00153	
5' flanking	434 (547)	4	10	0.00383	0.00332	-0.35638
coding	441	2	2	0.00140	0.00118	-0.29297
3' flanking	593 (594)	6	16	0.00553	0.00437	-0.61194
***hinb-1* landrace sample (n = 15)**						
Overall	1569 (1582)	5	21	0.00412	0.00464	0.52060
Silent	1229.79		19	0.00475	0.00536	
Synonymous	101.79		0	0.00000	0.00000	
Nonsynonymous	339.21		2	0.00181	0.00202	
5' flanking	434 (547)	3	9	0.00518	0.00578	0.43055
coding	441	2	2	0.00139	0.00155	0.30213
3' flanking	593 (594)	5	10	0.00518	0.00590	0.53132
***hinb-1* wild sample (n = 34)**						
Overall	1569 (1582)	22	86	0.01341	0.00924	-1.15921
Silent	1229.79		77	0.01531	0.01054	
Synonymous	101.79		3	0.00720	0.00488	
Nonsynonymous	339.21		9	0.00649	0.00454	
5' flanking	434 (547)	17	32	0.01463	0.01222	-0.58463
coding	441	10	12	0.00665	0.00462	-0.97411
3' flanking	593 (594)	18	42	0.01729	0.00997	-1.53076

Table 4.4. Estimates of nucleotide polymorphism within different germplasm samples across the *hinb-2* gene region. Number of haplotypes, number of mutations (η), and gene length were calculated after the omission of indels. Length of regions including indels are denoted in parentheses. Tajima's D statistic significance levels are denoted: * P<0.5 and **P<0.1.

Region	Length, bp	Haplotypes	η	θ/bp based on		D
				S_n	π	
***hinb-2* entire sample (n = 123)**						
Overall	1636 (1671)	30	102	0.01146	0.00582	-1.60516
Silent	1298.11		88	0.01259	0.00626	
Synonymous	102.11		9	0.01637	0.00786	
Nonsynonymous	335.89		13	0.00719	0.00407	
5' flanking	547 (578)	19	32	0.01086	0.00585	-1.38235
coding	440 (441)	16	23	0.00928	0.00500	-1.39612
3' flanking	649 (652)	22	47	0.01345	0.00636	-0.63105
***hinb-2* cultivated sample (n = 74)**						
Overall	1636 (1671)	8	30	0.00372	0.00221	-1.27716
Silent	1298.11		26	0.00405	0.00233	
Synonymous	102.11		3	0.00602	0.00304	
Nonsynonymous	335.89		4	0.00244	0.00177	
5' flanking	547 (578)	6	9	0.00327	0.00218	-0.87358
coding	440 (441)	6	7	0.00326	0.00205	-0.91213
3' flanking	649 (652)	6	14	0.00441	0.00235	-1.33359
***hinb-2* landrace sample (n = 15)**						
Overall	1636 (1671)	6	22	0.00413	0.00468	0.55504
Silent	1298.11		18	0.00425	0.00484	
Synonymous	102.11		3	0.00893	0.00793	
Nonsynonymous	335.89		4	0.00364	0.00406	
5' flanking	547 (578)	5	7	0.00393	0.00452	0.54095
coding	440 (441)	4	7	0.00488	0.00497	0.06301
3' flanking	649 (652)	4	8	0.00377	0.00462	0.82197
***hinb-2* wild sample (n = 34)**						
Overall	1636 (1671)	21	91	0.01331	0.00853	-1.36710
Silent	1298.11		80	0.01487	0.00987	
Synonymous	102.11		8	0.01898	0.00810	
Nonsynonymous	335.89		11	0.00796	0.00335	
5' flanking	547 (578)	16	30	0.01301	0.00902	-1.08394
coding	440 (441)	14	19	0.01054	0.00446	-1.95094*
3' flanking	649 (652)	17	42	0.01583	0.01088	-1.13090

```
hinb1_p1  MKTLFLLALLALVASTTFAQYSVGGGYNDVGGGGGSQQCPQERPNLGSCKDYVMERCFTMKDFPLTWPTKWW
hinb1_p2  MKTLFLLAILALVASTTFAQYSVGGGYNDVGGGGGSQQCPQERPNLGSCKDYVMERCFTMKDFPLTWPTKWW
hinb1_p3  MKTLFLLALLALVASTTFAQYSVGGGYNDVGGGGGSQQCPQERPNLGSCKDYVMERCFTMKDFPLTCPTKWW
hinb1_p4  MKTLFLLALLALVASTTFAQYSVGGGYNDVGGGGGSQQCPQERPNLGSCKDYVMERCFTMKDFPLTWPTKWW
hinb1_p5  MKTLFLLALLALVASTTFAQYSVGGGYNDVGGGGGSQQCPQERPDLGSCKDYVMERCFTMKDFPLTWPTKWW
hinb1_p6  MKTLFLLALLALVASTTFVQYSVGGGYNDVGGGGGSQQCPQERPNLGSCKDYVMERCFRMKDFPLTWPTKWW
hinb1_p7  MKTLFLLALLALVASTTFAQYSVGGGYNDVGGGGGSQQCPQERPNLGSCKDYVMERCFTMKDFPLTWPTKWW
hinb1_p8  MKTLFLLALLALVASTTFAQYSVGGGYNDIGGGGGSQQCPQERPNLGSCKDYVMERCFTMKDFPLTWPTKWW
hinb1_p9  MKTLFLLALLALVASTTFAQYSVGGGYNDVGGGGGSQQCPQERPNLGSCKDYVMERCFTMKDFPLTWPTKWW
hinb1_p10 MKTLFLLALLALVASTTFAQYSVGGGYNDVGGGGGSQQCPQERPNLGSCKDYVMERCFTMKDFPLTWPTKWW

hinb1_p1  KGGCEQEVREKCCQQLSQIAPQCRCDAIRGVIQGKLGGIFGIGGGDVFKQIQRAQILPSKCNMGAECKFPSG
hinb1_p2  KGGCEQEVREKCCQQLSQIAPQCRCDAIRGVIQGKLGGIFGIGGGDVFKQIQRAQILPSKCNMGADCKFPSG
hinb1_p3  KGGCEQEVREKCCQQLSQIAPQCRCDAIRGVIQGKLGGIFGIGGGDVFKQIQRAQILPSKCNMGADCKFPSG
hinb1_p4  KGGCEQEVREKCCQQLSQIAPQCRCDAIRGVIQGKLGGIFGIGGGDVFKQIQRAQILPSKCNMGADCKFPSG
hinb1_p5  KGGCEQEVREKCCQQLSQIAPQCRCDAIRGVIQGKLGGIFGIGGGDVFKQIQRAQILPSKCNMGADCKFPSG
hinb1_p6  KGGCEQEVREKCCQQLSQIAPQCRCDAIRGVIQGKLGGIFGIGGGDVFKQIQRAQILPSKCNMGADCKFPSG
hinb1_p7  KGGCEQEVREKCCQQLSQIAPQCRCDAIRGVIQGKLGAIFGIGGGDVFKQIQRAQILPSKCNMGVDCKFPSG
hinb1_p8  KGGCEQEVREKCCQQLSQIAPQCRCDAIRGVIQGKLGAIFGIGGGDVFKQIQRAQILPSKCNMGADCKFPSG
hinb1_p9  KGGCEQEVREKCCQQLSQIAPQCRCDAIRGVIQGKLSGIFGIGGGDVFKQIQRAQILPSKCNMGADCKFPSG
hinb1_p10 KGGCEQEVREKCCQQLSQIAPQCRCDAIRGVIQGKLGAIFGIGGGDVFKQIQRAQILPSKCNMGADCKFPSG

hinb1_p1  YYW*
hinb1_p2  YYW*
hinb1_p3  YYW*
hinb1_p4  YYW*
hinb1_p5  YYW*
hinb1_p6  YYW*
hinb1_p7  YYW*
hinb1_p8  YYW*
hinb1_p9  YYW*
hinb1_p10 YYW*
```

Figure 4.7. Multiple protein sequence alignment of *hinb-1*.

```
hinb2_p1  MKTLFLLALLALVASTTSAQYSVGGGYNDVGGGGGSQQCPQERPNLGSCKDYVMERCFTMKDFPVTWPTKWW
hinb2_p2  MKTLFLLALLALVASTTSAQYSVGGGYNDVGGGGGSQQCPQERPNLGSCKDYVMERCFTMKDFPVTWPTKWW
hinb2_p3  MKTLFLLALLALVASTTSAQYSVGGGYNDVGGGGGSQQCPQERPNLGSCKDYVMERCFTMKDFLVTWPTKWW
hinb2_p4  MKTLFLLALLALVASTTSAQYSVGGGYNDVGGGGGSQQCPQERPNLGSCKDYVMERCFTMKDFPVTWPTKWW
hinb2_p5  MKTLLLLLALLALVASTTSAQYSVGGGYNDVGGGGGSQQCPQERPNLGSCKDYMMERCFTMKDFPVTWPTKWW
hinb2_p6  MKTLFLLALLALVASTTSAQYSVGGGYNDVGGGGGSQQCPQERPNLGSCKDYVMERCFTMKDFPVTWPTKWW
hinb2_p7  MKTLFLLALLALVASTTSAQYSVGGGYNDVGGGGGSQQCPQERPNLGSCKDYVMERCFTMKDFPVTWPTKWW
hinb2_p8  MKTLFLLALLALVASTTSAQYSVGGGYNDVGGGGGSQQCPQERPNLGSCKDYVMERCFTMKDFPVTWPTKWW
hinb2_p9  MKTLFLLALLALVASTTSAQYSVGDGYNDVGGGGGSQQCPQERPNLGSCKDYVMERCFTMKDFPVTWPTKWW
hinb2_p10 MKTLFLLALLALVASTTSAQYSVGGGYNDVGGGGGSQQCPQERPNLGSCKDYVMERCFTMKDFPVTWPTKWW
hinb2_p11 MKTLFLLALLALVASTTSAQYSVGGGYNDVGGGGGSQQCPRERPNLGSCKDYVMERCFTMKDFPVTWPTKWW

hinb2_p1  KGGCEHEVREKCCQQLSQIAPHCRCDAIRGVIQGKLGGIFGIGGGAVFKQIQRAQILPSKCNMGADCKFPSG
hinb2_p2  KGGCEHEVREKCCQQLSQIAPHCRCDAIRGVIQGKLGGIFGIGGGAVFKQIQRAQILPSKCNMGVDCRFPSG
hinb2_p3  KGGCEHEVREKCCQQLSQIAPHCRCDAIRGVIQGKLGGIFGIGGGAVFKQIQRAQILPSKCNMGVDCRFPSG
hinb2_p4  KGGCEHEVREKCCQQLSQIAPHCRCDAIRGVIQGKLGGIFGIGGGDVFKQIQRAQILPSKCNMGADCKFPSG
hinb2_p5  KGGCEHEVREKCCQQLSQIAPHCRCDAIRGVIQGKLGGIFGIGGGAVFKQIQRAQILPSKCNMGADCKFPSG
hinb2_p6  KGGCEHEVREKCCQQLSQIAPHCRCDAIRGVIQGKLGGIFGIGGGAVFKQIQRA*ILPSKCNMGADCKFPSG
hinb2_p7  KGGCEHEVREKCCQQLSQIAPHCRCDAIRGVIQGKLGGIFGIGGGVVFKQIQRAQILPSKCNMGADCKFPSG
hinb2_p8  KGGCEHEVREKCCEQLSQIAPHCRCDAIRGVIQGKLGGIFGIGGGAVFKQIQRAQILPSKCNMGADCKFPSG
hinb2_p9  KGGCEHEVREKCCQQLSQIAPHCRCDAIRGVIQGKLGGIFGIGGGAVFKQIQRAQILPSKCNMGADCKFPSG
hinb2_p10 KGGCEHEVREKCCQQLSQIAPHCRCDAIRGVIQGKLGGIFGIGGGAVFKQIQRAQILPSKCNMGADCKFHSG
hinb2_p11 KGGCEHEVREKCCQQLSQIAPHCRCDAIRGVIQGKLGGIFGIGGGAVFKQIQRGQILPSKCNMGADCKFPSG

hinb2_p1  YYW*
hinb2_p2  YYW*
hinb2_p3  YYW*
hinb2_p4  YYW*
hinb2_p5  YYW*
hinb2_p6  YYW*
hinb2_p7  YYW*
hinb2_p8  YYW*
hinb2_p9  YYW*
hinb2_p10 YYW*
hinb2_p11 YYW*
```

Figure 4.8. Multiple protein sequence alignment of *hinb-2*.

second *hinb-1* protein variant (*hinb-1_p2*) was perfectly correlated with the second *hinb-2* protein variant (*hinb-2_p2*) in both the wild (f = 0.0588) and landrace (f = 0.0200) samples. However, this correlation is incomplete in the cultivated sample as 2 of the 63 lines containing *hinb-1_p1* (f = 0.851) were altered by an additional replacement substitution in *hinb-2_p2*. The landrace and cultivated samples each contained one additional unique protein variant for both *hinb-1* and *hinb-2*. In contrast, the wild sample contained 7 additional *hinb-1* protein variants and 7 additional *hinb-2* protein variants.

4.2.2.4 PG2

A total of 594 bp of exon 3 were sequenced from the *PG2* gene. Together with two indels, 37 SNPs including 8 singletons defined 29 individual haplotypes that could be sorted into 24 separate haplotype groups (Figure 4.9). Although no intron or flanking sequence was obtained for comparison with the coding sequence, the diversity level for *PG2* was similar to those obtained for the coding regions of the other four genes analyzed. Substitutions were highly biased to synonymous sites resulting in a 3 to 1 ratio of synonymous to replacement change (Table 4.5). This bias was reflected in the 10-fold greater diversity level at synonymous sites. Three major protein variants were found at intermediate frequencies in the cultivated germplasm (Figure 4.10). All three variants were also present within the 9 protein variants in the wild sample. Three null alleles were also present as a result of two different frameshift mutations: a 2 bp deletion in Morex, Union, Volla, and Vogelsanger Gold and a 1 bp deletion in Plaisant.

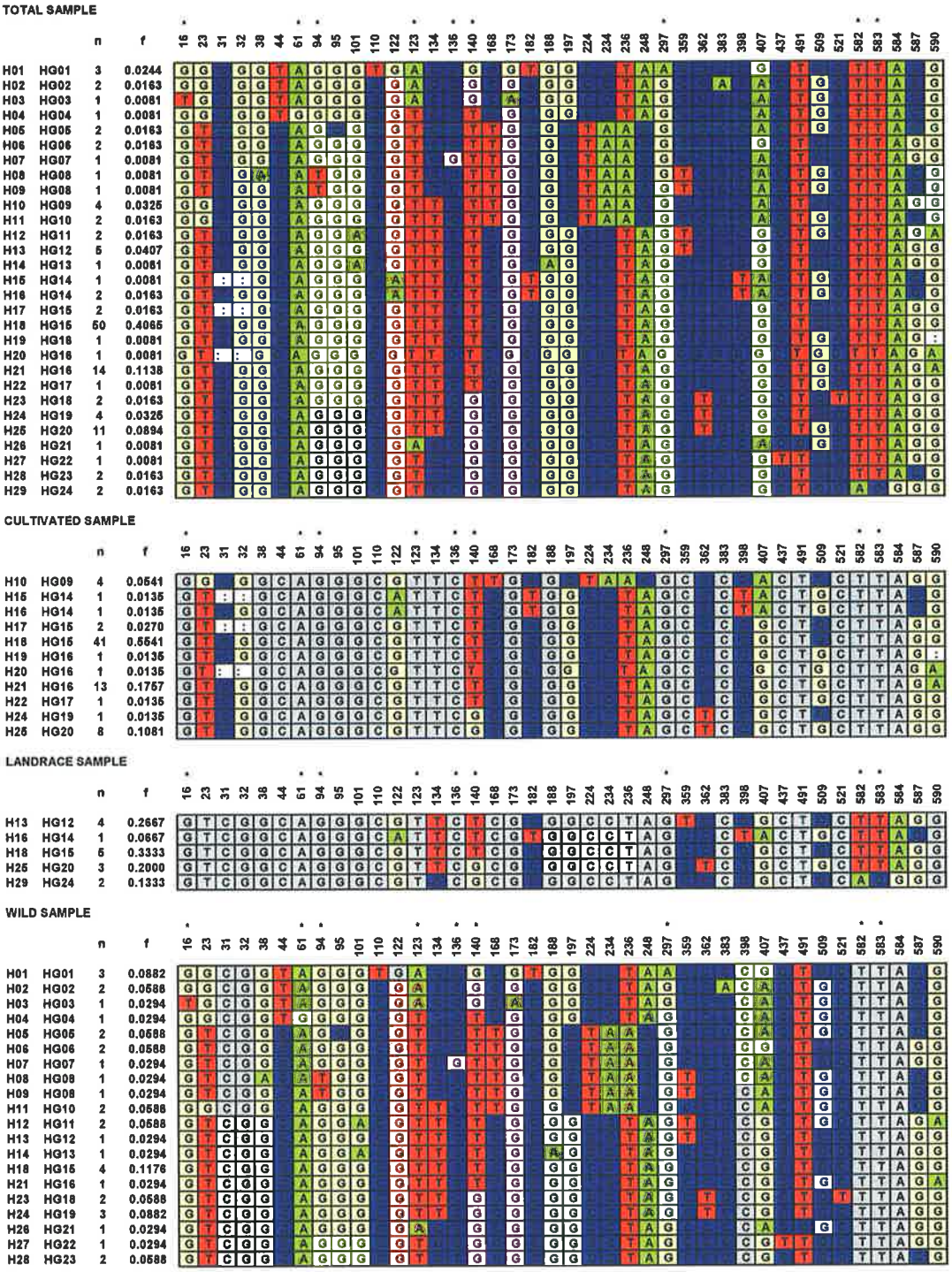


Figure 4.9. Multiple alignment of a 594 bp region of *PG2* exon 3. Only the positions of polymorphisms are indicated. Positions are labeled by both the distance away from the A in the start codon (position 1) and by the position in the consensus of the alignment. The haplotype (H) and haplotype group (HG) designations, the number of lines within each haplotype, and haplotype frequencies (f) are included. Positions of nonsynonymous change are indicated (*). Positions fixed within a given subset although polymorphic in the sample as a whole have been grayed.

```
PG_p1  QGGGRLPSSSSKLTDAWQSESDFCTSFGGDRSVCLSGSTVSFNATHPSASPKGVCLERIDNGSYAYLNMVPHPDG
PG_p2  QGGGRLPSSSSKLTDAWQSESDFCTSFGGDRSVCLSGSTVSFNATHPSASPKGVCLERIDNGSYAYLNMVPHPDG
PG_p3  QGGGRLPSSSSKLTDAWQSESDFCTSFGGDRSVCLSGSTVSFNATQPSASPKGVCLERIDNGSYAYLNMVPHPDG
PG_p4  QGGGRLPSSXSKLTDAWQSESDFCTSFGGDRSVCLSGSTVSFNATHPSASPKGVCLERIDNGSYAYLNMVPHPDG
PG_p5  QGGGRLPSSXSKLTDAWQSESDFCTSFGGDRSVCLSGSTVSFNATHPSASPKGVCLERIDNGSYAYLNMVPHPDG
PG_p6  QGGGRLPSSSSKLTDAWQSESDFCTSFGGDRSVCLSGSTVSFNATHPSASPKGVCLERIDNGSYAYLNMVPHPDG
PG_p7  QGGGRLPSSSSKLTDAWQSESDFCTSFGGDRSVCLSGSTVSFNATQPSASPKGVCLERIDNGSYAYLNMVPHPDG
PG_p8  QGGGRLPSSSSKLTDAWQSESDFCTSFGGDRSVCLSGSTVTFNATQPSASPKGVCLERIDNGSYAYLNMVPHPDG
PG_p9  QGGGMLPSSSSKLTDAWQSESDFCTSFGGDRSVCLSGSTVTFNATQPSASPKGVCLERIDNGSYAYLNMVPHPDG
PG_p10 QGGGRLPSSSSKLTDAWQSESDFCTSFGGDRSVCLSGSTVSFNASHPSASPKGVCLERIDNGSYAYLNMVPHPDG
Pg_p11 QGGGRLPSSSSKLTDAWQSGSDFCTSFGGDRSVCLSGSTVSFNATHPSASPKGVCLERIDNGSYAYLNMVPHPDG
PG_p12 QGGGRLPSSSSKLTDAWQSESDFCTSFGGDRSVCLSGSTVTFNATQPSASPKGVCLERIDNGSYAYLNMVPHPDG
PG_p13 QGGGRLPSSSSKLTDAWQSESDFCTSFGGDLSVCLSGSTVSFNATHPSASPKGVCLERIDNGSYAYLNMVPHPDG

PG_p1  SNRVFLGTQAGKILLATVPEQGSGGTLQFDEAGQFVDLTDQVHFDSTFGLMGMAFHPDFATNGRFFASYNCDRTK
PG_p2  SNRVFLGTQAGKILLATVPEQGSGGTLQFDEAGQFVDLTDQVHFDSTFGLMGMAFHPDFATNGRFFASYNCDRTK
PG_p3  SNRVFLGTQAGKILLATVPEQGSGGTLQFDEAGQFVDLTDQVHFDSTFGLMGMAFHPDFATNGRFFASYNCDRTK
PG_p4  SNRVFLGTQAGKILLATVPEQGSGGTLQFDEAGQFVDLTDQVHFDSTFGLMGMAFHPDFATNGRFFASYNCDRTK
PG_p5  SNRVFLGTQAGKILLATVPEQGSGGTLQFDEAGQFVDLTDQVHFDSTFGLMGMAFHPDFATNGRFFASYNCDRTK
PG_p6  SNRVFLGTQAGKILLATVPEQGSGGTLQFDEAGQFVDLTDQVHFDSTFGLMGMAFHPDFATNGRFFASYNCDRTK
PG_p7  SNRVFLGTQAGKILLATVPEQGSGGTLQFDEAGQFVDLTDQVHFDSTFGLMGMAFHPDFATNGRFFASYNCDRTK
PG_p8  SNRVFLGTQAGKILLATVPEQGSSGTLQFDEAGQFVDLTDQVHFDSTFGLMGMAFHPDFATNGRFFASYNCDRTK
PG_p9  SNRVFLGTQAGKILLATVPEQGSGGTLQFDEAGQFVDLTDQVHFDSTFGLMGMAFHPDFATNGRFFASYNCDRTK
PG_p10 SNRVFLGTQAGKILLATVPEQGSGGTLQFDEAGQFVDLTDQVHFDSTFGLMGMAFHPDFATNGRFFASYNCDRTK
Pg_p11 SNRVFLGTQAGKILLATVPEQGSGGTLQFDEAGQFVDLTDQVHFDSTFGLMGMAFHPDFATNGRFFASYNCDRTK
PG_p12 SNRVFLGTQAGKILLATVPEQGSGGTLQFDEAGQFVDLTDQVHFDSTFGLMGMAFHPDFATNGRFFASYNCDRTK
PG_p13 SNRVFLGTQAGKILLATVPEQGSGGTLQFDEAGQFVDLTDQVHFDSTFGLMGMAFHPDFATNGRFFASYNCDRTK

PG_p1  SPSCSGRCSCNSDVGCDPSKLGTDNGAQPCQYQVVVSEYSAKGLSSN
PG_p2  SPSCSGRCSCNSDVGCDPSKLGTDNGAQPCQYQVVVSEYSAKGLSSN
PG_p3  SPSCSGRCSCNSDVGCDPSKLGTDNGAQPCQYQVVVSEYSAKGLSSN
PG_p4  SPSCSGRCSCNSDVGCDPSKLGTDNGAQPCQYQVVVSEYSAKGLSSN
PG_p5  SPSCSGRCSCNSDVGCDPSKLGTDNGAQPCQYQVVVSEYSAKGLSSN
PG_p6  SPSCSGRCSCNSDVGCDPSKLGTDNGAQPCQYQVVVSEYSAKGLSXN
PG_p7  SPSCSGRCSCNSDVGCDPSKLGTDNGAQPCQYQVVVSEYSAKGTSSN
PG_p8  SPSCSGRCSCNSDVGCDPSKLGTDNGAQPCQYQVVVSEYSAKGLSSN
PG_p9  SPSCSGRCSCNSDVGCDPSKLGTDNGAQPCQYQVVVSEYSAKGLSSN
PG_p10 SPSCSGRCSCNSDVGCDPSKLGTDNGAQPCQYQVVVSEYSAKGLSSN
Pg_p11 SPSCSGRCSCNSDVGCDPSKLGTDNGAQPCQYQVVVSEYSAKGLSSN
PG_p12 SPSCSGRCSCNSDVGCDPSKLGTDNGAQPCQYQVVVSEYSAKGLSSN
PG_p13 SPSCSGRCSCNSDVGCDPSKLGTDNGAQPCQYQVVVSEYSAKGLSSN
```

Figure 4.10. Multiple protein sequence alignment of *PG2*.

