Physical Mapping and Genomic Characterization of Wheat Quality Loci *Glu-B1* and *Ha*

BY

RAJA RAGUPATHY

A Thesis Submitted to the Faculty of Graduate Studies in Partial Fulfilment of the Requirements for the Degree of DOCTOR OF PHILOSOPHY

Department of Plant Science

University of Manitoba

Winnipeg, Canada

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES
*****
COPYRIGHT PERMISSION

Physical Mapping and Genomic Characterization of Wheat Quality Loci Glu-B1 and Ha

BY

Raja Ragupathy

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirement of the degree

Of

Doctor of Philosophy

FOREWORD

This thesis follows the Sandwich thesis model outlined by the Department of Plant Science, Faculty of Graduate studies, University of Manitoba. Manuscripts follow the style of the Theoretical and Applied Genetics journal. The thesis begins with a general introduction and literature review. Three manuscripts each contain an abstract, an introduction, materials and methods, results and discussion. The thesis ends with a general discussion, references and appendices.

ABSTRACT

Ragupathy, Raja. Ph.D., The University of Manitoba, July, 2008. Physical Mapping and Genomic Characterization of Wheat Quality Loci *Glu-B1* and *Ha*.

Major Professor; Dr. Sylvie Cloutier.

Bread wheat has a genome of 16000 Mb, ~100 times the size of Arabidopsis. High fractions of transposable elements add to its complexity. Though transposable elements were described as junk DNA, genome projects indicated their role in evolution of genes and genomes. In this study, retroelement mediated evolution of two loci namely, *Glu-B1* encoding high molecular weight glutenin subunit (HMW-GS) 7 and *Ha* encoding puroindolines have been demonstrated.

Sequencing of a BAC clone encompassing the *Glu-B1* locus in hexaploid wheat cultivar Glenlea revealed a 10.3 kb duplication harbouring a duplicate copy of *Bx7* gene leading to its overexpression (Bx7^{OE}) associated with strong dough. A collection of 412 wheat accessions was assessed for overexpression by SDS-PAGE and confirmed by RP-HPLC, and also for the presence of four diagnostic DNA markers. Forty three accessions found to have overexpression phenotype produced the diagnostic PCR amplicons, lines lacking overexpression did not. The results indicated that the overexpression is likely the result of the *Bx7* gene duplication driven by the insertion of an LTR copia retrotransposon named Sasanda at the *Glu-B1* locus. Discovery of three tetraploid accessions with Bx7^{OE} indicated that *Bx7* gene duplication was a pre hexaploidization event.

The Sasanda retroelement family was also characterized in the cultivar Glenlea. The copy number was estimated at ~347 elements per haploid genome. The element is at least 1.2 to 1.8 million years old and evolved into a minimum of five to nine sub families. Estimates of insertion time of 89 members based on the divergence of LTRs indicated identical LTRs in 49 elements suggesting recent transposition activity.

Sequencing of three BAC clones encompassing the *Ha* loci from the homoeologous A-, B- and D-genomes from cultivar Glenlea was carried out and sequences of 172 kb, 168 kb and 70 kb were obtained. Sequence analysis revealed retroelement driven deletion of *Pina* and *Pinb* genes in the A- and B-genomes and the *Pina* gene in the D-genome, leading to the hardness endosperm phenotype in hexaploid wheat in general and cultivar Glenlea in particular. Sequence comparions based on the A-genomes provided additional evidence for the involvement of more than one tetraploid ancestor in the origin of hexaploid wheat and therefore independent hexaplodization events.

ACKNOWLEDGEMENTS

I sincerely thank my Advisor Dr. Sylvie Cloutier for the exceptional guidance provided throughout the course of the research projects in both scientific as well as technical aspects. The numerous discussions nurtured and fine-tuned my analytical and communication skills, fuelling my passion for science in general and biology in particular. I am also most grateful and indebted for her friendship as an ideal mentor which made my stay in Canada a life altering experience.

I sincerely thank Prof. Murray Balance for serving in my advisory committee and for his valuable advice in the course of the investigations. I immensely benefited and was inspired by listening to him regarding the importance of detail and quality of data in pursuit of research excellence.