Table 4.5. Estimates of nucleotide polymorphism within different germplasm samples across the *PG2* gene region. Number of haplotypes, number of mutations (η), and gene length were calculated after the omission of indels. Length of regions including indels are denoted in parentheses. Tajima's D statistic significance levels are denoted: * P<0.5 and **P<0.1.

Region	Length, bp	Haplotypes	η	θ/bp based on		D
				S_n	π	
***PG2* entire sample (n = 123)**						
Overall	591 (594)	23	36	0.01133	0.00694	-1.17237
Synonymous	143.06		27	0.03505	0.02264	
Nonsynonymous	438.94		9	0.00381	0.00196	
***PG2* cultivated sample (n = 74)**						
Overall	591 (594)	6	15	0.00521	0.00336	-1.02410
Synonymous	143.06		14	0.02002	0.01162	
Nonsynonymous	438.94		1	0.00046	0.00073	
***PG2* landrace sample (n = 15)**						
Overall	591 (594)	4	12	0.00622	0.00488	-0.83954
Synonymous	143.06		9	0.01906	0.01296	
Nonsynonymous	438.94		3	0.00208	0.00229	
***PG2* wild sample (n = 34)**						
Overall	591 (594)	20	33	0.01359	0.01389	0.08042
Synonymous	143.06		26	0.04380	0.04514	
Nonsynonymous	438.94		7	0.00384	0.00382	

4.2.3 Diversity among Germplasm

The ratio of silent site diversity (θ_{silent}) in the cultivated sample with respect to the landrace and wild samples was calculated for all gene regions in order to compare the levels of diversity within the different genepools. θ_{silent} is expected to be more accurate for inter-population comparisons because silent sites are less susceptible to changes in mutation rate as a result of selection. The average ratio, calculated by the summation of θ_{silent} across all loci prior to ratio determination, was 0.89 and 0.32 with respect to the landrace and wild samples (Table 4.6). As would be expected from a domestication bottleneck, all of the gene regions had a lower θ_{silent} value in the cultivated germplasm as compared to the wild sample and only one gene region, *PG2*, had a higher θ_{silent} value relative to the landrace sample. Although similar average ratios were obtained using π_{silent} as the diversity statistic (0.68 and 0.28 respectively; Table 4.6), these values estimated lower levels of diversity in the cultivated sample with respect to the levels observed in the other two genepools. This deviated from expectations under the assumption of a domestication bottleneck.

Table 4.6. Estimated levels of diversity for each gene region within the different barley genepools and the level of diversity (%) within the cultivated material with respect to the diversity levels within the landrace and wild samples.

	hinb-2			*hinb-1*			*hina*		
	CV	**LR**	**HS**	**CV**	**LR**	**HS**	**CV**	**LR**	**HS**
π_{silent}	2.33	4.84	9.87	3.55	5.36	10.54	2.56	6.70	9.18
θ_{silent}	4.05	4.25	14.87	4.34	4.75	15.31	3.35	5.85	9.81
% diversity (π)		48%	24%		66%	34%		38%	28%
% diversity (θ)		95%	27%		91%	28%		57%	34%
	GSP			*PG2*			*Average*		
	CV	**LR**	**HS**	**CV**	**LR**	**HS**	**CV**	**LR**	**HS**
π_{silent}	3.31	4.77	8.37	11.62	12.96	45.12	4.67	6.93	16.62
θ_{silent}	2.47	4.57	22.23	20.02	19.06	43.80	6.85	7.70	21.20
% diversity (π)		69%	40%		90%	26%		67%	28%
% diversity (θ)		54%	11%		105%	46%		89%	32%

CV – cultivated
LR – landrace
HS - wild

A domestication bottleneck is caused by a sharp decrease in population size as a result of the initial selection process of individual wild plants displaying favorable agricultural traits. Domesticated populations consequently sustain a loss of genetic diversity inversely proportional to the number of individuals involved during domestication and are usually characterized by a scarcity of rare sequence variants. Tajima's estimator of nucleotide diversity (π) is based on the average number of pairwise comparisons in a sample and, therefore, is largely determined by common variants (Tajima, 1983). In contrast, Watterson's estimator (θ) can be based on either the number of segregating sites (S) or the total number of mutations (η) within a sample and consequently does not account for allele frequency (Watterson, 1975). Therefore, diversity calculations using θ are much more influenced by either an excess or scarcity of rare variants and should be more sensitive to the effects of domestication. Nevertheless, only two of the five genes (*GSP* and *hinb-1*) demonstrated lower levels of diversity in the cultivated sample when calculated by θ_{silent} than π_{silent} relative to its wild ancestor (Table 4.6).

GSP, *hinb-1*, and *hinb-2* were the only genes that demonstrated a higher Tajima's D value in the cultivated sample than in the wild germplasm and only *GSP* had a higher value relative to the landrace sample (Tables 4.1-4.5). Tajima's D statistic measures the difference between the two different estimates of nucleotide polymorphism: nucleotide diversity (π) and Watterson's estimator (Tajima, 1989). Under the null hypothesis that diversity is solely influenced by mutation and random genetic drift, these two measures of diversity are expected to be equal. However, discrepancies between the estimators arise when the pattern of diversity is marked by an excess (negative Tajima's D) or scarcity (positive Tajima's D) of rare alleles. Statistics relative to the landrace material may have been affected by the limited sample scheme restricted to Fertile Crescent and the small sample size. Sample sizes greater than n = 50 have been recommended in order to achieve a reasonable power of detection (Simonsen *et al.*, 1995).

4.2.4 Signatures of Selection

Three standard tests for deviation from the neutral model were employed to investigate signatures of selection. The McDonald-Kreitman test is based on the expectation that both the synonymous and nonsynonymous sites within the coding region of a gene should share the same evolutionary history as a result of tight linkage (McDonald *et al.*, 1991). Therefore, the ratio of replacement to synonymous variation should be similar for both polymorphisms within species and fixed substitutions between species if the variation observed at both types of sites results from neutral mutations. This test can also be calculated substituting silent site variation for synonymous variation when non-coding flanking sequence is available. The corresponding genomic *Triticum tauschii* sequences for interspecific comparison were obtained from the public databases for all gene regions with one exception (Genbank accession numbers AF177219, AY159805, and AY159804). A *Triticum aestivum* cDNA (Genbank accession number BJ278101) homologous to *PG2* was used in the absence of any genomic wheat sequence for the region. Two gene regions tested yielded a significant result at the 5% confidence level:

hina using silent sites and *hinb-1* using synonymous sites (Table 4.7). The larger sample set was reanalyzed as three separate subsets corresponding to the cultivated, landrace, and wild samples. The significant results at both loci could be attributed to the patterns of diversity within the wild germplasm.

Table 4.7. McDonald-Krietman (MK) Test statistics and significance values.

		MK_{silent}		MK_{syn}	
		G value	**P-value**	**G value**	**P-value**
hinb-2	ALL	0.413	0.52028	0.250	0.61720
	CV	0.380	0.53779	0.123	0.72587
	LR	0.903	0.34194	0.124	0.72502
	HS	0.207	0.64882	0.233	0.62916
hinb-1	ALL	0.018	0.89273	4.744	0.02939*
	CV	0.163	0.68638	N/A	N/A
	LR	0.029	0.86501	N/A	N/A
	HS	0.006	0.93913	3.924	0.04759*
hina	ALL	4.876	0.02723*	0.047	0.82873
	CV	0.935	0.33356	0.304	0.58114
	LR	0.713	0.39837	0.131	0.71762
	HS	4.744	0.02939*	0.012	0.91212
GSP	ALL	0.421	0.51632	1.616	0.20364
	CV	0.263	0.60801	0.007	0.93263
	LR	0.514	0.47344	0.033	0.85487
	HS	0.351	0.55336	0.944	0.33115
PG2	ALL	N/A	N/A	0.034	0.85266
	CV	N/A	N/A	N/A	N/A
	LR	N/A	N/A	0.409	0.52267
	HS	N/A	N/A	2.059	0.15134

* significance level of $0.01<P<0.05$

The Hudson, Kreitman, and Aguadé (HKA) test also looks for correlation in the degree of variation that exists between and within populations (Hudson *et al.*, 1987). It is based on the expectation that in the absence of selection, both intraspecies polymorphism and interspecies divergence are proportional to the mutation rate and, therefore, the ratio of the two should be constant across all loci. All possible pairwise comparisons among the 4 gene regions with available homologous *T. tauschii* sequence were performed using the HKA test. *hinb-1* demonstrated significant results ($P<0.05$) in all pairwise comparisons using synonymous sites except for that with its paralogous sequence, *hinb-2* (Table 4.8).

None of the other pairwise comparisons yielded significant results at the 5% confidence level. Again, the larger sample set was reanalyzed as three separate subsets corresponding to the cultivated, landrace, and wild samples. All three subsets were significant at the 5% confidence level with respect to *hina* while only the landrace and wild samples were significant at the 5% confidence level in comparison with *GSP* (Table 4.8).

As mentioned above, significant negative Tajima's D values can also be an indication of directional selection. If a selective sweep persists to completion then all variation seen at the affected locus/loci will be the result of new mutations arising after fixation of the given genetic background. The recovery phase will be marked by an excess of rare variants (negative Tajima's D) denoted by a high proportion of singleton polymorphisms (Braverman *et al.*, 1995). In contrast, a partial selective sweep can be detected by a loss of rare alleles (positive Tajima's D) and diversity would appear as haplotype groups composed of deviations from the surviving ancestral haplotypes. *GSP* demonstrated significant ($P<0.01$) results due to an influx of singleton polymorphisms from the wild sample (Table 4.1). One wild accession, 181679, contributed 43 of the 102 singleton sequence variants (42%) within the wild germplasm assayed. To test if the sequence variants present in this line alone were enough to account for the excess of rare alleles, statistics were recalculated after its removal from the sample. Although the evidence for selection is not quite as strong, significant Tajima's D values were obtained at the 5% confidence level.

One other gene region is worth mentioning with respect to Tajima's D values. Despite the inability to detect a significant departure from the neutral model across the entire gene region of *hinb-2*, a significant ($P<0.05$) negative value was obtained in the wild sample when only the coding region was considered (Table 4.4).

Table 4.8. HKA Test statistics and significance values. Test was performed using all sites (red), silent sites (black), and synonymous sites (blue).

	hinb-2		*hinb-1*		*hina*	
	X^2	P-value	X^2	P-value	X^2	P-value
Total sample (n = 123)						
hinb-1	0.018	0.8921				
	0.001	0.9799				
	2.810	0.0937				
hina	0.039	0.8436	0.004	0.9503		
	0.246	0.6196	0.227	0.6335		
	0.808	0.3688	6.154	0.0131*		
GSP	1.299	0.2544	1.637	0.2007	1.788	0.1812
	1.724	0.1891	1.828	0.1764	3.213	0.0730
	1.781	0.1821	8.236	0.0041**	0.160	0.6890
Cultivated sample (n = 74)						
hinb-1	0.055	0.8143				
	0.037	0.8474				
	3.146	0.0761				
hina	0.113	0.7368	0.329	0.5661		
	0.054	0.8171	0.183	0.6686		
	0.082	0.7744	4.021	0.0449*		
GSP	0.249	0.6177	0.065	0.7983	0.731	0.3926
	0.220	0.6387	0.074	0.7853	0.513	0.4739
	0.012	0.9127	3.457	0.0630	0.031	0.8599
Landrace sample (n = 15)						
hinb-1	0.141	0.7074				
	0.033	0.8548				
	2.799	0.0943				
hina	0.771	0.3798	1.552	0.2128		
	0.793	0.3732	1.181	0.2771		
	0.866	0.3522	5.363	0.0206*		
GSP	0.013	0.9085	0.248	0.6186	0.667	0.4140
	0.038	0.8455	0.148	0.7004	0.570	0.4503
	0.170	0.6804	3.858	0.0495*	0.272	0.6019
***Hordeum spontaneum* sample (n = 34)**						
hinb-1	0.000	0.9883				
	0.000	0.9967				
	1.900	0.1681				
hina	0.000	0.9921	0.000	0.9962		
	0.097	0.7549	0.102	0.7497		
	1.013	0.3141	5.245	0.0220*		
GSP	1.074	0.300	1.114	0.2913	1.099	0.2945
	1.305	0.2532	1.317	0.2512	2.068	0.1504
	1.331	0.2487	5.809	0.0159*	0.016	0.8995

* significance level of 0.01<P<0.05
** significance level of 0.001<P<0.01

4.2.5 Recombination

The four-gametic test was used to determine the minimum number of recombination events (R_M) required to explain the pattern of diversity within each gene region. This

algorithm, described in Hudson and Kaplan (1985), predicts a recombination event only when all four possible gametal types are found within the sample. Therefore, this test is a conservative representation of recombination often underestimating the total number of recombination events throughout the history of the sample.

The nine recombination events detected within the *GSP* gene region were well distributed with three events each in the 5' flanking, coding, and 3' flanking regions. A similar distribution was noted for the 10 detected events within the *hina* gene region. The detection of 18 of the 19 events could be accounted for by the diversity patterns within the wild germplasm. Only one recombination event, located within *hina*, was detected within the landrace material and no recombination was detected for either gene within the cultivated sample.

Despite covering approximately twice the sequence length of either of the above mentioned genes, the gene region spanning both duplicated hordoindoline-b copies exhibited less evidence of recombination. The 5' flanking region of *hinb-1* and the 3' flanking region of *hinb-2* contained 3 and 4 respective recombination events. Only 1 recombination event was detected within the entire internal 2 kb segment starting and ending with the *hinb-1* and *hinb-2* gene coding regions, respectively. Despite the overall reduced recombination rate within the region, more recombination was detected within the cultivated and landrace samples with two events each. All but one event was detected within the wild material.

PG2 demonstrated an increased level of recombination exhibiting an equal number of recombination events as detected within the coding regions of the other four genes combined. All 6 events were detected within the wild sample with only 1 event detected in both of the other subsamples.

4.3 Discussion

4.3.1 Diversity within the Region Containing the Hardness Locus

This study investigated the nucleotide diversity in four separate gene regions located within the contig containing the barley *Ha* locus with three main goals:

1) to determine the level of diversity within the cultivated germplasm relative to landrace and wild samples,

2) to gain a better understanding of forces of evolution acting upon the different genes within the region

3) to compare the patterns of diversity seen within different germplasm.

In general, the region harboring the hardness locus was very diverse with one SNP every 15 to 24 bp depending on the inclusion or exclusion of singleton sequence variants. However, this frequency dropped by over 60% when only cultivated material was considered. When all loci were considered together, the level of diversity in the cultivated sample was 89% of that detected in the landrace sample (Table 4.6). Detection of only a marginal reduction of diversity could reflect the limited sampling scheme restricted to landraces from the Fertile Crescent and the small sample size. In contrast, the level of diversity found in the cultivated material was substantially diminished relative to the wild sample, containing only 32% of the level of variation detected within the later.

These results demonstrate a notable deviation from comparisons of other cereal crop plants and their closest wild ancestors. Several different studies have estimated that cultivated maize has retained ~70% of the proportion of diversity found in its closest wild ancestors (Goloubinoff *et al.*, 1993; Hilton *et al.*, 1998; Eyre-Walker *et al.*, 1998; White *et al.*, 1999; Tenaillon *et al.*, 2001). RFLP studies at the *RbcS* locus in hexaploid wheat also indicate only a 30% reduction in diversity relative to the wild diploid species (Galili *et al.*, 2000). A similar reduction was found in both cultivated rice and sorghum using isozyme data (Oka, 1988; Morden *et al.*, 1990). These reports are consistent with the suggestion that the use of cereal species in subsistence farming may have prevented a

decrease in population size sufficient to cause a severe domestication bottleneck (reviewed in Buckler *et al.*, 2001; Buckler *et al.*, 2002). However, certain genes have demonstrated a substantial decrease in diversity levels as a consequence of their direct involvement during the domestication process. The promoter region of *teosinte branched1* in domesticated maize showed a 97% reduction in variation as compared to its closest wild relative (Wang *et al.*, 1999). This gene was instrumental in the development of a single stalk maize plant (Doebley *et al.*, 1997).

Diversity levels within the region surrounding the *Ha* locus in barley exhibited an intermediate position between the high reduction of genetic diversity at loci critical to domestication and the moderate loss of diversity at otherwise neutral loci. Several considerations must be taken in the interpretation of these results. With maize as the exception, studies comparing the diversity levels of cultivated germplasm relative to their closest wild ancestors have been limited to only a few genetic loci. It is, therefore, hard to assess the true level of overall diversity reduction throughout the genome. In addition, these studies have been conducted using varying sample strategies and different marker systems which both have an ultimate effect on the levels of diversity detected. Furthermore, local genomic levels of diversity could deviate drastically from the genome average as a result of different evolutionary forces acting independently on specific chromosomal regions. The bulk of diversity studies in barley have focused on the wild species *Hordeum spontaneum* and few comparative studies exist relative to the cultivated species. Combined results from several loci involved in adaptive variation demonstrated similar levels of diversity in both cultivated and wild germplasm (Marmiroli *et al.*, 1999). This contrasts the results from within the sequence surrounding the *Ha* locus. However, several aspects of this study make it difficult to draw a direct comparison:

1) loci analyzed were among candidates for selection in the wild germplasm

2) marker system used (long primer-PCR, LP-PCR) has reduced the power to detect levels of diversity

3) wild material was limited to Israeli germplasm

Until assays of multiple loci are conducted using a similar experimental design, comprehension of the patterns of diversity between cultivated germplasm and their wild ancestors will remain limited.

4.3.2 Patterns of Nucleotide Diversity and Signatures of Selection within Hordeum spontaneum

Several studies of nucleotide diversity have indicated reduced genetic variation in regions of low recombination (Aguadé *et al.*, 1989; Stephan *et al.*, 1989; Begun *et al.*, 1992). Two potential explanations for this effect are selective sweeps and background selection. A "selective sweep" or "hitch-hiking effect" occurs when neutral sites become fixed in a population because of proximity to an advantageous mutation. In contrast, background selection occurs when neutral polymorphisms are eliminated from a population because of proximity to deleterious mutations. Since both of these processes are a consequence of linkage, this reduction of genetic variation is amplified in selfing species as a result of substantially reduced effective recombination rates (Kaplan *et al.*, 1989; Charlesworth *et al.*, 1993). This is exemplified in *Leavenworthia* and *Caenorhabditis* where a greater than 2-fold and 10-fold respective reduction in nucleotide diversity was found in inbreeding relative to outbreeding species (Liu *et al.*, 1998; Liu *et al.*, 1999). Furthermore, stochastic theory shows that species maintaining high levels of self-fertilization can demonstrate a decrease of up to 50% in genetic variation compared to their out-crossing relatives (Pollak *et al.*, 1992). In addition to decreased levels of genetic variation, the reduced effective recombination rate associated with inbreeders also predicts homogeneous patterns of genetic diversity extending across large chromosomal segments as a result of shared evolutionary history.