I sincerely thank Prof. Gary Fulcher for serving in my advisory committee and for his valuable advice for my research projects. It was always motivating and intellectually enriching experience to listen to his words from the basic science of evolution to practical science.

I sincerely thank Prof. Genyi Li for serving in my advisory committee and for his valuable advice for my thesis projects. It was always captivating to interact with him because of his exceptional interest in the progress of the projects and suggestions for their successful completion.

I give my special thanks to Dr. Gopalan Selvaraj for his willingness and commitment to serve as the external examiner of the thesis.

I also sincerely thank Dr. Gavin Humphreys and Dr. George Fedak of AAFC (Canada), Mr. Dallas Kessler, PGRC (Canada), Dr. Harold Bockelman, USDA-NSGC (USA), CIMMYT (Mexico), RICP (Czech Rep.) and AWCC (Australia) for kindly providing the seeds used in the study. I also sincerely thank Dr. Odean Lukow, Prof. Harry Sapirstein and Dr. Hamid Naeem for help with protein analyses.

I give my special thanks to Travis Banks for his excellent support regarding Bioinformatic analyses.

I give my special thanks to Andrzej Walichnowski for manuscript review and suggestions.

I am also sincerely thankful to Elsa Reimer, Andrzej Walichnowski, Natasa Radovanovic, Kathy Adams and Malgorzata Prochownik for their excellent technical assistance. I give my special thanks to Mike Shillinglaw for outstanding graphic support.

Funding for the projects through the Canadian Crop Genomics Initiative from Agriculture and Agriculture and Agri-Food Canada for my Advisor is acknowledged.

Financial assistance for Raja Ragupathy from the University of Manitoba Graduate Fellowship, Manitoba Graduate Scholarship, William B. Malchy Graduate Fellowship, Clarence Bougardus Sharp Memorial Scholarship and International Graduate Student Scholarship is acknowledged with gratitude.

I thank our former Dean Dr. P. Selvaraj, the current Dean Dr. L. Nadarajan, and Professor and Head of the Department, Dr. K. Paramasivam for helping me to come to Canada on extraordinary leave for study purpose and their enthusiasm and encouragement in the progress and completion of my research.

I also sincerely thank my former advisor Prof. T. S.Raveendran, inspirational Teachers Dr. R. Govindarasu, Dr. K. Omar Hattab, Dr. T. Mohapatra, Dr. B. M. Prasanna and Dr. A. K. Singh.

I thank Profs. Peter McVetty, Sylvie Cloutier, Genyi Li, Dilantha Fernando and Brian Fristensky for teaching me advances of plant science and bioinformatics and broadening my intellectual horizons. Also special thanks are due to Messrs. Martha, Bev, Vicki, Debbie and Barbara of the Plant Science Department for their excellent administrative support.

I owe my sincere thanks to the Dean of the Faculty of Graduate Studies, Prof. John (Jay) Doering for providing an ideal environment for graduate students to focus on their research and the staff of the FGS for their administrative support.

I give my special thanks to Jonathan for helping me while thesis compilation.

I thank my friends Drs. Minu, Sundar, Manick, Anbumani, Suvira, Rajesh, Huang, Saravanan, Alice and Messrs. Sentha, Rajavelu,Vadivambal, Satya, Nithya, Reem, Ramanna, Zining and Suresh for their help and morale support, and all the other people whom I have met over the past four years who have helped me to achieve this milestone.

Finally, I thank my family for understanding my goals in pursuit of the highest academic degree and other worthy career ambitions and always encouraging me to realise my dreams.

Dedicated to

My Parents,
My wife Sriletchoumy
and my son Rishi Kartikeyen

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

BAC	Bacterial Artificial Chromosome
BARE	Barley Retroelement
BLAST	Basic Local Alignment Search Tool
CWES	Canada Western Extra Strong Wheat
HMW-GS	High Molecular Weight-Glutenin Subunits
IN	*Integrase*
IRAP	Inter Retrotransposon Amplification Polymorphism
Kb	Kilobases
LTR	Long Terminal Repeat
Mb	Megabases
MEGA4	Molecular Evolutionary Genetic Analyses (Version 4)
MYA	Million Years Ago
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
QTL	Quantitative Trait Loci
REMAP	Retrotransposon-Microsatellite Amplification Polymorphism
RP-HPLC	Reversed-Phased High-Performance Liquid Chromatography
RT	*Reverse Transcriptase*
SDS-PAGE	Sodium Dodecyl Sulphate-Poly Acrylamide Gel Electrophoresis
SSAP	Sequence Specific Amplification Polymorphism
TaBAC	*Triticum aestivum* Bacterial Artificial Chromosome
TREP	Triticeae Repetitive Element Database
WIS	Wheat Insertion Sequence