The wild material was marked by high levels of nucleotide diversity ranging from θ = 0.01062 to 0.02011 at the *hina* and *GSP* gene regions, respectively (Tables 4.1-4.5). This

diversity was arranged into a large number of distinct haplotypes (20-26) each occurring in no more than four lines (Figures 4.1, 4.3, 4.5, 4.6, and 4.9). A relatively high level of recombination was determined despite high levels of inbreeding with the smallest number of recombination events observed within the *hinb* duplicated region ($R_m = 8$).

Further partitioning of the gene regions and the analysis of different types of sites within the gene regions revealed somewhat contrasting patterns of nucleotide diversity. Four out of the five gene regions showed evidence of deviation from the neutral model of evolution. However, in each case a different pattern of nucleotide diversity could explain this departure. Supporting evidence at the *hinb-1* locus was the strongest with significant statistical results from both the HKA and McDonald Krietman tests (Tables 4.7 and 4.8). The HKA results were significant only when synonymous sites were considered, in contrast to the inclusion of all silent sites, suggesting that positive selection has been operating predominantly within the coding sequence. This was further supported by a greater than 2-fold reduction in nucleotide diversity in the coding region relative to the flanking sequence (Table 4.3). Further mutational constraints were indicated within the coding region by a 3:1 replacement to synonymous substitution ratio.

The *hina* gene region also showed significant test results using the McDonald Krietman test. However, in contrast to *hinb-1* these results were only significant using silent sites instead of synonymous sites (Table 4.7). This suggests that selective constraints have been operating outside of the coding region. This was further supported by the fact that both the 5' and 3' flanking regions contained a greater than 2-fold reduction in diversity levels relative to the flanking regions of the other genes analyzed (Tables 4.1-4.5). A reduction in the diversity level in the 5' flanking region could be the result of selective constraints acting on the regulatory components of the gene. A relatively high level of sequence homology (78%) has been demonstrated between the first 400 bp upstream of the *pina* and *pinb* start codons with a sharp decline (55%) within the next immediate

400 bp upstream (Lillemo *et al.*, 2002). In addition, several putative regulatory units, such as TATA and CAAT boxes, a transcription start site, and two endosperm specific boxes, were shown to be conserved between the wheat and barley homologs of *pinb* and *hinb* and no significant homology extended further than 544 bp upstream of the respective start codons (Darlington *et al.*, 2001). Furthermore, all regulatory units necessary for tissue specific expression were found to be located within the first 400 bp upstream of the start codon in *pinb* (Digeon *et al.*, 1999). These results indicate that the 503 bp of 5' flanking sequence analyzed here are likely to contain the majority if not all the necessary 5' components for regulation and expression of *hina*. The reduction of diversity in the 3' translated region could also be a result of selective constraints on translational control elements within the region (reviewed in Kuersten *et al.*, 2003; Mazumder *et al.*, 2003).

Although both the HKA and McDonald Krietman tests failed to detect any deviation from the neutral model at either the *GSP* or *hinb-2* gene regions; both regions demonstrated significant negative Tajima's D values. This indication of an excess of rare sequence variants was found across the entire *GSP* gene region but was restricted to the coding portion of *hinb-2*. Significant negative Tajima's D values are consistent with the recovery stage after the completion of a population bottleneck and recent demographic expansion. However, both evolutionary events should have an equal effect on all nucleotide sites and gene regions. An excess of rare variants can also be indicative of a selective sweep. In contrast to a population bottleneck, selective sweeps only affect diversity patterns at loci linked to the advantageous alleles. It is possible that selective constraints at these two gene regions are relatively weak compared to those seen at *hinb-1* and *hina*.

Sequence diversity results in wild barley at the *adh1* and *adh2* loci also demonstrated contrasting gene histories despite close proximity and tight linkage (Cummings *et al.*, 1998; Lin *et al.*, 2002). Although both loci were shown to be marked with an excess of

rare sequence variants, several key differences were observed. At the *adh2* locus, these variants showed no frequency restriction or bias to gene region or type of sequence site. In contrast, substantial bias was noted at the *adh1* locus for singleton replacement substitutions. Furthermore, a 1.6-fold increase of nucleotide diversity was found at *adh2* relative to *adh1*. Diversity levels were higher still at the *adh3* locus where the pattern of diversity sharply contrasted to both of the other paralogs. Here diversity was arranged into two major haplotype groups marked by geographical substructure (Lin *et al.*, 2001). Examples of heterogeneous patterns of nucleotide diversity in wild barley are not limited to the alcohol dehydrogenase gene family. A study including six additional nuclear genes revealed patterns of nucleotide variation which demonstrated a range of diversity levels from $\theta = 0.00115$ at the *Pepc* locus to $\theta = 0.01542$ at *Adh3* (Morrell *et al.*, 2003). Furthermore, different individual genes were representative of the three different diversity levels (low, moderate, and high) and demonstrated both the presence and absence of underlying population substructure (Morrell *et al.*, 2003).

4.3.3 Patterns of Nucleotide Diversity and Signatures of Selection within Hordeum vulgare

The general pattern of diversity observed within the wild sample was contrasted to that in the cultivated material. Low to moderate levels of nucleotide diversity were found ranging from $\theta = 0.00242$ to 0.00521 at the *GSP* and *PG2* gene regions, respectively (Tables 4.1-4.5). This diversity was arranged into a small number of haplotypes (5 to 11) demonstrating a range of frequencies (Figures 4.1, 4.3, 4.5, 4.6, and 4.9). No more than 2 recombination events were detected within any gene region and no recombination events were observed within *GSP*. Several lines of evidence were consistent with a domestication bottleneck:

1) θ_{silent} indicated a decrease in diversity levels relative to the wild sample in all five gene regions

2) the cultivated material demonstrated higher Tajima's D values at 3 of the 5 gene regions

3) diversity within the cultivated material was marked by a notable reduction in haplotype number

4) haplotypes within the cultivated material were predominately at intermediate frequencies

5) a limited number of protein variants were carried forward to the cultivated material for all genes

The landrace material demonstrated moderate levels of genetic diversity (θ = 0.00412 to 0.00622) organized into a small number of haplotypes (5 to 7) found at low to intermediate frequencies (Tables 4.1-4.5 and Figures 4.1, 4.3, 4.5, 4.6, and 4.9). Although these patterns of diversity were intermediate position between the cultivated and wild samples, the limited sample size restricted to material collected from the Fertile Crescent, make it difficult to access whether the patterns of diversity within landrace material are more similar to that in their cultivated or wild ancestors.

Of the five different gene regions analyzed in the cultivated material, only one revealed any evidence of deviation from the neutral model of evolution. *hinb-1* showed significant HKA test results when synonymous sites were considered suggesting that selective constraints have been operating predominantly within the coding sequence (Table 4.8). This was further supported by a greater than 2-fold reduction in nucleotide diversity in the coding region relative to the flanking sequence (Table 4.3). Further mutational constraints were indicated within the coding region by the complete absence of synonymous substitutions despite the presence of 3 replacement substitutions. Although, less pronounced, this pattern was comparable to the one observed within the wild germplasm. Similar patterns between the two sample sets were also observed at the *hina* gene region where diversity levels were elevated in the coding sequence relative to the

flanking regions (Table 4.2). As seen for *hinb-1*, this pattern was notably less prominent in the cultivated germplasm.

Several aspects must be considered when interpreting the different patterns of nucleotide diversity within the cultivated material. As previously mentioned, all five gene regions appeared to have sustained a loss of genetic diversity as a result of domestication. Likewise, human intervention arising from the processes of selection and domestication can have a notable impact on the magnitude and pattern of LD, ultimately increasing the length of chromosomal segments sharing similar evolutionary histories. An 8-fold decrease in the minimal number of recombination events was detected in the cultivated sample relative to the wild sample. Furthermore, a notable increase in the extent of high LD was observed across the *Ha* locus in the cultivated material (see Chapter 5). It is, therefore, possible that the presence of strong purifying selection at one locus could have potentially masked the effect on the presence of weaker selective constraints at nearby linked loci as a consequence of hitchhiking or linkage drag. The presence of similar levels of nucleotide diversity at the probable site of selection (*hinb-1*) and the immediate 5' and 3' gene regions (*hinb-2* and *hina* respectively) are consistent with this model.

4.3.4 Relationship of Putative Function to Signatures of Selection

Two putative gene functions have been attributed to the hordoindoline genes: grain texture and defense response. Their involvement in determining grain texture is implied by unbreakable association with the hardness QTL, perfect correlation of individual haplotypes of the wheat orthologs with grain texture, and successful puroindoline transformation studies (Rouves *et al.*, 1996; Jolly *et al.*, 1996; Sourdille *et al.*, 1996; Giroux *et al.*, 1997; Giroux *et al.*, 1998; Beecher *et al.*, 2001; Morris *et al.*, 2001; Krishnamurthy *et al.*, 2001b; Beecher *et al.*, 2002a). Their role in microbial defense response was first suggested by shared structural characteristics with thionins, α-amylase inhibitors, and non-specific lipid transfer proteins (ns-LTPs). *In vitro* assays

demonstrated that puroindolines alleles could cause significant inhibition of fungal growth even at low dosages (Dubreil *et al.*, 1998). In addition, rice plants containing puroindoline transgenes exhibited 53.5% and 22.3% increases in resistance to rice blast (*Magnaporthe grisea*) and sheath blight (*Rhizoctonia solani*), respectively (Krishnamurthy *et al.*, 2001a).

On the assumption that both of these functions are indeed pleiotropic effects of the same genes, it is reasonable to speculate that the evidence of selection within the wild germplasm is related to a role in microbial defense. Interest in grain texture is founded upon its role in directing the end product quality of cereal crops. Therefore, this function does not appear to contribute to a selective advantage to the species. Despite numerous studies investigating the correlation of puroindoline alleles with grain texture, equivalent tests have not been performed to establish a correlation of specific alleles with different degrees of microbial resistance. It is, therefore, possible to speculate that although certain alleles may be advantageous for defense responses in wild habitats, a completely different set of alleles may provide the desired end product effect in cultivated material. In this manner, selective constraints for alleles of high frequency within the wild material could have been shifted to rare alleles carried forward in early domestication. Although evidence of barley subsistence farming can be dated as far back as 17,000 BC (Kislev *et al.*, 1992), fermentation of barley grain for alcohol production probably did not arise until 3,000 to 5,000 BC (Edney, 1996). Furthermore, the introduction of malting quality as a key trait of interest within breeding programs did not arise until the late 1950's with the development of micro-malting techniques and small-scalc malting tests in the 1970's (Whitmore *et al.*, 1957). Therefore, it is likely that sufficient time has not passed for the detection of significant levels of selection among genes related to malting quality.

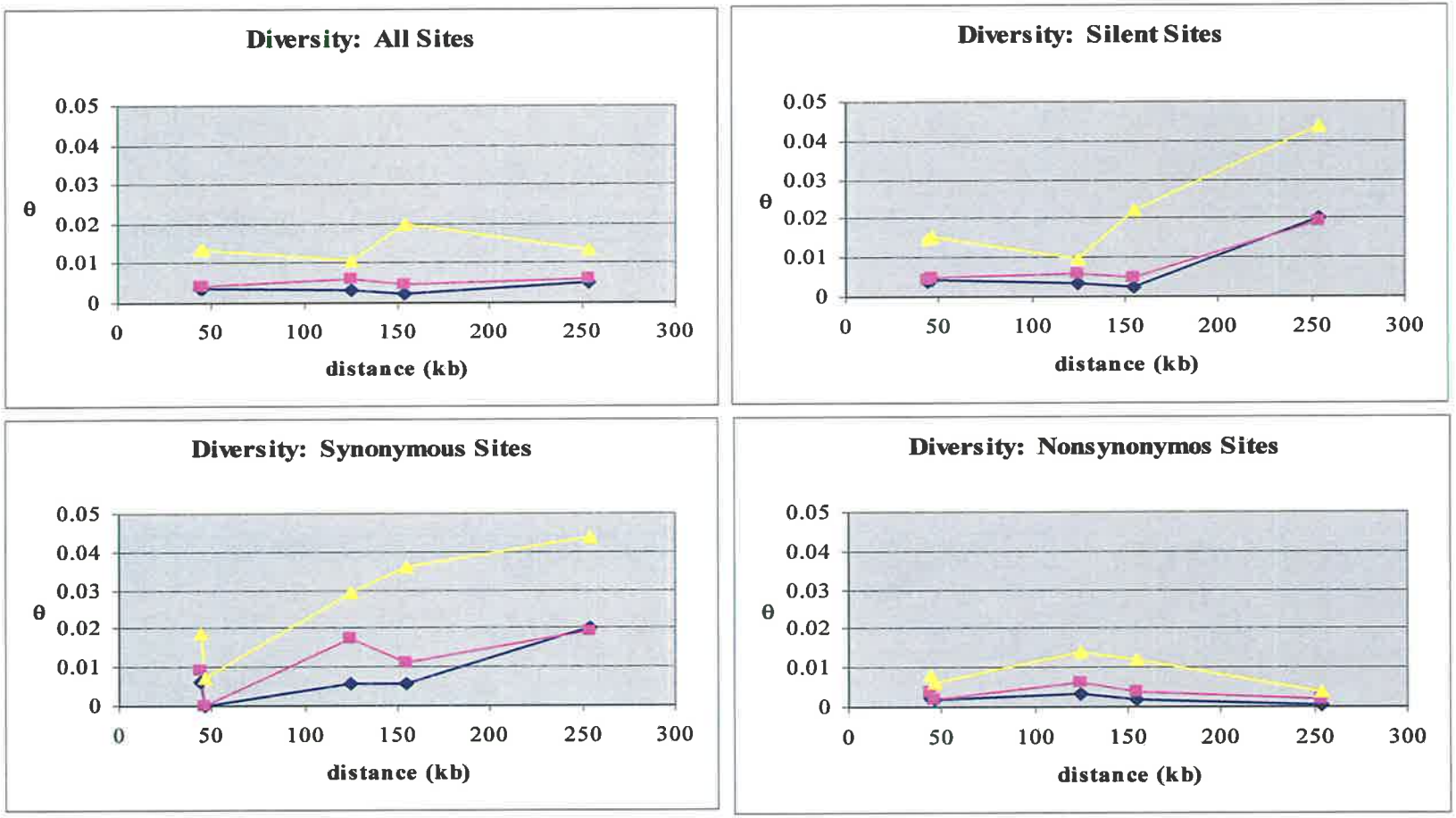


Figure 4.11. Plots of the levels of diversity for each gene region with respect to their physical location along the sequenced region of barley. Results for all three sample sets are shown: cultivated (blue), landrace (pink), and wild (yellow). Observations exemplify the utility of partitioning sequence data to acquire a more comprehensive understanding of the evolutionary history of the individual genes and the genomic region as a whole.

4.3.5 Conclusions

The utilization of high-throughput DNA sequencing allowed high-resolution and detailed determination of the underlying genetic diversity present at the hardness locus. The ability to accurately partition sequence information allowed direct comparison between loci, regions of the gene, and sequence sites for an understanding of gene history (summarized in Figure 4.11). Observations at the *Ha* locus indicated the presence of unique variation within the wild germplasm suggesting this material could be a potential resource for the introgression of novel alleles into breeding programs. Furthermore, contrasting patterns of genetic diversity within the wild and cultivated material could aid in the different stages of association studies. Despite both species maintaining similar breeding systems, a substantially higher rate of recombination in the wild material allowed more freedom for the independent evolution of proximal genes. This suggests that the extent of LD within the wild population will also be reduced compared to the cultivated germplasm. Such differences suggest that association studies could be designed around specific germplasm appropriate for either global or fine-scale mapping. In order to better evaluate the feasibility of this approach, a more direct analysis of the level of LD within and between the gene regions was undertaken. The results are presented in Chapter 5.

CHAPTER 5: INVESTIGATION OF THE EXTENT AND MAGNITUDE OF LOCAL LINKAGE DISEQUILIBRIUM ACROSS THE REGION HARBORING THE *HA* LOCUS IN BARLEY

5.1 Introduction

Determining the pattern and magnitude of linkage disequilibrium (LD) is an important step towards understanding population history, ascertaining regional levels of recombination, and designing effective association mapping strategies. LD studies have dominated the recent human genetics literature providing invaluable insight into LD structure and maintenance in human populations (reviewed in Goldstein, 2001; Gabriel *et al.*, 2002). The human genome is organized into large consecutive haplotype blocks (10-100 kb) interrupted by short regions with limited haplotype structure (1-7 kb; Taillon-Miller *et al.*, 2000; Jeffreys *et al.*, 2001; Rioux *et al.*, 2001; Daly *et al.*, 2001). Although recombination is known to occur within individual haplotype blocks, these events are relatively infrequent compared with the frequency of recombination within the hotspots defining block boundaries (Jeffreys *et al.*, 2001; Daly *et al.*, 2001; Reich *et al.*, 2001). LD extends across the entire length of each block. However, the magnitude of LD varies among blocks depending on their gene histories (Reich *et al.*, 2001). In general, haplotype block borders are consistent across populations and the same common haplotypes (3-5), albeit at different frequencies, exist in the majority of human populations. However, the extent of LD within these blocks varies among populations predominantly as a result of the breakdown of larger blocks into smaller blocks due to differing evolutionary histories (Reich *et al.*, 2001).

Two major conclusions were drawn from these results that influenced the future direction of human LD mapping approaches. The first was a switch at the genome-wide level from pairwise comparisons of individual markers to the application of haplotype analysis (Johnson *et al.*, 2001). A major disadvantage of pairwise comparisons between individual markers is that a negative association with one SNP cannot exclude the possibility of a

positive association with other near-by polymorphisms. Likewise, a given SNP demonstrating a positive association can not be considered a definitive causative mutation. In addition, by creating a SNP map based solely on the physical spacing of markers, the underlying haplotype diversity may not be fully described and important associations may go undetected. Although haplotype blocks will not allow the discovery of precise causative mutations, it will allow researchers to rapidly focus on well defined chromosomal regions. The use of haplotypes reduces the obscurity created by both the age and history of the individual markers and eliminates the short-comings of statistical approaches involving the pairwise comparisons of single polymorphisms. The second significant outcome from the human LD studies was the implication that populations with contrasting evolutionary histories could be utilized in a two-tiered approach to ultimately accomplish both genome-wide and fine-scale mapping. Populations demonstrating high levels of LD extending across large haplotype blocks could be used to rapidly define the specific chromosomal regions of interest. This would then be complemented by studies employing populations more suited to fine-scale mapping to locate individual candidate genes and causative mutations.

Despite the possibility that LD mapping could be a revolutionary tool in association studies, map-based cloning techniques, and the development of markers for selection in breeding, there have been relatively few reports on the underlying nature of LD in plant systems. The available studies have predominantly focused on the outbreeding crop species maize and the inbreeding model system *Arabidopsis thaliana* (reviewed in Gaut *et al.*, 2003; Flint-Garcia *et al.*, 2003) revealing a greater than 100-fold difference in the extent of LD between these two species. One of the main factors contributing to this vast difference can be attributed to alternate breeding systems. Although LD will naturally decay with distance, this process occurs at a considerably slower rate in inbreeding systems because effective recombination is reduced. Some of the world's most important crops such as rice, soybean, barley, and wheat are inbreeders. The impact of inbreeding

on the magnitude and pattern of LD in crop plants will also be influenced by human intervention through the processes of selection and domestication. Furthermore, the haploid genome size of *Arabidopsis* (125 Mb) is fifty times smaller than that of wheat and barley (5,300 Mb), mainly due to the recent amplification of repetitive DNA. These two factors suggest that studies of LD in *Arabidopsis* may not provide a comprehensive picture of the patterns and magnitude of LD in crop species with large genomes and a history of strong selective breeding.

This chapter presents the first study of localized LD in an inbreeding crop species. The evaluation of the extent of LD across the 212 kb sequenced region of barley harboring candidate genes for grain texture provided insight into the effect of genome structure and organization on LD maintenance and decay. In addition, the analysis of LD across independent gene regions provided supporting evidence for the presence of selection among genes observed in the patterns of nucleotide diversity reported in Chapter 4. Furthermore, the comparison of LD across different populations with contrasting histories allowed the significance of sampling to be quantified as a factor influencing the direction of future LD mapping approaches.

5.2 Measures of Linkage Disequilibrium

Several review articles have discussed the strengths and weaknesses of the various statistical methods for measuring LD (Lewontin, 1988; Devlin *et al.*, 1995; Jorde, 2000). The basic component of these methods is the frequency difference between the observed and expected (assuming independence of alleles) haplotypes:

$$D = \pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB} = 0$$

where π is the frequency of the different gametal types (haplotypes) between two bialellic loci A and B. In the complete absence of association D = 0. Unfortunately, the dependency of D upon allele frequencies excludes comparison with other loci and organisms. This is because the presence of low frequency alleles can dramatically skew

overall D values. Consequently, various measurements have been developed in an attempt to normalize D.

The two most commonly used measures of LD reported in the literature are D′ and r^2. D′ is calculated as a proportion of the maximum possible value of D based on the observed allele frequencies (Lewontin, 1964):

$$D' = D/\min(\pi_A\pi_b, \pi_a\pi_B) \text{ for } D < 0$$

$$D' = D/\max(\pi_A\pi_B, \pi_a\pi_b) \text{ for } D > 0$$

By taking the absolute value of D′, one can gain the desired effect of values ranging from $|D'| = 0$ (random inheritance) and $|D'| = 1$ (complete association). However, the deviation from complete association will only be recognized if all possible haplotypes are observed in the sample as evidence of a past recombination event. Therefore, D′ values will often tend to overestimate the level of association when relatively small sample sizes are employed.

r (also seen as Δ in the literature) is calculated as D scaled by the product of the allele frequencies at the two loci (Hill *et al.*, 1968):

$$r = D/\sqrt{(\pi_A\pi_a\pi_B\pi_b)}$$

As positive and negative values are based on arbitrary allele designation, the squared value is usually reported. Although D′ does a better job adjusting for allele frequencies and, therefore, is less affected by mutational history, r^2 is more predictable for small sample sizes making it the generally favored statistic. For this reason and to provide consistency with other plant LD literature, all values of LD reported here were measured with r^2.

5.3 Results

5.3.1 Patterns of LD within Candidate Gene Regions

The four gene regions described in Chapter 4, namely *hinb*, *hina*, *GSP*, and *PG2*, were analyzed to determine the level of LD (r^2) between informative sites (rare allele with f > 0.1). Plots of LD as a function of distance (bp) are shown for the three sample sets; cultivated, landrace, and *Hordeum spontaneum* (wild) in Figures 5.1 to 5.3. High levels of association extended across the entire 3373 bp *hinb* gene region containing both *hinb-1* and *hinb-2* in the cultivated sample with relatively few pairwise comparisons demonstrating low association. A scarcity of low association values was also observed at the *hina* region. However, in contrast to the *hinb* region, there appeared to be evidence of LD decay after 1000 bp. The plot of LD across the *GSP* gene revealed a high level of association across the entire 1805 bp region. However, this pattern also differed from that observed at the *hinb* region as association values demonstrating a range of magnitudes were found between pairs of sites at varying, intermediate distances within the gene region. An insufficient number of informative sites existed within the *PG2* region in the cultivated sample to accurately assess LD.

The overall extent of LD observed within gene regions in the landrace sample were similar to those observed within the cultivated material (Figure 5.2); high levels of association stretched across each gene region. However, a substantial number of pairwise comparisons within the *hinb* and *hina* gene regions gave moderate to low association values that were not detected in the cultivated material. Furthermore, there was a complete absence of intermediate association values in the *GSP* region. This bimodal distribution could be attributed to an elevated level of high association values as a result of small sample size. In sample sizes less than twenty, only singleton polymorphisms are excluded by the criterion of informative sites used in this study (rare allele with f > 0.1). Therefore, rare alleles occurring in the landrace sample (n = 15) are less likely to be corrected for and association values may be skewed toward deceptively high levels.

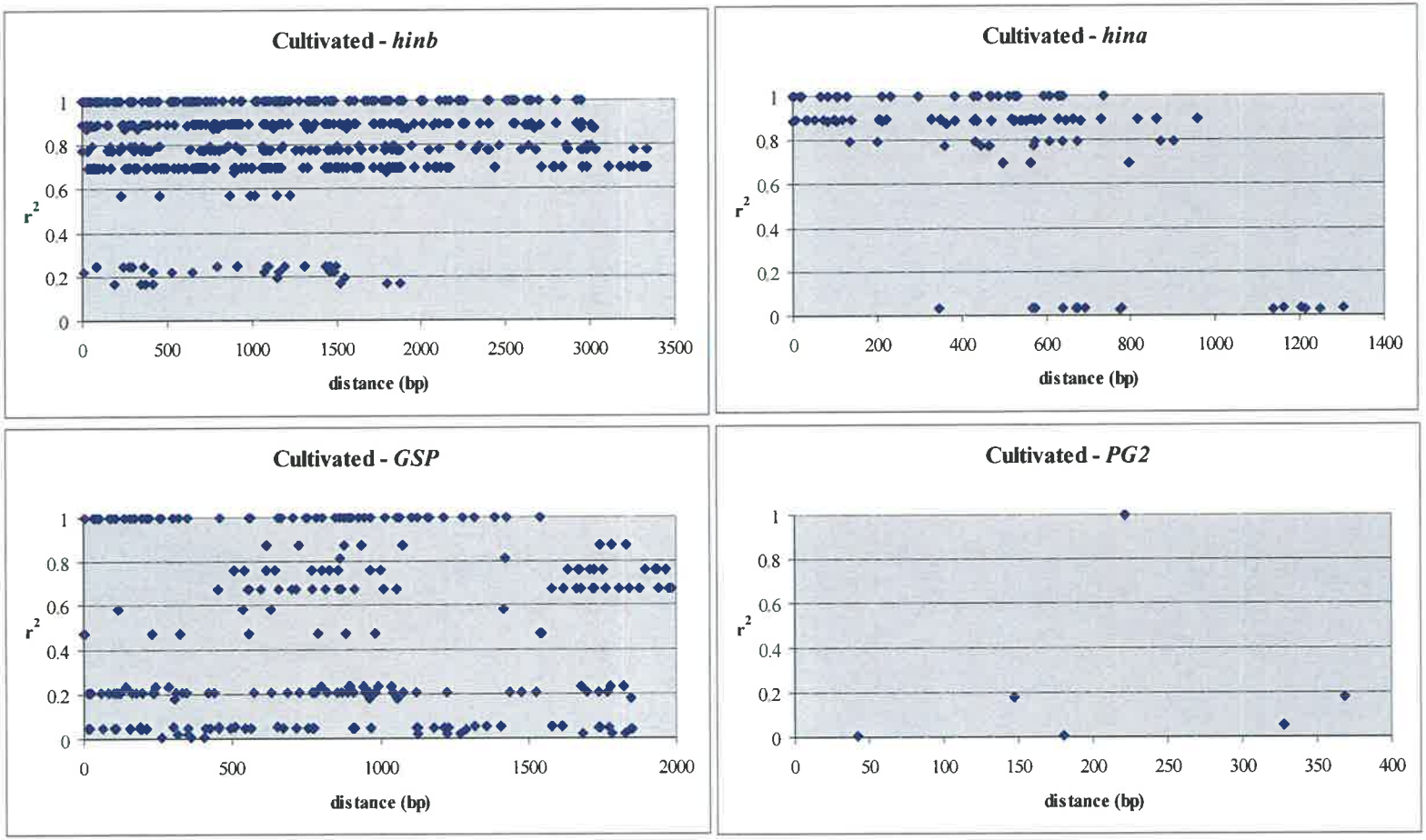


Figure 5.1. Plots of LD (r^2) as a function of distance (bp) between informative ($f > 0.1$) polymorphic sites in four different gene regions in the cultivated sample.

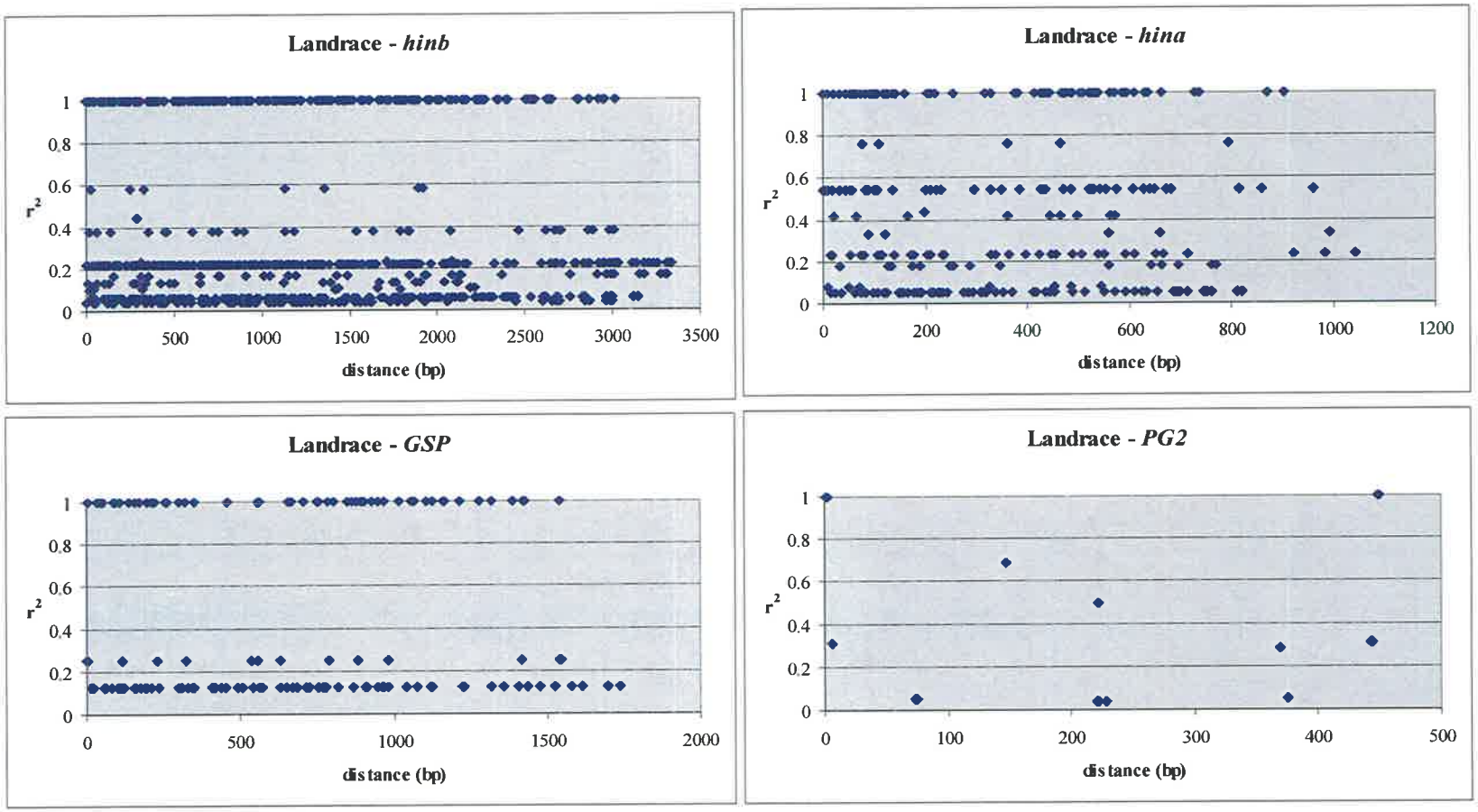


Figure 5.2. Plots of LD (r^2) as a function of distance (bp) between informative ($f > 0.1$) polymorphic sites in four different gene regions in the landrace sample.

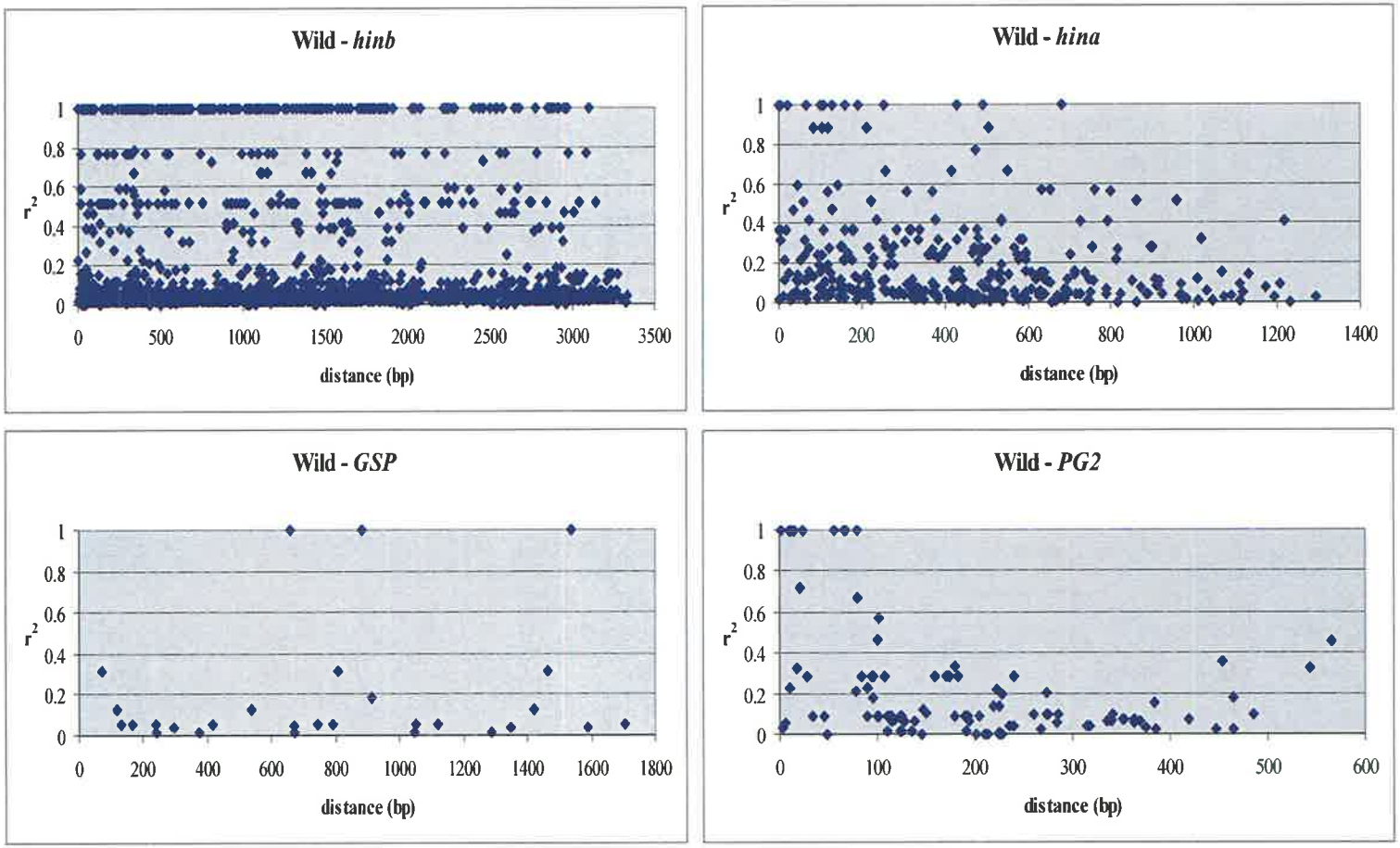


Figure 5.3. Plots of LD (r^2) as a function of distance (bp) between informative ($f > 0.1$) polymorphic sites in four different gene regions in the wild sample.

Although a few additional informative SNPs were available for association analysis in the *PG2* region, the paucity of these events still prevented accurate assessment of LD.

Similar to the pattern in the cultivated material, high levels of association extended across the entire *hinb-1* and *hinb-2* gene region in the wild sample (Figure 5.3). However, a substantial number of intermediate and low association values were also observed at a range of distances within the gene region. Likewise, a more distinct pattern of LD decay within the *hina* region was observed in the wild sample relative to the cultivated material with complete decaying to less than 0.2 by 1100 bp. An even greater rate of decay was observed in the *PG2* region where association values dropped below 0.2 by 400 bp. Although the number of informative SNPs within the *GSP* region was low, a general trend of low association values is apparent.

5.3.2 Patterns of LD across a Contiguous 212 kb Region

In order to extend the analysis of LD beyond within gene comparison, estimates of association were also determined between the different gene regions. The genes are nonuniformly distributed across the 212 kb contiguous sequence (Figure 5.4). Therefore, the distribution of distances between markers is broken into seven distinct groups:

1. within gene comparisons (1-4 kb)
2. between markers within *hina* and *GSP* (28-32 kb)
3. between markers within *hinb* and *hina* (77-83 kb)
4. between markers within *GSP* and *PG2* (98-101 kb)
5. between markers within *hinb* and *GSP* (107-113 kb)
6. between markers within *hina* and *PG2* (128-131 kb)
7. between markers within *hinb* and *PG2* (207-212 kb)

Estimates of LD between all pairs of informative sites are summarized in Figure 5.5.

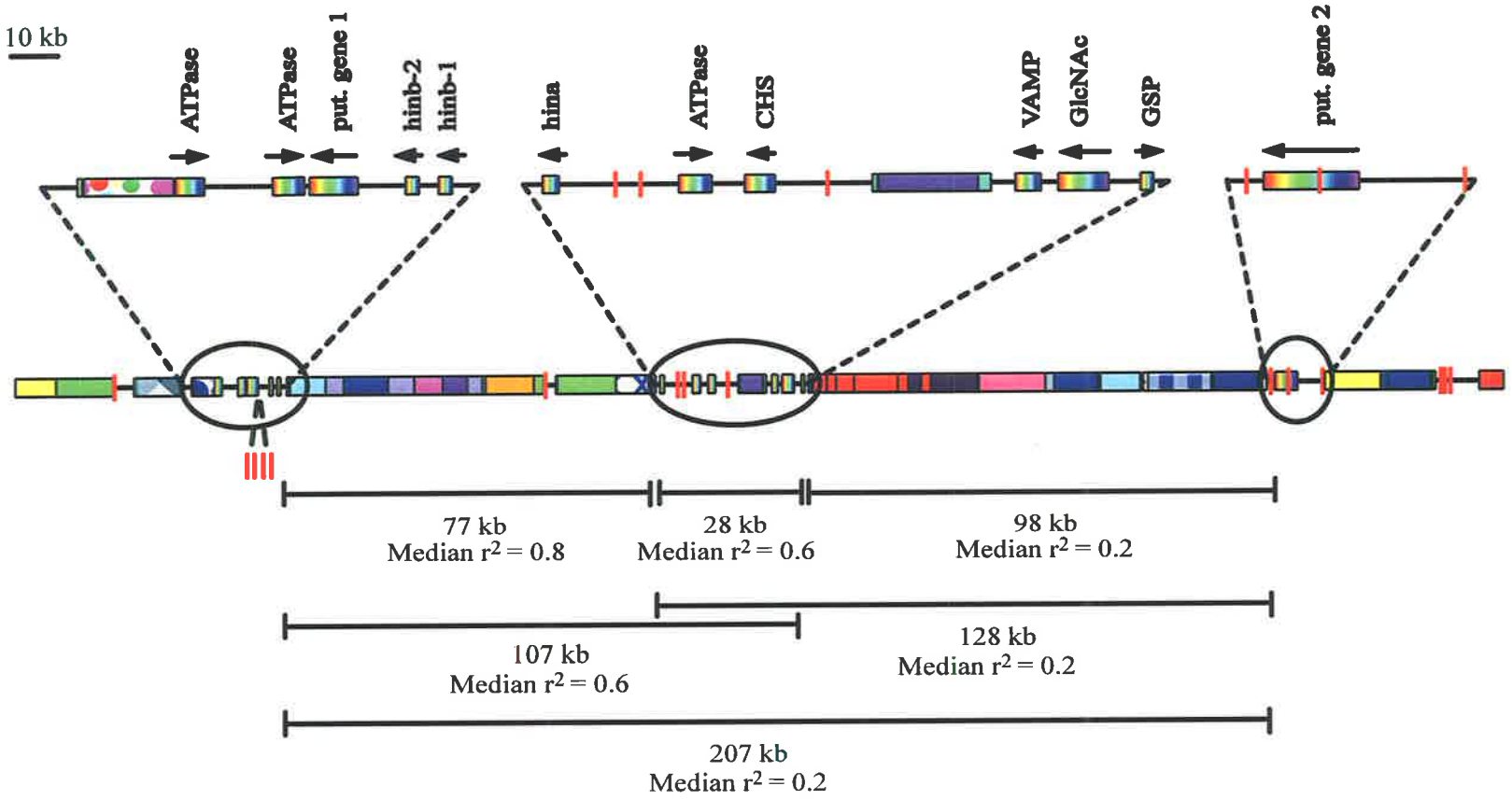


Figure 5.4. Graphical representation of the intergenic space between the different genes located in the sequence surrounding the *Ha* locus. Median LD values across these regions for the cultivated sample are indicated. Coding sequence is represented by rainbow boxes and arrows designate gene orientation. Repetitive sequence is coded similar to the legend in Figure 3.1

Although high levels of association were found to stretch across the entire region in the cultivated sample, the presence of numerous intermediate and low level association values obscured any obvious pattern of LD maintenance or decay (Figure 5.5A). In order to gain a better visualization of the relationship between association and physical distance, the data were summarized by calculating the median association value for each group and plotting it against the corresponding median distance (Figure 5.6A). Although a gradual decay of LD was observed with distance, significant median LD values extended across the entire 212 kb region and the median level for each group never decayed below 0.2.

To determine if the pattern and level of LD observed in the cultivated material is maintained in ancestral samples of barley, pairwise association was calculated between all informative sites within the sequence generated for the landrace and wild samples across the same gene regions. Although a few pairwise comparisons at distances greater than 100 kb demonstrated perfect association ($r^2 = 1$) in the landrace sample, it is likely that these were spurious associations detected as a consequence of the small sample size and inability to recognize and remove rare polymorphisms (Figure 5.5B). In general, LD decreased as a function of increasing distance in both ancestral samples. This pattern was particularly distinct in the wild material (Figure 5.5C). Significant median LD values at a level above 0.1 extended as far as 83 kb in the landrace sample and complete decay was seen by 98 kb (Figure 5.6B). In contrast, complete equilibrium was observed in the wild sample with no median values at a level above 0.1 (Figure 5.6C).