LIST OF APPENDICES

Appendix **Page**

CHAPTER 1

GENERAL INTRODUCTION

1.1 Introduction

Wheat is one of the most important crops for mankind in terms of its utility as a food crop and as a classical plant genetic system, especially for understanding the evolution of polyploidy. The grain quality of released varieties is important for domestic consumption, processing into products (value addition) and for exports. Hence, a critical understanding of the genome organization of agronomically important loci is needed, to complement the efforts of molecular plant breeding, to eventually design new strategies to breed varieties with novel quality traits (Sorrells 2007). Also, understanding the evolution of structural organization of loci *per se* at the DNA sequence level, has biological significance because it serves to provide insights into the molecular mechanisms of genome evolution with implications to understand the origin of genome complexity and genome dynamics, as a whole.

Wheat seeds are comprise of starches, proteins, lipids and other components, all of which play a role in determining seed quality. Of all the cereal grains, wheat grain is unique because wheat flour alone has the ability to form dough that exhibits the rheological properties required for the production of leavened bread and for the wider diversity of products that have been developed by taking advantage of this attribute (Weegels et al. 1996; Shewry et al. 2006). These unique properties of the wheat grain are determined primarily by the storage proteins of its endosperm namely the gluten complex (Shewry et al. 1992; Anjum et al. 2007). Protein content and, particularly, protein composition are critical in determining the rheological property of the flour. High molecular weight (HMW) glutenins are the most important fraction of the gluten complex and are encoded by the genes located at the *Glu-1* loci of homoeologous chromosome 1 (Payne et al. 1987). Each locus consists of tightly linked paralogues, encoding x- and y-type subunits. Among the six HMW glutenin subunits of bread wheat, most cultivars express only 3-5 subunits with unequal contribution to quality. Subunit Dx5 which is paired with subunit Dy10 has been ranked the best subunit pair presumably because Dx5 contains an extra cysteine residue thereby allowing branching of the gluten macro polymer (Anderson and Green 1989; Radovanovic et al. 2002). Subunit Bx7 in Glenlea wheat was also demonstrated to be associated with improved gluten strength. This association was correlated with the overexpression of this subunit (D'Ovidio et al. 1997). Authors of other studies on Glenlea and other cultivars that overexpress the Bx7 subunit hypothesized that the overexpression was the result of a gene duplication at the *Glu-B1* locus (Cloutier et al. 2005). This phenomenon has not been reported for any other glutenins.

Another important characteristic of the wheat grain is its endosperm texture, referred to as hardness. This physical property is the most important one of the wheat grain for determining its end-use properties. For example, flour from hard wheats is preferred for bread making while flour from soft wheats is ideal for manufacturing cookies and cakes (Tipples et al. 1994; Bhave and Morris 2008a). In contrast, semolina obtained from very hard wheat such as durum is best suited for pasta. Hard and soft endosperm textures are primarily starch-related properties. However, puroindolines, members of small lipid-binding proteins seem to play a key role in determining hardness. Soft-textured grains require less grinding energy than hard-textured grains during the milling process to derive flour from endosperm. As a consequence, comparatively smaller proportion of starch granules become physically damaged. Undamaged starch granules absorb less water than damaged granules and consequently soft wheat flour has lower water absorption capacity than hard wheat flour. Puroindoline-a (*Pina*) and puroindoline-b (*Pinb*) genes are located at the *Ha* locus on the chromosome 5 homoeologues, except in the 5A and 5B of polyploid wheat where they are absent. Also located at this locus are the grain softness protein (*Gsp-1*) genes, which are present in the homoeologues of 5A and 5B in polyploid wheat. These genes also encode lipid binding proteins but do not seem to play a role in grain hardness (Bhave and Morris 2008a).