5.3.3 LD and its Relation to Genome Organization

Although the overall trend reveals a gradual decrease of LD with physical distance, there is also an undulating pattern in the levels of association among the different between gene comparison groups in all three sample sets studied (Figure 5.5). This pattern is even more

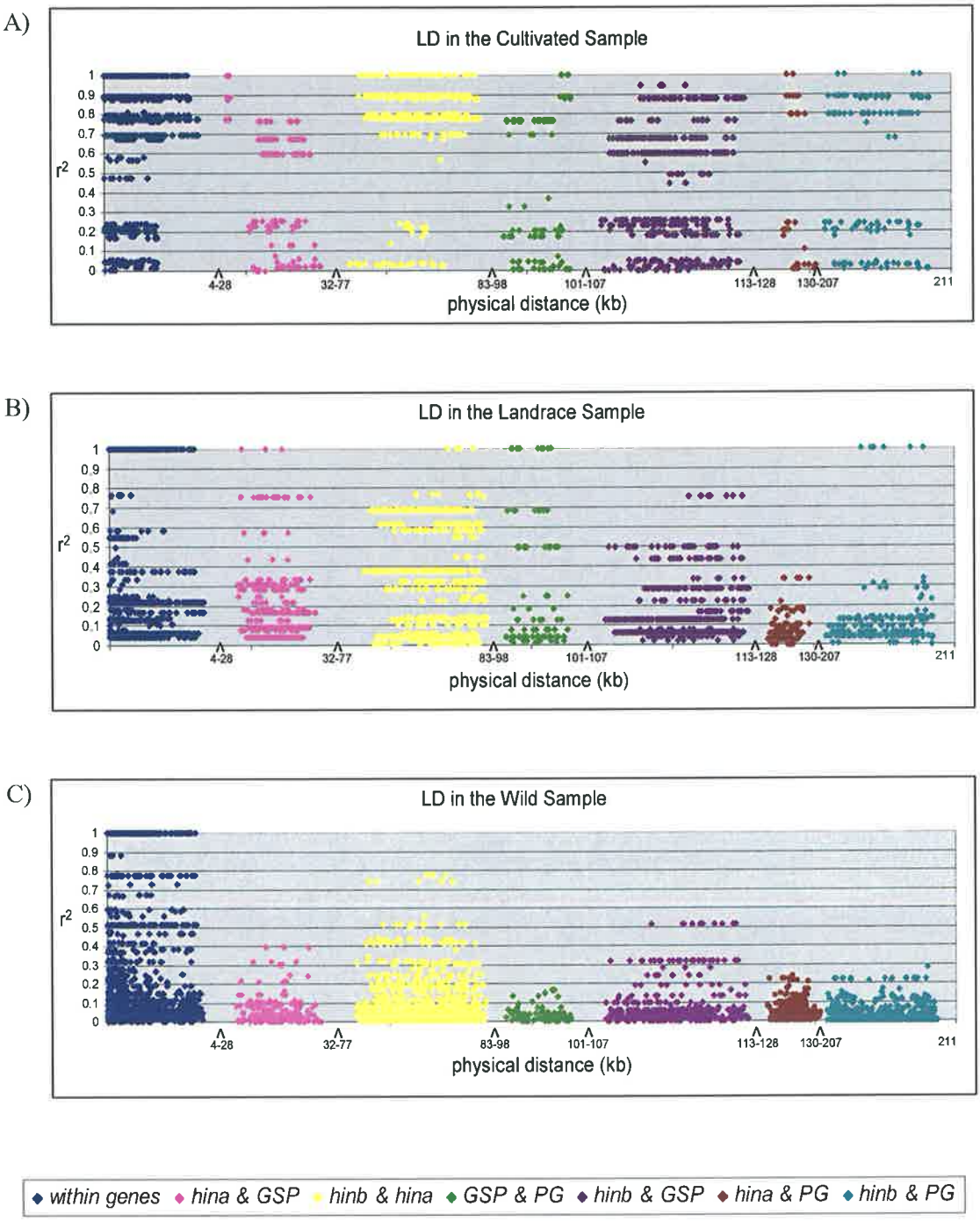


Figure 5.5. Plots of LD (r^2) as a function of distance (kb) for the A) cultivated, B) landrace, and C) wild samples.

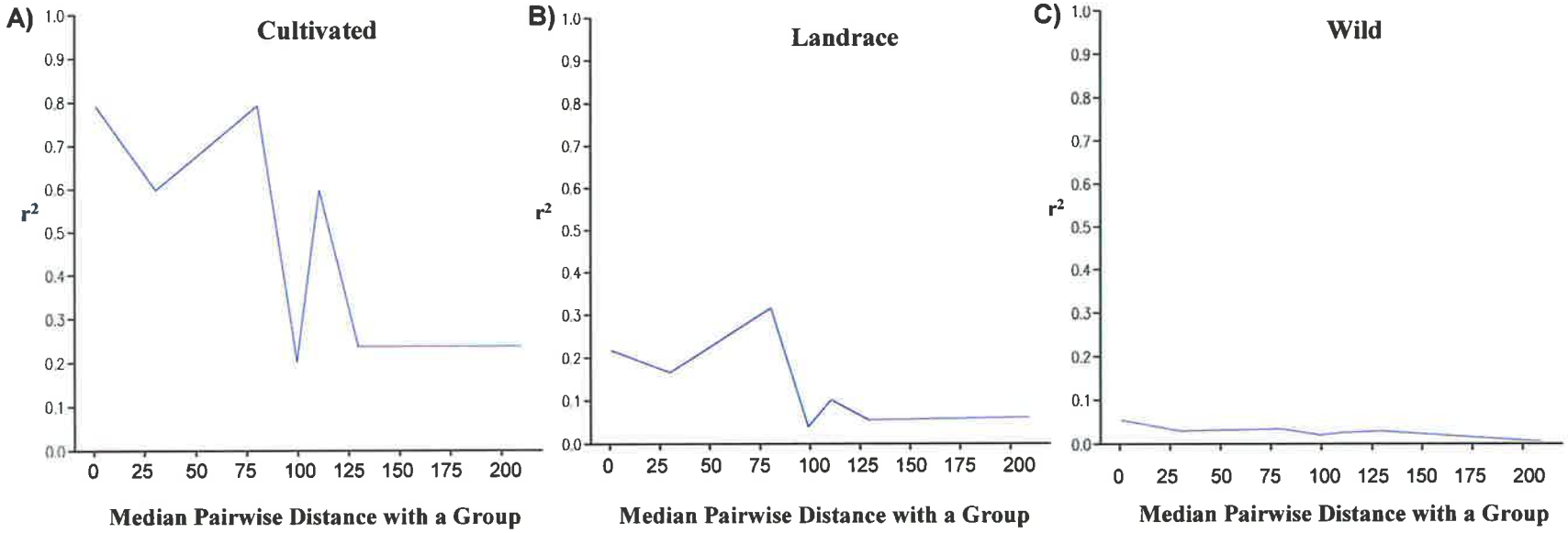


Figure 5.6. Plots of the median association value for each group of pairwise comparisons against the corresponding median distance. Groups are as follows: within gene comparisons (1-4 kb), comparisons between markers within *hina* and *GSP* (28-32 kb), comparisons between markers within *hinb* and *hina* (77-83 kb), comparisons between markers within *GSP* and *PG2* (98-101 kb), comparisons between markers within *hinb* and *GSP* (107-113 kb), comparisons between markers within *hina* and *PG2* (128-131 kb), and comparisons between markers within *hinb* and *PG2* (207-212 kb).

striking when the median LD value plots are considered (Figure 5.6). In an effort to explain this pattern, focus was turned to the genome composition and organization of the region. The sequence spanning the candidate grain texture genes includes several different patterns of genome organization that are typical of small grained cereals including both solitary genes and gene clusters separated by stretches of nested repetitive element insertions (Figure 5.4). The presence of either singular or nested transposable elements has previously been implicated as a mechanism for recombination suppression in several eukaryotic genomes and, therefore, could have an impact of local levels of LD (Charlesworth *et al.*, 1994; The Arabidopsis Genome Initiative, 2000; Yao *et al.*, 2002; Fu *et al.*, 2002b).

Two groups of pairwise comparisons involved sites from genes separated by large expanses of nested repetitive sequence: *hinb* and *hina* and *GSP* and *PG2*. The *hinb* and *hina* gene regions are separated by 77 kb, 93% of which is composed of transposable elements (Figure 5.4). This region occupies a 2.5-fold greater intergenic interval than that observed between *hina* and *GSP* which contains only one small ~5 kb retrotransposon (Figure 5.4). Nevertheless, association values between pairwise sites flanking the 77 kb transposable element cluster were higher in all three sample sets than between the 28 kb intergenic region composed primarily of low copy genic sequence. This result suggests recombination suppression in the former region (Figures 5.5 and 5.6).

The intergenic space between *GSP* and *PG2* was also primarily composed of nested transposable elements (96%) and spanned a region even larger than that observed between *hina* and *hinb* (96 kb; Figure 5.4). However, in contrast to the *hina* and *hinb* region, the *GSP* and *PG2* region demonstrated one of the lowest median association values of all groups considered. This suggests that the extensive regional expansion caused by element insertion has had negligible suppression on the recombination between these two genes (Figures 5.5 and 5.6). Therefore, the presence/absence of repetitive DNA

cannot solely account for the punctuated pattern of LD observed across the entire sequenced region.

5.3.4 Impact of Selection on Local Levels of LD

Another possible cause for the undulating pattern of LD is contrasting histories for the different gene regions. In all three sample sets, the *hinb-1* gene demonstrated the strongest evidence for selection (see Section 4.2.4). It is striking, therefore, that both prominent peaks present on the plots of median LD values correspond to pairwise comparisons involving sites located within the *hinb-1* gene region (Figure 5.6). Although median LD values are too low to observe this pattern in the wild germplasm, plots of the 95 percentile are consistent with this observation (data not shown). In contrast, no evidence for selection was observed for the *PG2* gene region regardless of sample set analyzed or test statistic employed (see Section 4.2.4). Median values of pairwise comparison groups involving sites located within the *PG2* gene demonstrate the lowest LD values observed.

5.4 Discussion

5.4.1 Contrasting Evolutionary Histories of Different Germplasm Samples as a Tool for Association Mapping Strategies

A significant "Catch-22" surrounding the utility of LD mapping studies lies in the conflicting requirements for a realistic genome-wide mapping approach consisting of a minimum set of genetic markers and the ability to perform high resolution mapping. Several studies have suggested that a strategy based on sampling individuals with contrasting population histories could be one route to achieving an initial low resolution analysis followed by candidate-gene association approaches in populations especially suited to high resolution mapping (Reich *et al.*, 2001; Nordborg *et al.*, 2002). The results obtained from this study present an encouraging outlook for accomplishing both global and fine-scale mapping. A limited genome scan based on polymorphic markers could be

adopted with cultivated germplasm, where LD extends beyond a few hundred kb, to narrow the interval in which genes controlling complex traits may reside. The ability to resolve candidate regions to a few hundred kb should greatly reduce the number of chromosome walking steps needed for positional cloning. This would be complemented by fine-scale mapping utilizing landrace and wild germplasm, where LD is more rapidly decayed relative to the cultivated material, to identify and assist in the validation of candidate genes.

Similar analyses have been reported in other plant systems. Initial studies in the outcrossing species maize indicate that the extent of LD decay differs from within a few hundred to 2000 base pairs depending on whether landraces or cultivated material is analyzed (Remington *et al.*, 2001; Tenaillon *et al.*, 2001; Ching *et al.*, 2002; Palaisa *et al.*, 2003). Although these levels are appropriate for fine-scale mapping, a marker density at the 100 bp level would be necessary for genome-wide mapping and the isolation of genes controlling complex traits. Small isolated populations of the inbreeding species *A. thaliana* also exhibit drastic differences in the extent of LD ranging from 1 cM to >50 cM (Nordborg *et al.*, 2002). These results provide a more optimistic scenario, relative to maize, for genome-wide mapping and would involve a marker density at the 100 kb scale. However, fine-scale mapping would be difficult due to the extended range of LD. Therefore, the observations in the region surrounding the *Ha* locus reported here represent the first example of different patterns of LD detected among samples of an inbreeding plant that are directly applicable to a two-tiered association mapping approach.

5.4.2 Contrasting Gene Histories Generate a Punctuated Pattern of LD

The plots of LD relative to physical distance did not show a smooth progression of decreasing association values with increasing distance (Figures 5.5 and 5.6). Instead, an undulating pattern of LD was observed with several regions of notable LD increase at intermediate distances (Figures 5.5 and 5.6). This observation is similar to that described

for humans where the "haplotype-block" model of LD has gained prominence (Jeffreys *et al.*, 2001). One plausible explanation for this punctuated pattern of LD is the presence of contrasting gene histories within the same local chromosomal region. Indeed, the different patterns of nucleotide diversity observed among the genes analyzed within the region of study demonstrated evidence of different degrees of selection (see Section 4.2.4). Furthermore, individual plots of association limited to comparisons within gene regions demonstrated different patterns of LD extent and magnitude (Figures 5.1-5.3). In all three sample sets, the region harboring the two *hinb* gene copies demonstrated the highest levels of association with negligible signs of LD decay (Figures 5.1-5.3). This is consistent with evidence that suggests that the *hinb-1* region was subjected to past directional selection (see Section 4.2.4). In contrast, LD was found to decay within only a few hundred base pairs at the *PG2* gene region in the wild germplasm (Figure 5.3). Limited LD maintenance within this gene region is supported by the lack of evidence to suggest any past selection (see Section 4.2.4). These local patterns of selection can help to explain the undulating pattern of LD across the entire region. The peaks of high LD levels corresponded perfectly to associations involving the putatively selected *hinb-1* gene region (Figures 5.5 and 5.6). Likewise, dips of low LD corresponded to associations involving the assumed neutral *PG2* gene region (Figures 5.5 and 5.6).

Evidence of changes in local levels of LD as a result of contrasting gene history have been previously reported at the different *adh* loci in *Hordeum spontaneum* (Lin *et al.*, 2001; Lin *et al.*, 2002). However, such examples are not limited to *Hordeum* species. At the region surrounding the *Arabidopsis FRI* gene, significant association levels were observed for pairwise-association comparisons up to ~250 kb apart (Nordborg *et al.*, 2002). The *FRI* gene has been implicated in flowering-time control and may have been subjected to local adaptative selection constraints accounting for the large extent of LD (Johanson *et al.*, 2000). In contrast, the extent of observed LD was reduced by as much as a factor of ten at both the *CLV2* and *RPS5 Arabidopsis* loci (Tian *et al.*, 2002; Shepard

et al., 2003). *CLV2* is involved in the regulation of the development of the shoot meristem and different *RPS5* alleles are believed to determine ability to recognize a *Pseudomonas syringae* strain for appropriate defense response (Shepard *et al.*, 2003). Both loci demonstrate evidence of balancing selection which is consistent with a lower levels of LD found in the region (Tian *et al.*, 2002; Shepard *et al.*, 2003). Breeding selection can also leave a mark in the patterns of LD within a plant system. Although the *Y1* phytoene synthase gene and the *PSY2* putative second phytoene synthase gene in maize are closely related, the two demonstrate drastically different nucleotide diversity levels with *Y1* displaying a greater than 10-fold increase in the extent of LD relative to *PSY2* (Palaisa *et al.*, 2003). This is predicted to be a result of both human selection for yellow endosperm, textured grain for increase carotenoid content and nutritional value and white endosperm texture for preferred end product color (Palaisa *et al.*, 2003).

5.4.3 The Role of Transposable Elements in Observed LD Patterns

Several studies in plant species indicate that recombination is predominately active in gene-rich chromosomal regions (Gill *et al.*, 1996a; Gill *et al.*, 1996b; Schnable *et al.*, 1998; Faris *et al.*, 2000; Kunzel *et al.*, 2000). Furthermore, single transposable elements and clusters of nested repetitive elements within plant species are either recombinationally inert or recombination suppressors (Yao *et al.*, 2002; Fu *et al.*, 2002b). As a consequence, successful fine-scale mapping of mutations based on association mapping may be largely dependent upon the genomic regions that surround candidate genes.

The intergenic region located between *hinb-1* and *hina* is composed predominantly of nested transposable elements (Figure 5.4). Although sites between these regions are separated by a distance greater than 77 kb, high levels of association extend across the intergenic region with a median value of 0.8 (Figures 5.4). Based on this observation alone it would be plausible to conclude that the presence of these elements could be

having an adverse effect on the recombination between these two gene regions. A similar region of clustered repetitive elements exists between *GSP* and *PG2*. The level of LD observed across this intergenic space drastically contrasts with that found between *hinb* and *hina*, showing minimal evidence of strong association and a median value of 0.2. This suggests that the presence of large segments of transposable sequence was not sufficient to suppress recombination between the genes characterized in this region of the barley genome.

Several points must be remembered in interpreting these results. The first is the assumption that all the lines represented in this study have the same or very similar genome organization and content. This assumption has already been proven dangerous in the comparison of the bronze locus in two maize inbred lines where the location and extent of element insertion as well as gene content were found to be highly variable (Fu *et al.*, 2002a). It cannot yet be determined whether this situation is the rule or the exception as this remains the only study where the lack of microcolinearity within a plant species has been reported.

The second consideration is that the results reported here cannot distinguish between recombination events occurring within the repetitive or low copy regions between the genes analyzed. Only exon 3 of *PG2* was included in the nucleotide diversity and association analysis. Therefore, the 4 kb of low copy sequence, including the later portion of the gene and the 3' flanking region, could harbor a recombination hotspot accounting for the low levels of association. This would be consistent with the observation that exon 3 of *PG2* contained the highest level of recombination observed within all regions analyzed. Similarly, although the transposable elements in the intergenic region between *hinb-1* and *hina* could be having an effect on the local recombination rate, it would be difficult to distinguish this effect from that of regional selection based on the information available here.

Whether or not transposable elements play a direct role in generating patterns of LD, it is interesting to keep such possibilities in mind when comparing LD results from *Arabidopsis* and cultivated barley. While the extent of LD in both inbreeding plant species appears to demonstrate a similar rate of LD decay (>200 kb), only 3 to 12 genes in barley have been found within the same range of about 25 genes in *Arabidopsis* (Panstruga *et al.*, 1998; Shirasu *et al.*, 2000; The Arabidopsis Genome Initiative, 2000; Dubcovsky *et al.*, 2001; Rostoks *et al.*, 2002; Wei *et al.*, 2002; Yan *et al.*, 2002; Gu *et al.*, 2003). This could prove to be an instance where the large portion of the barley genome inhabited by repetitive sequence provides an opportunity for association mapping by limiting the number of candidate genes and subsequent validation steps needed for positional cloning.

5.4.4 Conclusions

This work represents a detailed study into the levels and patterns of local LD within an inbreeding crop species. Although additional studies of both local and global LD need to be performed to get a more comprehensive assessment of the lower and upper limits to the extent of LD in barley, the results presented here indicate a positive future for the use of LD mapping in barley association genetics. The examination of different genepools with contrasting evolutionary histories has highlighted the potential for direct association mapping strategies based on the careful selection of appropriate germplasm. Furthermore, different gene histories within the 212 kb region resulted in a punctuated pattern of LD indicating that association analysis could also be a valuable tool for locating genes involved in local adaptation and the domestication process.

CHAPTER 6: CONCLUSIONS AND FUTURE DIRECTIONS

6.1 Context

The beginning of the 21st century is an exciting time for genome science and its application to plant breeding. Extensive resources including vast EST collections, large insert libraries, robust genetic and physical maps, structured mutant populations, and forward and reverse genetic platforms are becoming more readily available. Coupled with bioinformatics and automated high-throughput technologies, these tools provide the necessary framework for investigating genome organization, gene content, and the patterns and level of diversity present between and within species. In addition, these approaches have established a context for greater reproducibility across research laboratories at an international scale enabling direct comparisons of experimental results and increased confidence in the interpretation of data and overall collaboration. Furthermore, the application of such tools to include organisms of large genome size provides a means for the direct analysis of the species of interest without over reliance upon model systems. The use of model organisms, namely rice and *Arabidopsis*, has been invaluable for comparative genomics studies. Access to the fully sequenced genomes of these model plant species has enabled successful positional cloning of genes from large genome species (Yan *et al.*, 2003). However, rapid evolution among the grass genomes has resulted in the breakdown of microcolinearity at the genic level to a greater extent then previously revealed through genetic mapping (reviewed in Bennetzen *et al.*, 2003). This presents complications and limitations in the use of model species for identifying candidate genes located in regions that have undergone considerable evolution relative to the model organism since speciation. The scope of these limitations has been touched upon in the findings of this thesis.

6.2 Principal Findings

The primary objective of this study was to assess the feasibility of LD mapping for fine-scale association studies in an inbreeding crop species. Several complementary inter-

related approaches were employed for the realization of this goal. Chapter 3 describes the exploitation of the large insert Morex BAC library to generate a physical contig of the region harboring the *Ha* locus in barley. This represents the largest sequenced barley region encompassing a greater than 300 kb region sequenced to ten times coverage. The characterization of greater than 85% of the region added to the growing knowledge of the organization and content of the barley genome. Comparison with the colinear rice region revealed a complex evolutionary history arising since the divergence of the species involving numerous small chromosomal rearrangements, several gene duplications, and at least one translocation event. The breakdown of microcolinearity in this region exemplifies the limitations of rice as a model organism for the application of comparative genomics in association mapping and positional cloning approaches. These findings stress the importance of implementing genomic studies directly in the species of interest. Nevertheless, comparative genomics with both rice and *Arabidopsis* was proven to be a valuable resource for the determination of gene structure by providing an additional reference where ESTs were unavailable and gene prediction programs proved ambiguous and inconsistent.

Detailed characterization and nucleotide sequence of the region enabled nucleotide diversity studies. The sequence information presented in Chapter 4 represents, to my knowledge, the most comprehensive study of haplotype content and diversity in barley. Four different gene regions harboring five genes were analyzed. These represented a combined total of 7121 bp of aligned sequence excluding indels. With the exception of *PG2*, sequence information was obtained equally for 5' flanking, coding, and 3' flanking regions. The inclusion of flanking sequence data allowed comparisons between different regions of the genes and different types of nucleotide sites in order to obtain an understanding of gene history. Furthermore, this study extended the analysis of barley diversity by including both the cultivated and landrace *Hordeum vulgare* genepools to

complement previous studies on *H. spontaneum* (Cummings *et al.*, 1998; Lin *et al.*, 2001; Lin *et al.*, 2002; Morrell *et al.*, 2003).

Direct DNA sequencing provided a robust method for obtaining an accurate and comprehensive measure of genome diversity within and between barley genepools. As expected, a comparison of the observed diversity levels among the different samples revealed unique variation in the wild genepool confirming that this material could be a source for introgression of novel alleles into breeding programs. In addition, the different patterns of diversity provided insight into the effect different evolutionary forces have had on local genic regions. Although heterogeneous patterns of diversity have previously been observed in *H. spontaneum*, these results were based on independent genes spread throughout the barley genome (Morrell *et al.*, 2003). The results in Chapter 4 demonstrate that the levels and patterns of diversity continue to show a degree of heterogeneity even among genes in close proximity despite the predominance of self-fertilization. Furthermore, differences in the patterns of diversity are also observable in the cultivated material despite the homogenization of diversity levels. This homogenization is most likely attributable to selective constraints unique to the cultivation and domestication process.

The within gene polymorphic data (Chapter 4) combined with precise knowledge of gene position and its relation to genome content and organization (Chapter 3) established a platform for studying the levels and patterns of local linkage disequilibrium (LD) across the sequenced region. The results presented in Chapter 5 represent the first detailed study of the local (short-range) levels and patterns of LD in an inbreeding crop species. The cultivated genepool demonstrated substantially higher levels in both the magnitude and extent of LD relative to the *H. spontaneum* sample; the landrace material revealed an intermediary level of LD decay. These observations are consistent with the diversity data from Chapter 4 indicating that a higher rate of recombination in the wild material has

allowed more freedom for the independent evolution of neighboring genes. These findings suggest new opportunities for future association mapping studies in barley. Through the careful selection of appropriate germplasm, the barley community could direct both global and fine-scale association mapping strategies. Furthermore, the ability to detect different gene histories using both the diversity data and the LD studies indicates that revealing these underlying patterns could be a valuable tool for identifying candidate genes involved in local adaptation and the domestication process. Once candidate genes for traits of interest are both identified and validated, markers that demonstrated high association with such genes could be employed in breeding programs for the marker-assisted selection of desirable alleles.

6.3 Future Directions

There are two main research directions that would serve to complement and extend the work presented in this thesis. The first involves the validation of LD mapping for locating and identifying candidate genes. Although the extent and magnitude of LD observed within the different barley genepools indicated the potential for successful association approaches, until phenotypic information is applied to the marker data and direct gene-trait/marker-trait correlations are calculated, the actual utility of this approach remains uncertain. The puroindoline genes were suggested to have a role in grain texture through the unbroken correlation of haplotype and phenotype (Giroux *et al.*, 1997; Giroux *et al.*, 1998; Morris *et al.*, 2001). These initial findings were further supported by successful transformation of soft-textured alleles into both hard wheat and rice backgrounds (Krishnamurthy *et al.*, 2001b; Beecher *et al.*, 2002a). However, the implications of their barley counterparts, hordoindolines, in grain texture rests predominantly on the presumed shared function of orthologs and their association with a major endosperm-texture QTL (Beecher *et al.*, 2002b). Therefore, the application of LD mapping could provide considerable insight into the degree of association of the different components located in the region harboring the barley *Ha* locus by providing the

necessary increased resolution. The candidate genes could then be validated through numerous functional studies including the exploitation of structured mutation populations for reverse genetics and transformation assays.

The second focus for future research involves the extension of LD-studies to obtain a genome-wide understanding of the extent and magnitude of barley LD through an investigation of additional local regions of the barley genome. Several independent forces, due to contrasting population and gene histories, can have a profound effect on the patterns of LD observed within and between different populations of a species and among different regions of a given genome. Estimates of LD along chromosome 1 in maize indicate that LD can be affected by local recombination rates (Tenaillon *et al.*, 2002). Furthermore, a heterogeneous distribution of recombination has been observed along individual barley chromosomes in which recombination was primarily restricted to a few small regions of the genome (Kunzel *et al.*, 2000). These findings emphasize the need for more extensive LD studies on a whole genome level to gain a more comprehensive assessment of the lower and upper limits to the extent of LD in barley.

Chromosomal composition and location are not the only factors affecting local levels of LD. A marked contrast in the observed levels of LD has been noted in several *Arabidopsis* studies where different selective pressures, such as directional and balancing selection, are believe to have had a role in shaping the patterns of nucleotide diversity (Nordborg *et al.*, 2002; Tian *et al.*, 2002; Shepard *et al.*, 2003). Such patterns of selection are likely to be observed in wild genepools that have been subjected to different regional adaptive pressures and in the cultivated genepool which has been exposed to human intervention. Complementary studies of local LD would, therefore, add to the understanding of the best approaches for applying LD mapping to a wide range of genomic regions demonstrating contrasting gene histories.

The results of this study can also be used for directing future experimentation in other crop species. One of the primary findings of this study was the indication that appropriate selection of germplasm could aid in directing association studies for both global and fine-scale mapping. Although the different populations of the inbreeding plant species *A. thaliana* also demonstrated large differences in the extent of LD, none of the populations tested were amenable to fine-scale association studies (Nordborg *et al.*, 2002). Further research, therefore, should include additional species with high degrees of self-fertilization to acquire a better understanding of the effects of inbreeding on the extent of LD. Studies of this nature are vital to determine whether the use of different genepools for directing association studies is also applicable for other important inbreeding crop species, including wheat, rice, and sorghum.

Finally, the present study utilized sequence information to reveal the underlying patterns of nucleotide diversity within and between different genepools, to obtain a better perspective on the evolutionary forces acting upon the local region, and to evaluate the use of natural diversity for association mapping approaches. However, these are only a few scientific questions that can be answered using this approach. The haplotype information could be exploited to trace the origins of cultivated barley back to the cradle of domestication. In this manner the number and location of independent domestication events contributing to cultivated barley could be determined.

LITERATURE CITED

Aguadé, M., Miyashita, N., and Langley, C. H. (1989) Reduced Variation in the Yellow-Achaete-Scute Region in Natural Populations of *Drosophila melanogaster*. *Genetics* 122: 607-615.

Ahn, S. and Tanksley, S. D. (1993) Comparative Linkage Maps of the Rice and Maize Genomes. *Proceedings of the National Academy of Sciences of the United States of America* 90: 7980-7984.

Akhunov, E. D., Goodyear, A. W., Geng, S. *et al.* (2003) The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. *Genome Research* 13: 753-763.

Allard RW. 1999. *Principles of Plant Breeding.* second ed. New York: John Wiley & Sons, Inc.

Alleman, M. and Kermicle, J. L. (1993) Somatic Variegation and Germinal Mutability Reflect the Position of Transposable Element Dissociation Within the Maize R-Gene. *Genetics* 135: 189-203.

Allison, M. J. (1986) Relationships Between Milling Energy and Hot Water Extract Values of Malts from Some Modern Barleys and Their Parental Cultivars. *Journal of the Institute of Brewing* 92: 604-607.

Altschul, S. F., Madden, T. L., Schaffer, A. A. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25: 3389-3402.

Anandalakshmi, R., Pruss, G. J., Ge, X. *et al.* (1998) A viral suppressor of gene silencing in plants. *Proceedings of the National Academy of Sciences of the United States of America* 95: 13079-13084.

Arumuganathan, K. and Earle, E. D. (1991) Nuclear DNA content of some important plant species. *Plant Molecular Biology Reporter* 9: 211-215.

Bakhella, M., Hoseney, R. C., and Lookhart, G. L. (1990) Hardness of Moroccan Wheats. *Cereal Chemistry* 67: 246-250.

Barlow, K. K., Butterose, M. S., Simmonds, D. H. *et al.* (1973) The Nature of the Starch-Protein Interface in Wheat Endosperm. *Cereal Chemistry* 50: 443-454.

Barrett, J. H. (1992) Genetic-Mapping Based on Radiation Hybrid Data. *Genomics* 13: 95-103.

Baudry, E., Kerdelhue, C., Innan, H. *et al.* (2001) Species and recombination effects on DNA variability in the tomato genus. *Genetics* 158: 1725-1735.

Bauer, E., Weyen, J., Schiemann, A. *et al.* (1997) Molecular mapping of novel resistance genes against Barley Mild Mosaic Virus (BaMMV). *Theoretical and Applied Genetics* 95: 1263-1269.

Baurnert, M., Maycox, P. R., Navone, F. *et al.* (1989) Synaptobrevin: an integral membrane protein of 18,000 Daltons present in small synaptic vesicles of rat brain. *EMBO J* 8: 379-384.

Becker, J. and Heun, M. (1995) Barley Microsatellites - Allele Variation and Mapping. *Plant Molecular Biology* 27: 835-845.

Beecher, B., Bettge, A., Smidansky, E. *et al.* (2002a) Expression of wild-type pinB sequence in transgenic wheat complements a hard phenotype. *Theoretical and Applied Genetics* 105: 870-877.

Beecher, B., Bowman, J., Martin, J. M. *et al.* (2002b) Hordoindolines are associated with a major endosperm-texture QTL in Barley (*Hordeum vulgare*). *Genome* 45: 584-591.

Beecher, B., Smidansky, E. D., See, D. *et al.* (2001) Mapping and sequence analysis of barley hordoindolines. *Theoretical and Applied Genetics* 102: 833-840.

Begun, D. J. and Aquadro, C. F. (1992) Levels of Naturally Occurring DNA Polymorphism Correlate with Recombination Rates in *Drosophila melanogaster*. *Nature* 356: 519-520.

Bennett, M. D. and Leitch, I. J. (1995) Nuclear DNA Amounts in Angiosperms. *Annals of Botany* 76: 113-176.

Bennett, M. D. and Leitch, I. J. (1997) Nuclear DNA amounts in angiosperms - 583 new estimates. *Annals of Botany* 80: 169-196.

Bennett, M. D., Smith, J. B., and Heslop-Harrison, J. S. (1982) Nuclear DNA Amounts in Angiosperms. *Proceedings of the Royal Society of London Series B-Biological Sciences* 216: 179-199.

Bennetzen, J. L. (2000a) Comparative sequence analysis of plant nuclear genomes: Microcolinearity and its many exceptions. *Plant Cell* 12: 1021-1029.

Bennetzen, J. L. (2000b) Transposable element contributions to plant gene and genome evolution. *Plant Molecular Biology* 42: 251-269.

Bennetzen, J. L. and Kellogg, E. A. (1997) Do plants have a one-way ticket to genomic obesity? *Plant Cell* 9: 1509-1514.

Bennetzen, J. L. and Ma, J. (2003) The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis. *Current Opinion in Plant Biology* 6: 128-133.

Bennetzen, J. L. and Ramakrishna, W. (2002) Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. *Plant Molecular Biology* 48: 821-827.

Blochet, J. E., Chevalier, C., Forest, E. *et al.* (1993) Complete Amino-Acid-Sequence of Puroindoline, A New Basic and Cystine-Rich Protein with A Unique Tryptophan-Rich Domain, Isolated from Wheat Endosperm by Triton X-114 Phase Partitioning. *Febs Letters* 329: 336-340.

Botstein, D., White, R., Skolnick, M. *et al.* (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* 32: 314-331.

Braverman, J. M., Hudson, R. R., Kaplan, N. L. *et al.* (1995) The Hitchhiking Effect on the Site Frequency-Spectrum of DNA Polymorphisms. *Genetics* 140: 783-796.

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

Brown, A. H. D., Zohary, D, and Nevo, E. (1978) Outcrossing rates and heterozygosity in natural populations of *Hordeum spontaneum* Koch in Israel. *Heredity* 41: 49-62.

Brunner, S., Keller, B., and Feuillet, C. (2003) A large rearrangement involving genes and low-copy DNA interrupts the microcollinearity between rice and barley at the Rph7 locus. *Genetics* 164: 673-683.

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

Buckler, E. S., Thornsberry, J. M., and Kresovich, S. (2001) Molecular diversity, structure and domestication of grasses. *Genetical Research* 77: 213-218.

Bureau, T. E. and Wessler, S. R. (1992) Tourist - A Large Family of Small Inverted Repeat Elements Frequently Associated with Maize Genes. *Plant Cell* 4: 1283-1294.

Bureau, T. E. and Wessler, S. R. (1994a) Stowaway - A New Family of Inverted Repeat Elements Associated with the Genes of Both Monocotyledonous and Dicotyledonous Plants. *Plant Cell* 6: 907-916.

Bureau, T. E., White, S. E., and Wessler, S. R. (1994b) Transduction of a Cellular Gene by a Plant Retroelement. *Cell* 77: 479-480.

Chantret, N., Center, A., Sabot, F. *et al.* (2004) Sequencing of the *Triticum monococcum hardness* locus reveals good microcolinearity with rice. *Molecular & General Genetics* 271: 377-386.

Charlesworth, B., Morgan, M. T., and Charlesworth, D. (1993) The Effect of Deleterious Mutations on Neutral Molecular Variation. *Genetics* 134: 1289-1303.

Charlesworth, B., Sniegowski, P., and Stephan, W. (1994) The Evolutionary Dynamics of Repetitive DNA in Eukaryotes. *Nature* 371: 215-220.

Charrier, B., Foucher, F., Kondorosi, E. *et al.* (1999) Bigfoot: a new family of MITE elements characterized from the *Medicago* genus. *Plant Journal* 18: 431-441.

Chen, M. S., SanMiguel, P., and Bennetzen, J. L. (1998) Sequence organization and conservation in sh2/a1-homologous regions of sorghum and rice. *Genetics* 148: 435-443.

Chen, Y. A. and Scheller, R. H. (2001) SNARE-mediated membrane fusion. *Nature Reviews Molecular Cell Biology* 2: 98-106.

Ching, A., Caldwell, K. S., Jung, M. *et al.* (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *Bmc Genetics* 3: art-19.

Clark, L. G., Zhang, W. P., and Wendel, J. F. (1995) A Phylogeny of the Grass Family (Poaceae) Based on Ndhf - Sequence Data. *Systematic Botany* 20: 436-460.

Conway, D. J., Roper, C., Oduola, A. M. J. *et al.* (1999) High recombination rate in natural populations of *Plasmodium falciparum*. *Proceedings of the National Academy of Sciences of the United States of America* 96: 4506-4511.

Corder, E. H., Saunders, A. M., Risch, N. J. *et al.* (1994) Protective Effect of Apolipoprotein-E Type-2 Allele for Late Onset Alzheimer Disease. *Nature Genetics* 7: 180-184.

Crepet, W. L. and Feldman, G. D. (1991) The Earliest Remains of Grasses in the Fossil Record. *American Journal of Botany* 78: 1010-1014.

Cummings, M. P. and Clegg, M. T. (1998) Nucleotide sequence diversity at the alcohol dehydrogenase 1 locus in wild barley (*Hordeum vulgare* ssp. *spontaneum*): An evaluation of the background selection hypothesis. *Proceedings of the National Academy of Sciences of the United States of America* 95: 5637-5642.

Daly, M. J., Rioux, J. D., Schaffner, S. E. *et al.* (2001) High-resolution haplotype structure in the human genome. *Nature Genetics* 29: 229-232.

Darlington, H. F., Rouster, J., Hoffmann, L. *et al.* (2001) Identification and molecular characterisation of hordoindolines from barley grain. *Plant Molecular Biology* 47: 785-794.

Davis, M. P., Franckowiak, J. D., Konishi, T. *et al.* (1997) New and revised Barley Gentic Stock descriptions. *Barley Genetics Newsletter* 26: 22-516.

Dear, P. H. and Cook, P. R. (1993) Happy Mapping - Linkage Mapping Using A Physical Analog of Meiosis. *Nucleic Acids Research* 21: 13-20.

Depicker, A. and VanMontagu, M. (1997) Post-transcriptional gene silencing in plants. *Current Opinion in Cell Biology* 9: 373-382.

DeScenzo, R. A. and Wise, R. P. (1996) Variation in the ratio of physical to genetic distance in intervals adjacent to the Mla locus on barley chromosome 1H. *Molecular and General Genetics* 251: 472-482.

Devlin, B. and Risch, N. (1995) A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping. *Genomics* 29: 311-322.

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

Digeon, J. F., Guiderdoni, E., Alary, R. *et al.* (1999) Cloning of a wheat puroindoline gene promoter by IPCR and analysis of promoter regions required for tissue-specific expression in transgenic rice seeds. *Plant Molecular Biology* 39: 1101-1112.

Distelfeld, A., Uauy, C., Olmos, S. *et al.* (2004) Microcolinearity between a 2-cM region encompassing the grain protein content locus *Gpc-6B1* on wheat chromosome 6B and a 350-kb region on rice chromosome 2. *Functional and Integrative Genomics* 4: 59-66.

Dixon, R. A., Harrison, M. J., and Paiva, N. L. (1995a) The Isoflavonoid Phytoalexin Pathway - from Enzymes to Genes to Transcription Factors. *Physiologia Plantarum* 93: 385-392.

Dixon, R. A., Lamb, C. J., Masoud, S. *et al.* (1996) Metabolic engineering: Prospects for crop improvement through the genetic manipulation of phenylpropanoid biosynthesis and defense responses - A review. *Gene* 179: 61-71.

Dixon, R. A. and Paiva, N. L. (1995b) Stress-Induced Phenylpropanoid Metabolism. *Plant Cell* 7: 1085-1097.

Doebley, J., Stec, A., and Hubbard, L. (1997) The evolution of apical dominance in maize. *Nature* 386: 485-488.

Drouin, G. and Dover, G. A. (1987) A Plant Processed Pseudogene. *Nature* 328: 557-558.

Druka, A., Kudrna, D., Han, F. *et al.* (2000) Physical mapping of the barley stem rust resistance gene rpg4. *Molecular and General Genetics* 264: 283-290.

Dubcovsky, J., Ramakrishna, W., SanMiguel, P. J. *et al.* (2001) Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. *Plant Physiology* 125: 1342-1353.

Dubreil, L., Compoint, J. P., and Marion, D. (1997) Interaction of puroindolines with wheat flour polar lipids determines their foaming properties. *Journal of Agricultural and Food Chemistry* 45: 108-116.

Dubreil, L., Gaborit, T., Bouchet, B. *et al.* (1998) Spatial and temporal distribution of the major isoforms of puroindolines (puroindoline-a and puroindoline-b) and non specific lipid transfer protein (ns-LTPle(1)) of *Triticum aestivum* seeds. Relationships with their *in vitro* antifungal properties. *Plant Science* 138: 121-135.

Dunning, A. M., Durocher, F., Healey, C. S. *et al.* (2000) The extent of linkage disequilibrium in four populations with distinct demographic histories. *American Journal of Human Genetics* 67: 1544-1554.

Edney, M. J. (1996) Barley. In Henry RJ, Kettlewell PS (eds) Cereal Grain Quality. London: Chapman and Hall. 113-151.

Elrouby, N. and Bureau, T. E. (2001) A novel hybrid open reading frame formed by multiple cellular gene transductions by a plant long terminal repeat retroelement. *Journal of Biological Chemistry* 276: 41963-41968.

Enoki, H., Izawa, T., Kawahara, M. *et al.* (1999) Ac as a tool for the functional genomics of rice. *Plant Journal* 19: 605-613.

Erkkila, M. J. (1999) Intron III-specific markers for screening of beta-amylase alleles in barley cultivars. *Plant Molecular Biology Reporter* 17: 139-147.

Everson, E. H. and Schaller, C. W. (1955) The genetics of yield differences associated with awn garbing in the barley hybrid (Lion x Atlas 10) x Atlas. *Agronomy Journal* 47: 276-280.

Ewing, B. and Green, P. (1998a) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* 8: 186-194.

Ewing, B., Hillier, L., Wendl, M. C. *et al.* (1998b) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* 8: 175-185.

Eyre-Walker, A., Gaut, R. L., Hilton, H. *et al.* (1998) Investigation of the bottleneck leading to the domestication of maize. *Proceedings of the National Academy of Sciences of the United States of America* 95: 4441-4446.

Faris, J. D., Haen, K. M., and Gill, B. S. (2000) Saturation mapping of a gene-rich recombination hot spot region in wheat. *Genetics* 154: 823-835.

Feuillet, C. and Keller, B. (1999) High gene density is conserved at syntenic loci of small and large grass genomes. *Proceedings of the National Academy of Sciences of the United States of America* 96: 8265-8270.

Feuillet, C. and Keller, B. (2002) Comparative Genomics in the grass family: Molecular characterization of grass genome structure and evolution. *Annals of Botany* 89: 3-10.

Fischbeck, G. (2001) Contribution of Barley to Agriculture. In Slafer GA et al (eds) Barley Science. London: Food Products Press. 1-14.

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

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

Fu, H. H., Zheng, Z. W., and Dooner, H. K. (2002b) Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. *Proceedings of the National Academy of Sciences of the United States of America* 99: 1082-1087.

Gabriel, S. B., Schaffner, S. F., Nguyen, H. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science* 296: 2225-2229.

Gale, M. D. and Devos, K. M. (1998) Comparative genetics in the grasses. *Proceedings of the National Academy of Sciences of the United States of America* 95: 1971-1974.

Galili, S., Avivi, Y., Millet, E. *et al.* (2000) RFLP-based analysis of three RbcS subfamilies in diploid and polyploid species of wheat. *Molecular and General Genetics* 263: 674-680.

Gaut, B. S. and Long, A. D. (2003) The Lowdown on Linkage Disequilibrium. *Plant Cell* 15: 1502-1506.

Gautier, M. F., Aleman, M. E., Guirao, A. *et al.* (1994) *Triticum aestivum* Puroindolines, 2 Basic Cystine-Rich Seed Proteins - cDNA Sequence Analysis and Developmental Gene Expression. *Plant Molecular Biology* 25: 43-57.

Gill, K. S., Gill, B. S., Endo, T. R. *et al.* (1996a) Identification and high-density mapping of gene-rich regions in chromosome group 5 of wheat. *Genetics* 143: 1001-1012.

Gill, K. S., Gill, B. S., Endo, T. R. *et al.* (1996b) Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. *Genetics* 144: 1883-1891.

Giroux, M. J., Clancy, M., Baier, J. *et al.* (1994) *De Novo* Synthesis of an Intron by the Maize Transposable Element Dissociation. *Proceedings of the National Academy of Sciences of the United States of America* 91: 12150-12154.

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

Glenn, G. M. and Saunders, R. M. (1990) Physical and Structural Properties of Wheat Endosperm Associated with Grain Texture. *Cereal Chemistry* 67: 176-182.

Goff, S. A., Ricke, D., Lan, T. H. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). *Science* 296: 92-100.

Goldstein, D. B. (2001) Islands of linkage disequilibrium. *Nature Genetics* 29: 109-111.

Goloubinoff, P., Paabo, S., and Wilson, A. C. (1993) Evolution of Maize Inferred from Sequence Diversity of An Adh2 Gene Segment from Archaeological Specimens. *Proceedings of the National Academy of Sciences of the United States of America* 90: 1997-2001.

Gordon, D., Abajian, C., and Green, P. (1998) Consed: A graphical tool for sequence finishing. *Genome Research* 8: 195-202.