Transposable elements, first characterized by McClintock while studying varigated kernel colour (McClintock 1956), were suggested as important players in the evolution of the architecture and function of the eukaryotic genomes (Wessler 2006a and 2006b, Lynch 2007). Retrotransposons are the most abundant class of transposable elements in plant genomes such as maize and wheat (SanMiguel et al. 1998; Moolhuijzen et al. 2007). Transposable elements drive chromosomal rearrangements by mechanisms of unequal homologous recombination and illegitimate recombination because of the presence of repeat domains such as LTRs and short tandem repeats distributed across the genome in high copy numbers (Bennetzen 2005). A body of evidence from sequence organization studies indicated significant correlation between the rearrangements/transposable element insertions and phenotypic changes, with implications in biological adaptations and shaping of agriculturally important loci (Jiang et al. 2004; Chantret et al. 2005; Wicker et al. 2007a, Gao et al. 2007).

The main objective of the project is to understand the structural genomic organization of two major loci namely, *Glu-B1* and *Ha*. Both loci comprise several genes and have undergone minor and major chromosomal rearrangements with impacts on phenotypes related to wheat quality.

1.2 Objectives

1.2.1 (A)

Tracing the evolutionary history of the *Glu-B1* locus of *Triticum aestivum* cv Glenlea which harbors a 10.3 kb duplication encompassing the gene encoding the HMW glutenin Bx7 subunit, by assessing the genomic organization of the locus in a number of *Triticum* accessions representing geographical and genetic diversity including diploid and tetraploid ancestral species.

1.2.1 (B)

Characterization of the family members of the intergene retrotransposon (Sasanda_EU157184-1) present at the *Glu-B1* locus in the genome of *Triticum aestivum* cv Glenlea.

1.2.2

Understanding the genomic organization of the homoeologous *Ha* loci that led to the absence of the *Pina* and *Pinb* genes in the A- and B-genomes and the *Pina* in the D-genome of hexaploid wheat cv Glenlea, with implications into the evolution of the wheat genomes (A, B and D).

CHAPTER 2

LITERATURE REVIEW

2.1 Wheat genome evolution

Wheat is an allohexaploid (2n=6x=42) with homoeologous A, B, D genomes having a basic chromosome number of seven. Available evidence from the archaeological, phytogeographical and molecular data indicate that bread wheat originated in the Fertile Crescent of the Near East, where domesticated tetraploid wheat (*Triticum durum* ssp *dicoccum*) overlapped with the natural habitat of diploid goat grass *Aegilops tauschii*, the D genome donor of hexaploid wheat, favouring intergeneric hybridization ~ 8000 years ago (Fig. 2.1, Feldman 2001; Salamini et al. 2002; Huang et al. 2002). The presence of more than one shared genome organization at some loci among accessions of *T. turgidum* and *T. aestivum* suggested the contribution of more than one tetraploid ancestor in the evolution of hexaploid wheat. Also, the presence of more than one haplotype at orthologous positions, among the D genomes of *Ae. tauschii* and *T. aestivum* indicate multiple origins for the genome of bread wheat, from independent crosses involving different genotypes of its progenitors (Dvorak et al. 1998; Talbert et al. 1998; Feldman 2001; Caldwell et al. 2004, Giles and Brown 2006, Gu et al. 2006).

2.2 Colinearity among the grass genomes

Comparison of genetic maps of the grass genomes revealed conservation of blocks of markers (macrocolinearity) and sequences (microcolinearity), indicating their shared