Goss, S. J. and Harris, H. (1975) New method for mapping genes in human chromosomes. *Nature* 255: 684-

Grandbastien, M. A. (1992) Retroelements in Higher Plants. *Trends in Genetics* 8: 103-108.

Graner, A. (1996) Molecular mapping of genes conferring disease resistance: the present state and future aspects. Scoles, G. J. and Rossnagel, B. G. Proceeding of the V International Oat Conference and the VII International Barley Genetics Symposium. 157-166. Saskatoon, University of Saskatchewan Extension Press.

Graner, A. and Bauer, E. (1993) Rflp Mapping of the Ym4 Virus-Resistance Gene in Barley. *Theoretical and Applied Genetics* 86: 689-693.

Graner, A., Jahoor, A., Schondelmaier, J. *et al.* (1991) Construction of An Rflp Map of Barley. *Theoretical and Applied Genetics* 83: 250-256.

Graner, A., Streng, S., Kellermann, A. *et al.* (1999) Molecular mapping and genetic fine-structure of the rym5 locus encoding resistance to different strains of the Barley Yellow Mosaic Virus Complex. *Theoretical and Applied Genetics* 98: 285-290.

Greenwell, P. and Schofield, J. D. (1986) A Starch Granule Protein Associated with Endosperm Softness in Wheat. *Cereal Chemistry* 63: 379-380.

Gu, Y. Q., Anderson, O. D., Londeore, C. F. *et al.* (2003) Structural organization of the barley D-hordein locus in comparison with orhtologous region of wheat genomes. *Genome* 46: 1084-1097.

Guyot, R, Yahiaoui, N, Feuillet, C. *et al.* (2004) In silico comparative analysis reveals a mosaic conservation of genes within a novel colinear region in wheat chromosome 1AS and rice chromosome 5S. *Functional and Integrative Genomics* 4: 47-58.

Hackett, C. A., Meyer, R. C., and Thomas, W. T. B. (2001) Multi-trait QTL mapping in barley using multivariate regression. *Genetical Research* 77: 95-106.

Hamrick JL, Godt MJW (1990) Allozyme diversity in plant species. In Brown AHD et al (eds) Plant Population Genetics, Breeding, and Genetic Resources. Sunderland, Massachusettes: Sinauer Associates Inc. 43-63.

Hamrick, J. L. and Godt, M. J. W. (1997) Allozyme diversity in cultivated crops. *Crop Science* 37: 26-30.

Han, F., Kilian, A., Chen, J. P. *et al.* (1999) Sequence analysis of a rice BAC covering the syntenous barley Rpg1 region. *Genome* 42: 1071-1076.

Han, F., Kleinhofs, A., Ullrich, S. E. *et al.* (1998) Synteny with rice: analysis of barley malting quality QTLs and rpg4 chromosome regions. *Genome* 41: 373-380.

Han, F., Romagosa, I., Ullrich, S. E. *et al.* (1997a) Molecular marker-assisted selection for malting quality traits in barley. *Molecular Breeding* 3: 427-437.

Han, F., Ullrich, S. E., Chirat, S. *et al.* (1995) Mapping of Beta-Glucan Content and Beta-Glucanase Activity Loci in Barley-Grain and Malt. *Theoretical and Applied Genetics* 91: 921-927.

Han, F., Ullrich, S. E., Kleinhofs, A. *et al.* (1997b) Fine structure mapping of the barley chromosome-1 centromere region containing malting-quality QTLs. *Theoretical and Applied Genetics* 95: 903-910.

Hastbacka, J., Delachapelle, A., Kaitila, I. *et al.* (1992) Linkage Disequilibrium Mapping in Isolated Founder Populations - Diastrophic Dysplasia in Finland. *Nature Genetics* 2: 204-211.

Hayes, P. M., Liu, B. H., Knapp, S. J. *et al.* (1993) Quantitative Trait Locus Effects and Environmental Interaction in A Sample of North-American Barley Germ Plasm. *Theoretical and Applied Genetics* 87: 392-401.

Hedrick, S. M., Cohen, D. I., Nielsen, E. A. *et al.* (1984) Isolation of Cdna Clones Encoding T-Cell-Specific Membrane- Associated Proteins. *Nature* 308: 149-153.

Helentjaris, T., Weber, D., and Wright, S. (1988) Identification of the Genomic Locations of Duplicate Nucleotide-Sequences in Maize by Analysis of Restriction Fragment Length Polymorphisms. *Genetics* 118: 353-363.

Heun, M., Kennedy, A. E., Anderson, J. A. *et al.* (1991) Construction of a Restriction Fragment Length Polymorphism Map for Barley (*Hordeum vulgare*). *Genome* 34: 437-447.

Hill, W. G. and Robertson, A. (1968) Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* 38: 226-231.

Hillier, L. W., Fulton, R. S., Fulton, L. A. *et al.* (2003) The DNA sequence of human chromosome 7. *Nature* 424: 157-1U2.

Hilton, H. and Gaut, B. S. (1998) Speciation and domestication in maize and its wild relatives: Evidence from the globulin-1 gene. *Genetics* 150: 863-872.

Hirochika, H. (2001) Contribution of the Tos17 retrotransposon to rice functional genomics. *Current Opinion in Plant Biology* 4: 118-122.

Hudson, R. R., Kreitman, M., and Aguadé, M. (1987) A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics* 116: 153-159.

Hugot, J. P., Chamaillard, M., Zouali, H. *et al.* (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 411: 599-603.

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

Jin, Y. K. and Bennetzen, J. L. (1994) Integration and Nonrandom Mutation of a Plasma-Membrane Proton ATPase Gene Fragment Within the Bs1 Retroelement of Maize. *Plant Cell* 6: 1177-1186.

Johanson, U., West, J., Lister, C. *et al.* (2000) Molecular analysis of FRIGIDA, a major determinant of natural variation in *Arabidopsis* flowering time. *Science* 290: 344-347.

Johnson, G. C. L., Esposito, L., Barratt, B. J. *et al.* (2001) Haplotype tagging for the identification of common disease genes. *Nature Genetics* 29: 233-237.

Jolly, C. J., Glenn, G. M., and Rahman, S. (1996) GSP-1 genes are linked to the grain hardness locus (H alpha) on wheat chromosome 5D. *Proceedings of the National Academy of Sciences of the United States of America* 93: 2408-2413.

Jolly, C. J., Rahman, S., Kortt, A. A. *et al.* (1993) Characterization of the Wheat Mr 15000 Grain-Softness Protein and Analysis of the Relationship Between Its Accumulation in the Whole Seed and Grain Softness. *Theoretical and Applied Genetics* 86: 589-597.

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

Kaplan, N. L., Hudson, R. R., and Langley, C. H. (1989) The Hitchhiking Effect Revisited. *Genetics* 123: 887-899.

Kearsey, M. J. and Farquhar, A. G. L. (1998) QTL analysis in plants; where are we now? *Heredity* 80: 137-142.

Keller, B. and Feuillet, C. (2000) Colinearity and gene density in grass genomes. *Trends in Plant Science* 5: 246-251.

Kerem, B. S., Rommens, J. M., Buchanan, J. A. *et al.* (1989) Identification of the Cystic Fibrosis Gene - Genetic Analysis. *Science* 245: 1073-1080.

Kidwell, M. G. and Lisch, D. R. (2000) Transposable elements and host genome evolution. *Trends in Ecology and Evolution* 15: 95-99.

Kilian, A., Chen, J., Han, F. *et al.* (1997) Towards map-based cloning of the barley stem rust resistance genes Rpgl and rpg4 using rice as an intergenomic cloning vehicle. *Plant Molecular Biology* 35: 187-195.

Kipling, D. and Warburton, P. E. (1997) Centromeres, CENP-B and Tigger too. *Trends in Genetics* 13: 141-145.

Kislev, M. E., Nadel, D., and Carmi, I. (1992) Epipalaeolithic (19,000 bp) Cereal and Fruit Diet at Ohalo-Ii, Sea of Galilee, Israel. *Review of Palaeobotany and Palynology* 73: 161-166.

Kleine, M., Michalek, W., Graner, A. *et al.* (1993a) Construction of a Barley (*Hordeum vulgare* L) YAC Library and Isolation of a Hor1-Specific Clone. *Molecular & General Genetics* 240: 265-272.

Kleine, M., Michalek, W., Graner, A. *et al.* (1993b) Construction of A Barley (Hordeum-Vulgare L) Yac Library and Isolation of A Hor1-Specific Clone. *Molecular & General Genetics* 240: 265-272.

Kleinhofs A, Han JH (2002) Molecular Mapping of the Barley Genome. In Slafer GA et al (eds) Barley Science. Binghamton, NY: Food Products Press. 31-63

Kleinhofs, A., Kilian, A., Maroof, M. A. S. *et al.* (1993) A Molecular, Isozyme and Morphological Map of the Barley (*Hordeum vulgare*) Genome. *Theoretical and Applied Genetics* 86: 705-712.

Kloeckener-Gruissem, B. and Freeling, M. (1995) Transposon-Induced Promoter Scrambling - A Mechanism for the Evolution of New Alleles. *Proceedings of the National Academy of Sciences of the United States of America* 92: 1836-1840.

Kooijman, M., Orsel, R., Hessing, M. *et al.* (1997) Spectroscopic characterisation of the lipid-binding properties of wheat puroindolines. *Journal of Cereal Science* 26: 145-159.

Koprek, T., McElroy, D., Louwerse, J. *et al.* (2000) An efficient method for dispersing Ds elements in the barley genome as a tool for determining gene function. *Plant Journal* 24: 253-263.

Krishnamurthy, K., Balconi, C., Sherwood, J. E. *et al.* (2001a) Wheat puroindolines enhance fungal disease resistance in transgenic rice. *Molecular Plant-Microbe Interactions* 14: 1255-1260.

Krishnamurthy, K. and Giroux, M. J. (2001b) Expression of wheat puroindoline genes in transgenic rice enhances grain softness. *Nature Biotechnology* 19: 162-166.

Kuersten, S. and Goodwin, E. B. (2003) The power of the 3 ' UTR: Translational control and development. *Nature Reviews Genetics* 4: 626-637.

Kumar, A. and Bennetzen, J. L. (1999) Plant retrotransposons. *Annual Review of Genetics* 33: 479-532.

Kunzel, G., Korzun, L., and Meister, A. (2000) Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. *Genetics* 154: 397-412.

Kurata, N., Moore, G., Nagamura, Y. *et al.* (1994) Conservation of Genome Structure Between Rice and Wheat. *Bio-Technology* 12: 276-278.

Kutcher, H. R., Bailey, K. L., Rossnagel, B. G. *et al.* (1996) Linked morphological and molecular markers associated with common root rot reaction in barley. *Canadian Journal of Plant Science* 76: 879-883.

La Rota, M. and Sorrells, M. (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.

Labuschagne, M. T. and van Vuuren, A. (2000) The inheritance and expression of grain texture in wheat, as measured by a microtome procedure. *Euphytica* 112: 261-265.

Lagercrantz, U. (1998) Comparative mapping between *Arabidopsis thaliana* and *Brassica nigra* indicates that Brassica genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. *Genetics* 150: 1217-1228.

Lahaye, T., Hartmann, S., Topsch, S. *et al.* (1998) High-resolution genetic and physical mapping of the Rar1 locus in barley. *Theoretical and Applied Genetics* 97: 526-534.

Langridge, P. and Barr, A. R. (2003) Preface. *Australian Journal of Agricultural Research* 54: i-iv.

Langridge, P., Karakousis, A., Collins, N. *et al.* (1995) A Consensus Linkage Map of Barley. *Molecular Breeding* 1: 389-395.

Law CN et al (1978) The Study of Grain Protein Control in Wheat using Whole Chromosome Substituion Lines. Seed Protein Improvement by Nuclear Techniques. Vienna, Austria: International Atomic Energy Agency. 483-502.

Le Guerneve, C., Seigneuret, M., and Marion, D. (1998) Interaction of the wheat endosperm lipid-binding protein puroindoline-a with phospholipids. *Archives of Biochemistry and Biophysics* 360: 179-186.

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.

Lewontin, R. C. and Kojima, K. (1964) The evolutionary dynamics of complex polymorphisms. *Evolution* 14: 458-472

Li, C. D., Langridge, P., Lance, R. C. M. *et al.* (1996) Seven members of the (1-3)-beta-glucanase gene family in barley (*Hordeum vulgare*) are clustered on the long arm of chromosome 3 (3HL). *Theoretical and Applied Genetics* 92: 791-796.

Li, C. D., Zhang, X. Q., Eckstein, P. *et al.* (1999) A polymorphic microsatellite in the limit dextrinase gene of barley (*Hordeum vulgare* L.). *Molecular Breeding* 5: 569-577.

Li, W. L. and Gill, B. S. (2002) The colinearity of the Sh2/A1 orthologous region in rice, sorghum and maize is interrupted and accompanied by genome expansion in the Triticeae. *Genetics* 160: 1153-1162.

Li, Y., Baldauf, S., Lim, E. K. *et al.* (2001) Phylogenetic analysis of the UDP-glycosyltransferase multigene family of *Arabidopsis thaliana*. *Journal of Biological Chemistry* 276: 4338-4343.

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

Lim, J. K. and Simmons, M. J. (1994) Gross Chromosome Rearrangements Mediated by Transposable Elements in *Drosophila melanogaster*. *Bioessays* 16: 269-275.

Lin, J. Z., Brown, A. H. D., and Clegg, M. T. (2001) Heterogeneous geographic patterns of nucleotide sequence diversity between two alcohol dehydrogenase genes in wild barley (*Hordeum vulgare* subspecies *spontaneum*). *Proceedings of the National Academy of Sciences of the United States of America* 98: 531-536.

Lin, J. Z., Morrell, P. L., and Clegg, M. T. (2002) The influence of linkage and inbreeding on patterns of nucleotide sequence diversity at duplicate alcohol dehydrogenase loci in wild barley (*Hordeum vulgare* ssp *spontaneum*). *Genetics* 162: 2007-2015.

Lin, X. Y., Kaul, S. S., Rounsley, S. *et al.* (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* 402: 761-768.

Lister, C., Jackson, D., and Martin, C. (1993) Transposon-Induced Inversion in *Antirrhinum* Modifies Nivea Gene Expression to Give A Novel Flower Color Pattern Under the Control of Cycloidea (Radialis). *Plant Cell* 5: 1541-1553.

Liu, F., Charlesworth, D., and Kreitman, M. (1999) The effect of mating system differences on nucleotide diversity at the phosphoglucose isomerase locus in the plant genus *Leavenworthia*. *Genetics* 151: 343-357.

Liu, F., Zhang, L., and Charlesworth, D. (1998) Genetic diversity in *Leavenworthia* populations with different inbreeding levels. *Proceedings of the Royal Society of London Series B-Biological Sciences* 265: 293-301.

Manninen, I. and Schulman, A. H. (1993) Bare-1, A Copia-Like Retroelement in Barley (*Hordeum vulgare* L). *Plant Molecular Biology* 22: 829-846.

Marillonnet, S. and Wessler, S. R. (1997) Retrotransposon insertion into the maize waxy gene results in tissue-specific RNA processing. *Plant Cell* 9: 967-978.

Marmiroli, N., Maestri, E., Liviero, L. *et al.* (1999) Application of genomics in assessing biodiversity in wild and cultivated barley. *Molecular Ecology* 8: S95-S106.

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

Mather, D. E., Tinker, N. A., LaBerge, D. E. *et al.* (1997) Regions of the genome that affect grain and malt quality in a North American two-row barley cross. *Crop Science* 37: 544-554.

Matsoukas, N. P. and Morrison, W. R. (1991) Breadmaking Quality of 10 Greek Breadwheats .2. Relationships of Protein, Lipid and Starch Components to Baking Quality. *Journal of the Science of Food and Agriculture* 55: 87-101.

Mattern PJ et al (1973) Location of genes for kernel properties in the wheat cultivar 'Cheyenne' using chromosome substitution lines. In Sears ER, Sears LMS (eds) Proceedings of the 4th International Wheat Genetics Symposium. Coulumbia, MO: Agricultural Experimental Station, University of Missouri. 703-707.

May, B. P., Liu, H., Vollbrecht, E. *et al.* (2003) Maize-targeted mutagenesis: A knockout resource for maize. *Proceedings of the National Academy of Sciences of the United States of America* 100: 11541-11546.

Mayer, K., Schuller, C., Wambutt, R. *et al.* (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* 402: 769-777.

Mazumder, B., Seshadri, V., and Fox, P. L. (2003) Translational control by the 3 '-UTR: the ends specify the means. *Trends in Biochemical Sciences* 28: 91-98.

McCallum, C. M., Comai, L., Greene, E. A. *et al.* (2000) Targeting induced local lesions in genomes (TILLING) for plant functional genomics. *Plant Physiology* 123: 439-442.

McDonald, J. F., Matyunina, L. V., Wilson, S. *et al.* (1997) LTR retrotransposons and the evolution of eukaryotic enhancers. *Genetica* 100: 3-13.

McDonald, J. H. and Kreitman, M. (1991) Adaptive Protein Evolution at the *Adh* Locus in *Drosophila*. *Nature* 351: 652-654.

Melis, R., Bradley, P., Elsner, T. *et al.* (1993) Polymorphic Ssr (Simple-Sequence-Repeat) Markers for Chromosome-20. *Genomics* 16: 56-62.

Mendel, G. (1865) Versuche über Pflanzenhybriden. *Verhandlungen des Naturforschenden Vereines in Brünn* 4: 3-47.

Miklos, G. L. G. and Rubin, G. M. (1996) The role of the genome project in determining gene function: Insights from model organisms. *Cell* 86: 521-529.

Miller, J. T., Dong, F. G., Jackson, S. A. *et al.* (1998) Retrotransposon-related DNA sequences in the centromeres of grass chromosomes. *Genetics* 150: 1615-1623.

Moore, G., Devos, K. M., Wang, Z. *et al.* (1995) Cereal Genome Evolution - Grasses, Line Up and Form A Circle. *Current Biology* 5: 737-739.

Morden, C. W., Doebley, J., and Schertz, K. F. (1990) Allozyme Variation Among the Spontaneous Species of *Sorghum* Section *Sorghum* (Poaceae). *Theoretical and Applied Genetics* 80: 296-304.

Morrell, P. L., Lundy, K. E., and Clegg, M. T. (2003) Distinct geographic patterns of genetic diversity are maintained in wild barley (*Hordeum vulgare* ssp *spontaneum*) despite migration. *Proceedings of the National Academy of Sciences of the United States of America* 100: 10812-10817.

Morris, C. F., Greenblatt, G. A., Bettge, A. D. *et al.* (1994) Isolation and Characterization of Multiple Forms of Friabilin. *Journal of Cereal Science* 20: 167-174.

Morris, C. F., Lillemo, M., Simeone, M. C. *et al.* (2001) Prevalence of puroindoline grain hardness genotypes among historically significant North American spring and winter wheats. *Crop Science* 41: 218-228.

Morris CF, Rose SP (1996) Wheat. In Henry RJ, Kettlewell PS (eds) Cereal grain quality. New York: Chapman and Hall. 3-54.

Morrison, W. R., Greenwell, P., Law, C. N. *et al.* (1992) Occurrence of Friabilin, A Low-Molecular-Weight Protein Associated with Grain Softness, on Starch Granules Isolated from Some Wheats and Related Species. *Journal of Cereal Science* 15: 143-149.

Mueller, U. G. and Wolfenbarger, L. L. (1999) AFLP genotyping and fingerprinting. *Trends in Ecology & Evolution* 14: 389-394.

Myouga, F., Tsuchimoto, S., Noma, K. *et al.* (2001) Identification and structural analysis of SINE elements in the *Arabidopsis thaliana* genome. *Genes & Genetic Systems* 76: 169-179.

Nei, M. (1987) Molecular Evolutionary Genetics. Columbia University Press, New York, NY.

Nei, M. and Miller, J. C. (1990) A Simple Method for Estimating Average Number of Nucleotide Substitutions Within and Between Populations from Restriction Data. *Genetics* 125: 873-879.

Nordborg, M. (2000) Linkage disequilibrium, gene trees and selfing: An ancestral recombination graph with partial self-fertilization. *Genetics* 154: 923-929.

Nordborg, M., Borevitz, J. O., Bergelson, J. *et al.* (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics* 30: 190-193.

Nordborg, M. and Donnelly, P. (1997) The coalescent process with selfing. *Genetics* 146: 1185-1195.

Oda, S. (1994) 2-Dimensional Electrophoretic Analysis of Friabilin. *Cereal Chemistry* 71: 394-395.

Oda, S., Komae, K., and Yasui, T. (1992) Relation Between Starch Granule Protein and Endosperm Softness in Japanese Wheat (*Triticum aestivum* L) Cultivars. *Japanese Journal of Breeding* 42: 161-165.

Oda, S. and Schofield, J. D. (1997) Characterisation of friabilin polypeptides. *Journal of Cereal Science* 26: 29-36.

Ogura, T. and Wilkinson, A. J. (2001a) AAA(+) superfamily ATPases: common structure-diverse function. *Genes to Cells* 6: 575-597.

Ogura, Y., Bonen, D. K., Inohara, N. *et al.* (2001b) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* 411: 603-606.

Oka HI. 1988. *Origin of cultivated rice.* New York: Elsevier Science Publishing Co.

Palaisa, KA, Morgante, M, Williams, M *et al.* (2003) Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. *Plant Cell* 15: 1795-1806.

Panstruga, R., Buschges, R., Piffanelli, P. *et al.* (1998) A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome. *Nucleic Acids Research* 26: 1056-1062.

Pardue, M. L., Danilevskaya, O. N., Traverse, K. L. *et al.* (1997) Evolutionary links between telomeres and transposable elements. *Genetica* 100: 73-84.

Parinov, S. and Sundaresan, V. (2000) Functional genomics in Arabidopsis: large-scale insertional mutagenesis complements the genome sequencing project. *Current Opinion in Biotechnology* 11: 157-161.