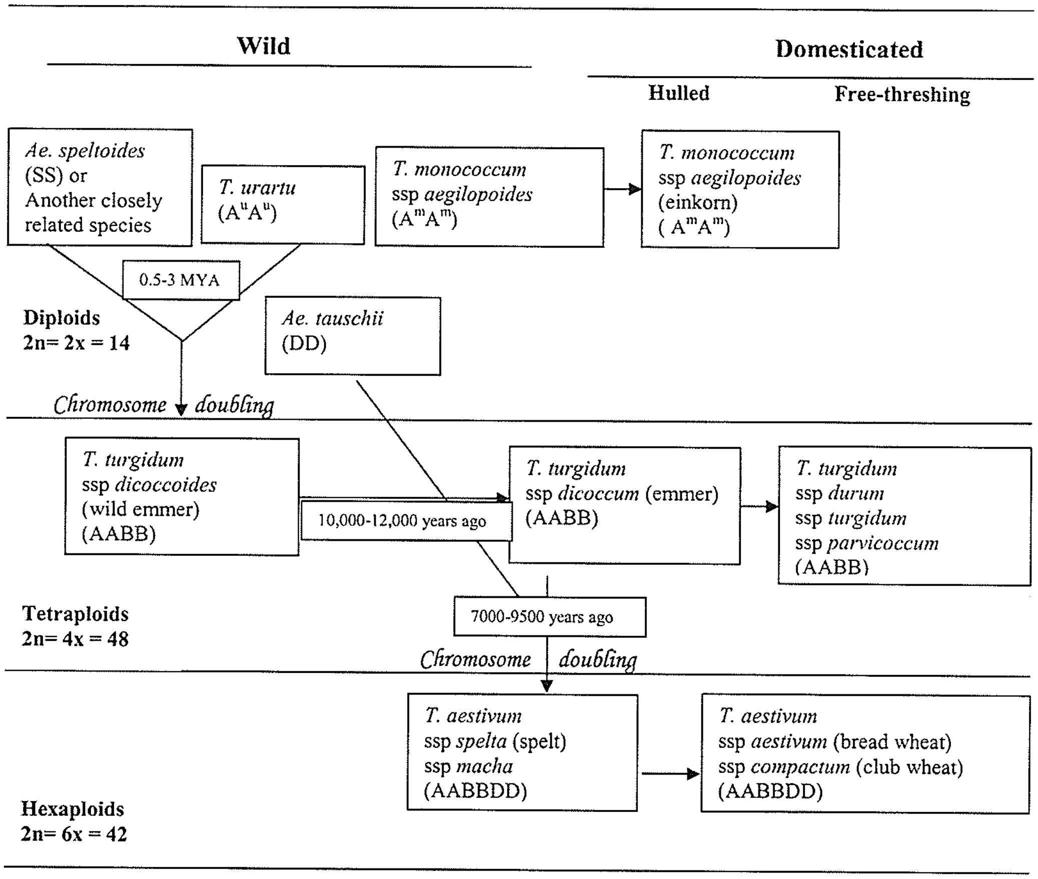

Wild
Domesticated
Hulled
Free-threshing
Ae. speltoides (SS) or Another closely related species
T. urartu (A^uA^u)
T. monococcum ssp aegilopoides (A^mA^m)
T. monococcum ssp aegilopoides (einkorn) (A^mA^m)
0.5-3 MYA
Diploids 2n= 2x = 14
Ae. tauschii (DD)
Chromosome doubling
T. turgidum ssp dicoccoides (wild emmer) (AABB)
10,000-12,000 years ago
T. turgidum ssp dicoccum (emmer) (AABB)
T. turgidum ssp durum ssp turgidum ssp parvicoccum (AABB)
Tetraploids 2n= 4x = 48
7000-9500 years ago
Chromosome doubling
T. aestivum ssp spelta (spelt) ssp macha (AABBDD)
T. aestivum ssp aestivum (bread wheat) ssp compactum (club wheat) (AABBDD)
Hexaploids 2n= 6x = 42

Figure 2.1 Evolution of the diploid, tetraploid and hexaploid wheat genomes (Eckardt 2001) © American Society of Plant Biologists; reprinted with permission.

ancestry and supporting the theory of the diversification of the grass genomes from a common ancestor 60 million years ago (Moore et al. 1995; Gale and Devos 1998; Keller and Feuillet 2000; Kellogg 2001; Tang et al. 2008). However, extensive comparative studies at the DNA sequence level among rice, sorghum, maize, wheat and barley revealed that there are many exceptions to the microcolinearity and mosaic patterns of conserved sequences (genes) amidst non-conserved sequences (intergenic regions) (Bennetzen 2000a; Song et al. 2002; Sorrels et al. 2003; Bennetzen and Ma 2003; Paterson et al. 2003; Bennetzen 2007).