Patanjali, S. R., Parimoo, S., and Weissman, S. M. (1991) Construction of A Uniform-Abundance (Normalized) Cdna Library. *Proceedings of the National Academy of Sciences of the United States of America* 88: 1943-1947.

Patel, S. and Latterich, M. (1998) The AAA team: related ATPases with diverse functions. *Trends in Cell Biology* 8: 65-71.

Paterson, A. H., Deverna, J. W., Lanini, B. *et al.* (1990) Fine Mapping of Quantitative Trait Loci Using Selected Overlapping Recombinant Chromosomes, in An Interspecies Cross of Tomato. *Genetics* 124: 735-742.

Pedersen, C., Giese, H., and Linde-Laursen, I. (1995) Towards an integration of the physical and the genetic chromosome maps of barley by *in situ* hybridization. *Hereditas* 123: 77-88.

Pickering, R. A. and Devaux, P. (1992) Haploid production: approaches and uses in plant breeding. In Shewry, P.R (Eds) Barley: Genetics, Biochemistry, Molecular Biology and Biotechnology. Cab International, Wallingford, UK. 519-548.

Pimpinelli, S., Berloco, M., Fanti, L. *et al.* (1995) Transposable Elements Are Stable Structural Components of *Drosophila melanogaster* Heterochromatin. *Proceedings of the National Academy of Sciences of the United States of America* 92: 3804-3808.

Pollak, E. and Sabran, M. (1992) On the Theory of Partially Inbreeding Finite Populations .3. Fixation Probabilities Under Partial Selfing When Heterozygotes Are Intermediate in Viability. *Genetics* 131: 979-985.

Pomeranz, Y., Bolling, H., and Zwingelberg, H. (1984) Wheat Hardness and Baking Properties of Wheat Flours. *Journal of Cereal Science* 2: 137-143.

Powell, W., Machray, G. C., and Provan, J. (1996a) Polymorphism revealed by simple sequence repeats. *Trends in Plant Science* 1: 215-222.

Powell, W., Morgante, M, Andre, C *et al.* (1996b) The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. *Molecular Breeding* 2: 225-238.

Presting, G. G., Malysheva, L., Fuchs, J. *et al.* (1998) A TY3/GYPSY retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes. *Plant Journal* 16: 721-728.

Pritchard, J. K. and Przeworski, M. (2001) Linkage disequilibrium in humans: Models and data. *American Journal of Human Genetics* 69: 1-14.

Pritchard, J. K. and Rosenberg, N. A. (1999) Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics* 65: 220-228.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* 155: 945-959.

Przeworski, M. (2002) The signature of positive selection at randomly chosen loci. *Genetics* 160: 1179-1189.

Qi, X., Stam, P., and Lindhout, P. (1998) Use of locus-specific AFLP markers to construct a high-density molecular map in barley. *Theoretical and Applied Genetics* 96: 376-384.

Raboy, V., Young, K. A., Dorsch, J. A. *et al.* (2001) Genetics and breeding of seed phosphorus and phytic acid. *Journal of Plant Physiology* 158: 489-497.

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

Rafalski, J. A. (2002b) Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Science* 162: 329-333.

Rahman, S., Jolly, C. J., Skerritt, J. H. *et al.* (1994) Cloning of A Wheat 15-KDa Grain Softness Protein (GSP) - GSP Is A Mixture of Puroindoline-Like Polypeptides. *European Journal of Biochemistry* 223: 917-925.

Ramsay, L. D., Jennings, D. E., Bohuon, E. J. R. *et al.* (1996) The construction of a substitution library of recombinant backcross lines in *Brassica oleracea* for the precision mapping of quantitative trait loci. *Genome* 39: 558-567.

Reich, D. E., Cargill, M., Bolk, S. *et al.* (2001) Linkage disequilibrium in the human genome. *Nature* 411: 199-204.

Remington, D. L., Thornsberry, J. M., Matsuoka, Y. *et al.* (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences of the United States of America* 98: 11479-11484.

Riggs, T. J., Hanson, P. R., and Start, N. D. (1981) Genetic improvement in yield of spring barley and associated chanves in plant phenotype. *Barley Genetics* 4th: 97-103.

Rioux, J. D., Daly, M. J., Silverberg, M. S. *et al.* (2001) Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn's disease. *Nature Genetics* 29: 223-228.

Roberts, M. A., Reader, S. M., Dalgliesh, C. *et al.* (1999) Induction and characterization of *Ph1* wheat mutants. *Genetics* 153: 1909-1918.

Ross, J., Li, Y., Lim, E. K. *et al.* (2001) Higher plant glycosyltransferases. *Genome Biology* 2: reviews3004.1-3004.6.

Rostoks, N, Park, Y-J, Ramakrishna, W *et al.* (2002) Genomic sequencing reveals gene content, genomic organization, and recombination relationships in barley. *Functional and Integrative Genomics* 2: 51-59.

Rouves, S., Boeuf, C., ZwickertMenteur, S. *et al.* (1996) Locating supplementary RFLP markers on barley chromosome 7 and synteny with homoeologous wheat group 5. *Plant Breeding* 115: 511-513.

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

Ruiz, M. T., Voinnet, O., and Baulcombe, D. C. (1998) Initiation and maintenance of virus-induced gene silencing. *Plant Cell* 10: 937-946.

Saeki, K., Miyazaki, C., Hirota, N. *et al.* (1999) RFLP mapping of BaYMV resistance gene rym3 in barley (*Hordeum vulgare*). *Theoretical and Applied Genetics* 99: 727-732.

Sakata, K., Nagamura, Y., Numa, H. *et al.* (2002) RiceGAAS: an automated annotation system and database for rice genome sequence. *Nucleic Acids Research* 30: 98-102.

Salamini, F., Ozkan, H., Brandolinin, A. *et al.* (2002) Genetics and geography of wild cereal domestication in the Near East. *Nature Reviews Genetics* 3: 429-441.

Salanoubat, M., Lemcke, K., Rieger, M. *et al.* (2000) Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. *Nature* 408: 820-822.

SanMiguel, P. and Bennetzen, J. L. (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Annals of Botany* 82: 37-44.

SanMiguel, P.J., Ramakrishna, W., Bennetzen, J.L. et al. (2002) Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A (m). *Functional and Integrative Genomics* 2: 70-80.

SanMiguel, P., Tikhonov, A., Jin, Y. K. *et al.* (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274: 765-768.

Sasaki, T., Matsumoto, T., Yamamoto, K. *et al.* (2002) The genome sequence and structure of rice chromosome 1. *Nature* 420: 312-316.

Savolainen, O., Langley, C. H., Lazzaro, B. P. *et al.* (2000) Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. *Molecular Biology and Evolution* 17: 645-655.

Schnable, P. S., Hsia, A. P., and Nikolau, B. J. (1998) Genetic recombination in plants. *Current Opinion in Plant Biology* 1: 123-129.

Schondelmaier, J., Jacobi, A., Fischbeck, G. *et al.* (1992) Genetic-Studies on the Mode of Inheritance and Localization of the Amo1 (High Amylose) Gene in Barley. *Plant Breeding* 109: 274-280.

Seko, H. and Kato, I. (1981) Breeding of Naked Barley for Protein Improvement. *Jarq-Japan Agricultural Research Quarterly* 14: 206-209.

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

Shepherd, N. S., Schwarzsommer, Z., Velspalve, J. B. *et al.* (1984) Similarity of the Cin1 Repetitive Family of *Zea mays* to Eukaryotic Transposable Elements. *Nature* 307: 185-187.

Sherman, J. D., Fenwick, A. L., Namuth, D. M. *et al.* (1995) A Barley RFLP Map - Alignment of 3 Barley Maps and Comparisons to Gramineae Species. *Theoretical and Applied Genetics* 91: 681-690.

Shewry, P. R., Beaudoin, F., Jenkins, J. *et al.* (2002) Plant protein families and their relationships to food allergy. *Biochemical Society Transactions* 30: 906-910.

Shields, R. (1993) Plant Genetics - Pastoral Synteny. *Nature* 365: 297-298.

Shirasu, K., Schulman, A. H., Lahaye, T. *et al.* (2000) A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Research* 10: 908-915.

Shirley, B. W. (1996) Flavonoid biosynthesis: 'New' functions for an 'old' pathway. *Trends in Plant Science* 1: 377-382.

Shizuya, H., Birren, B., Kim, U. J. *et al.* (1992a) Cloning and Stable Maintenance of 300-Kilobase-Pair Fragments of Human DNA in *Escherichia coli* Using An F-Factor-Based Vector. *Proceedings of the National Academy of Sciences of the United States of America* 89: 8794-8797.

Shizuya, H., Birren, B., Kim, U. J. *et al.* (1992b) Cloning and Stable Maintenance of 300-Kilobase-Pair Fragments of Human Dna in Escherichia-Coli Using An F-Factor-Based Vector. *Proceedings of the National Academy of Sciences of the United States of America* 89: 8794-8797.

Silvey, V. (1986) The contribution of new varieties to cereal yields in England and Wales between 1947 and 1983. *Journal of the National Institute of Agricultural Botany* 17: 155-168.

Simonsen, K. L., Churchill, G. A., and Aquadro, C. F. (1995) Properties of Statistical Tests of Neutrality for DNA Polymorphism Data. *Genetics* 141: 413-429.

Simpson, C. G., Thow, G., Clark, G. P. *et al.* (2002) Mutational analysis of a plant branchpoint and polypyrimidine tract required for constitutive splicing of a mini-exon. *Rna-A Publication of the Rna Society* 8: 47-56.

Sogaard, B. and Vonwettsteinknowles, P. (1987) Barley - Genes and Chromosomes. *Carlsberg Research Communications* 52: 123-196.

Sollner, T., Bennett, M. K., Whiteheart, S. W. *et al.* (1993) A Protein Assembly-Disassembly Pathway *In Vitro* That May Correspond to Sequential Steps of Synaptic Vesicle Docking, Activation, and Fusion. *Cell* 75: 409-418.

Sonnhammer, E. L. L. and Durbin, R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167: 1-10.

Sorrells, M. E., La Rota, M., Bermudez-Kandianis, C. E. *et al.* (2003) Comparative DNA sequence analysis of wheat and rice genomes. *Genome Research* 13: 1818-1827.

Sourdille, P., Perretant, M. R., Charmet, G. *et al.* (1996) Linkage between RFLP markers and genes affecting kernel hardness in wheat. *Theoretical and Applied Genetics* 93: 580-586.

Stam, P. (1993) Construction of Integrated Genetic-Linkage Maps by Means of A New Computer Package - Joinmap. *Plant Journal* 3: 739-744.

Stam, P. (1995) JoinMap (tm) version 2.0: Software for the calculation of genetic linkage maps.

Stephan, W. and Langley, C. H. (1989) Molecular Genetic-Variation in the Centromeric Region of the X- Chromosome in 3 *Drosophila ananassae* Populations .1. Contrasts Between the Vermilion and Forked Loci. *Genetics* 121: 89-99.

Swanston JS, Ellis RP (1999) Genetics and Breeding of Malt Quality Attributes. In Slafer GA et al (eds) Barley Science. New York: Food Products Press. 85-114.

Symes, K. J. (1965) The Inheritance of Grain Hardness in Wheat as Measured by the Particle Size Index. *Australian Journal of Agricultural Research* 16: 113-123.

Tabata, S., Kaneko, T., Nakamura, Y. *et al.* (2000) Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. *Nature* 408: 823-826.

Taillon-Miller, P., Bauer-Sardina, I., Saccone, N. L. *et al.* (2000) Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genetics* 25: 324-328.

Tajima, F. (1983) Evolutionary Relationship of DNA Sequences in Finite Populations. *Genetics* 105: 437-460.

Tajima, F. (1989) Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* 123: 585-595.

Tajima, F. (1993) Measurement of DNA Polymorphism. In Takahata, N and Clark, A.G. (eds), Mechanism of Moleculat Evolution, Sinauer Associates. Inc., Sunderland, MA. 37-59.

Takahashi, S., Inagaki, Y., Satoh, H. *et al.* (1999) Capture of a genomic HMG domain sequence by the En/Spm-related transposable element Tpn1 in the Japanese morning glory. *Molecular and General Genetics* 261: 447-451.

Tautz, D. and Renz, M. (1984) Simple Sequences Are Ubiquitous Repetitive Components of Eukaryotic Genomes. *Nucleic Acids Research* 12: 4127-4138.

Tenaillon, M. I., Sawkins, M. C., Anderson, L. K. *et al.* (2002) Patterns of diversity and recombination along chromosome 1 of maize (*Zea mays* ssp *mays* L.). *Genetics* 162: 1401-1413.

Tenaillon, M. I., Sawkins, M. C., Long, A. D. *et al.* (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp *mays* L.). *Proceedings of the National Academy of Sciences of the United States of America* 98: 9161-9166.

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

Theologis, A., Ecker, J. R., Palm, C. J. *et al.* (2000) Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*. *Nature* 408: 816-820.

Thomas WTB (2002) Molecular Marker-Assisted versus Conventional Selection. In Slafer GA et al (eds) Barley Science. Binghamton, NY: Food Products Press. 177-204.

Thomas, W. T. B., Powell, W., Swanston, J. S. *et al.* (1996) Quantitative trait loci for germination and malting quality characters in a spring barley cross. *Crop Science* 36: 265-273.

Thornsberry, J. M., Goodman, M. M., Doebley, J. *et al.* (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics* 28: 286-289.

Tian, D. C., Araki, H., Stahl, E. *et al.* (2002) Signature of balancing selection in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America* 99: 11525-11530.

Tikhonov, A. P., SanMiguel, P. J., Nakajima, Y. *et al.* (1999) Colinearity and its exceptions in orthologous adh regions of maize and sorghum. *Proceedings of the National Academy of Sciences of the United States of America* 96: 7409-7414.

Trimble, W. S., Cowan, D. M., and Scheller, R. H. (1988) VAMP-1: A synaptic vesicle-associated integral membrane protein. *Proceedings of the National Academy of Sciences of the United States of America* 85: 4538-4542.

Ullrich, S. E., Han, F., and Jones, B. L. (1997) Genetic complexity of the malt extract trait in barley suggested by QTL analysis. *Journal of the American Society of Brewing Chemists* 55: 1-4.

Vanin, E. F. (1985) Processed Pseudogenes - Characteristics and Evolution. *Annual Review of Genetics* 19: 253-272.

Varagona, M. J., Purugganan, M., and Wessler, S. R. (1992) Alternative Splicing Induced by Insertion of Retrotransposons Into the Maize Waxy Gene. *Plant Cell* 4: 811-820.

Venter, J. C., Adams, M. D., Myers, E. W. *et al.* (2001) The sequence of the human genome. *Science* 291: 1304-1351.

Vicient, C. M., Suoniemi, A., Anamthamat-Jonsson, K. *et al.* (1999) Retrotransposon BARE-1 and its role in genome evolution in the genus *Hordeum*. *Plant Cell* 11: 1769-1784.

Vivar, H. E. Building Multiple Disease Resistance into a High Yielding Platform. Louge, S. Proceedings of the 8th International Barley Genetics Symposium. I, 280-286. 2000. Glen Osmond, Department of Plant Science, Waite Campus, Adelaide University. 280-286.

von Bothmer, R., Jacobsen, N., Baden, C., *et al.* (1995) An ecogeographical study of the genus Hordeum. 2nd edition. Sytematic and ecogeographical studies on crop genepools 7. International Plant Genetic Resources Institute, Rome.

Vos, P., Hogers, R., Bleeker, M. *et al.* (1995a) AFLP - A New Technique for DNA Fingerprinting. *Nucleic Acids Research* 23: 4407-4414.

Vos, P., Hogers, R., Bleeker, M. *et al.* (1995b) Aflp - A New Technique for Dna-Fingerprinting. *Nucleic Acids Research* 23: 4407-4414.

Wall, J. D., Andolfatto, P., and Przeworski, M. (2002) Testing models of selection and demography in *Drosophila simulans*. *Genetics* 162: 203-216.

Wang, R. L., Stec, A., Hey, J. *et al.* (1999) The limits of selection during maize domestication. *Nature* 398: 236-239.

Wardrop, J., Snape, J., Powell, W. *et al.* (2002) Constructing plant radiation hybrid panels. *Plant Journal* 31: 223-228.

Watterson, G. A. (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* 7: 256-276.

Waugh, R., Bonar, N., Baird, E. *et al.* (1997) Homology of AFLP products in three mapping populations of barley. *Molecular & General Genetics* 255: 311-321.

Waugh, R., Dear, P. H., Powell, W. *et al.* (2002) Physical education - new technologies for mapping plant genomes. *Trends in Plant Science* 7: 521-523.

Weber, T., Zemelman, B. V., Mcnew, J. A. *et al.* (1998) SNAREpins: Minimal machinery for membrane fusion. *Cell* 92: 759-772.

Wei, F. S., Wong, R. A., and Wise, R. P. (2002) Genome dynamics and evolution of the Mla (powdery mildew) resistance locus in barley. *Plant Cell* 14: 1903-1917.

Weigel, D., Ahn, J. H., Blazquez, M. A. *et al.* (2000) Activation tagging in Arabidopsis. *Plant Physiology* 122: 1003-1013.

Weiner, A. M., Deininger, P. L., and Efstratiadis, A. (1986) Nonviral Retroposons - Genes, Pseudogenes, and Transposable Elements Generated by the Reverse Flow of Genetic Information. *Annual Review of Biochemistry* 55: 631-661.

Welsh, J. and Mcclelland, M. (1990) Fingerprinting Genomes Using Pcr with Arbitrary Primers. *Nucleic Acids Research* 18: 7213-7218.

Wendel, J. F. (2000) Genome evolution in polyploids. *Plant Molecular Biology* 42: 225-249.

Werner, J. E., Endo, T. R., and Gill, B. S. (1992) Toward A Cytogenetically Based Physical Map of the Wheat Genome. *Proceedings of the National Academy of Sciences of the United States of America* 89: 11307-11311.

Weyen, J., Bauer, E., Graner, A. *et al.* (1996) RAPD-mapping of the distal portion of chromosome 3 of barley, including the BaMMV/BaYMV resistance gene ym4. *Plant Breeding* 115: 285-287.

White, S. E. and Doebley, J. F. (1999) The molecular evolution of terminal earl, a regulatory gene in the genus *Zea*. *Genetics* 153: 1455-1462.

White, S. E., Habera, L. F., and Wessler, S. R. (1994) Retrotransposons in the Flanking Regions of Normal Plant Genes - A Role for Copia-Like Elements in the Evolution of Gene Structure and Expression. *Proceedings of the National Academy of Sciences of the United States of America* 91: 11792-11796.

Whitmore, E. T. and Sparrow, D. H. B. (1957) Laboratory micro-malting techniques. *Journal of the Institute of Brewing* 85: 262-265.

Wicker, T., Guyot, R, Yahiaoui, N *et al.* (2003a) CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements. *Plant Physiology* 132: 52-63.

Wicker, T., Matthews, D. E., and Keller, B. (2002) TREP: a database for Triticeae repetitive element. *Trends in Plant Science* 7: 561-562.

Wicker, T., Stein, N., Albar, L. *et al.* (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. *Plant Journal* 26: 307-316.

Wicker, T., Yahiaoui, N., Guyot, R., Schlagenhauf, E., Liu, Z.D., Dubcovsky, J., Keller, B. (2003b) Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A(m) genomes of wheat. *Plant Cell* 15: 1186-1197.

Williams, J. G. K., Kubelik, A. R., Livak, K. J. *et al.* (1990) Dna Polymorphisms Amplified by Arbitrary Primers Are Useful As Genetic-Markers. *Nucleic Acids Research* 18: 6531-6535.

Williamson, V. M. (1983) Transposable Elements in Yeast. *International Review of Cytology-A Survey of Cell Biology* 83: 1-25.

Witte, C. P., Le, Q. H., Bureau, T. *et al.* (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proceedings of the National Academy of Sciences of the United States of America* 98: 13778-13783.

Wolfe, K. H., Sharp, P. M., and Li, W. H. (1989) Mutation Rates Differ Among Regions of the Mammalian Genome. *Nature* 337: 283-285.

Woo, S. S., Jiang, J. M., Gill, B. S. *et al.* (1994a) Construction and Characterization of a Bacterial Artificial Chromosome Library of *Sorghum bicolor*. *Nucleic Acids Research* 22: 4922-4931.

Woo, S. S., Jiang, J. M., Gill, B. S. *et al.* (1994b) Construction and Characterization of A Bacterial Artificial Chromosome Library of Sorghum-Bicolor. *Nucleic Acids Research* 22: 4922-4931.

Wu, J., Yamagata, H., Hayashi-Tsugane, M. *et al.* (2004) Composition and structure of the centromeric region of rice chromosome 8. *Plant Cell* 16: 967-976.

Yan, L., Echenique, V., Busso, C. *et al.* (2002) Cereal genes similar to *Snf2* define a new subfamily that includes human and mouse genes. *Molecular & General Genetics* 268: 488-499.

Yan, L., Loukoianov, A., Tranquilli, G. *et al.* (2003) Positional cloning of the wheat vernalization gene VRN1. *Proceedings of the National Academy of Sciences of the United States of America* 100: 6263-6268.

Yao, H., Zhou, Q., Li, J. *et al.* (2002) Molecular characterization of meiotic recombination across the 140-kb multigenic a1-sh2 interval of maize. *Proceedings of the National Academy of Sciences of the United States of America* 99: 6157-6162.

Yu, J., Hu, S. N., Wang, J. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). *Science* 296: 79-92.

Yu, Y., Tomkins, J. P., Waugh, R. *et al.* (2000) A bacterial artificial chromosome library for barley (*Hordeum vulgare* L.) and the identification of clones containing putative resistance genes. *Theoretical and Applied Genetics* 101: 1093-1099.

Zhang, H. B., Choi, S. D., Woo, S. S. *et al.* (1996) Construction and characterization of two rice bacterial artificial chromosome libraries from the parents of a permanent recombinant inbred mapping population. *Molecular Breeding* 2: 11-24.