2.3 Glutenins and their genetic control

High molecular weight glutenins (HMW-GS) are the most important fractions of the gluten. They are encoded at the *Glu-1* loci on the long arms of the homoeologous chromosomes 1A, 1B and 1D (Payne et al. 1980, 1984, 1987). These loci are designated as *Glu-A1, Glu-B1* and *Glu-D1,* respectively. Each *Glu-1* locus contains two tightly linked paralogous genes encoding two different types of HMW-GS, namely x- and y-type subunits (Payne et al. 1981; Payne 1987; Shewry et al. 1992). Multiple alleles are identified at the *Glu-1* locus with three, eleven and six allelic forms so far described for *Glu-A1*, *Glu-B1* and *Glu-D1,* respectively (Li et al. 2006). Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) is used for profiling HMW-GS in different wheat lines.

In bread wheat, most cultivars express 3-5 subunits due to the silencing of some genes (Payne 1987). It appears that the gene encoding the Ay subunit is always silent. Not all subunits have the same effect on quality. It has been demonstrated that good bread making quality is associated with the presence of *Glu-D1-1d+Glu-D1-2b* encoding subunits Dx5 and Dy10, respectively (Payne et al. 1987; Radovanovic et al. 2002). Studies of the coding regions of the *Glu-1* alleles and its immediate upstream region have been carried out (Table 2.1). However, the comparative understanding of the regions covering more distal portions (~100 kb) of orthologues of HMW glutenin loci from diploid, tetraploid and hexaploid wheats emerged only very recently (Kong et al. 2004; Gu et al. 2006).

Table 2.1 List of orthologues of HMW glutenin loci whose sequences are available

Gene/Alleles	Accession number	Species	Cultivar	Reference
Glu-A1	-	*T. aestivum*	Cheyenne	Forde et al. 1985
Glu-D1	-	*T. aestivum*	Chinese Spring	Thompson et al.1985
Glu-D1-1a	-	*T. aestivum*	Chinese Spring	Sugiyama et al. 1985
Glu-B1-2b	-	*T. aestivum*	-	Halford et al. 1987
Glu-D1-1b and *Glu -D1-2b*	X12928	*T. aestivum*	Cheyenne	Anderson et al. 1989
Glu -A1-1b and *Glu- B1-1a*	X12929	*T. aestivum*	Cheyenne	Anderson et al. 1989
Glu-A1-1b	X61009	*T. aestivum*	Hope	Halford et al. 1992
Glu -D1-2	U39229	*T. tauschii*	-	Mackie et al. 1996
Glu-B1-1e	AJ437000	*T. durum*	Lira	Shewry et al. 2003
Glu D1-1 and *Glu-D1-2*	AF497474	*T. tauschii*	-	Anderson et al. 2003
Glu -B1-1	AY368673	*T. turgidum*	-	Kong et al. 2004
Glu- D^{t}*1* (a novel variant of *Glu- D1-2a*)	AY248704	*Ae. tauschii*	-	Yan et al. 2004
Glu-A1 *Glu-B1* *Glu-D1*	DQ537335 DQ537336 DQ537337	*T. aestivum*	Renan	Gu et al. 2006

2.4 The molecular basis of overexpression of the Bx7 gene in Glenlea

Certain wheat cultivars, such as Glenlea, Bluesky, Wildcat, Glenavon and Burnside, belonging to the Canada Western Extra Strong market class (CWES), are known to produce very strong and extensible dough. These exceptional dough properties enable CWES cultivars to be blended with wheats of lesser quality as well as being suited to the manufacturing of frozen dough products. Allelic variation at the locus encoding HMW-GS subunit Bx7 and their association with overexpression was detected quantitatively by reversed phase high performance liquid chromatography (Marchylo et al. 1992). In this study, the proportion of subunit Bx7 variant relative to the total amount of HMW glutenin subunit was significantly higher (41.2± 1.5%) than the proportion of the subunit Bx7 relative to the total amount of HMW glutenin subunit (27.1± 0.9%). Southern blot analysis in Red River 68, another cultivar overexpressing the Bx7 subunit, led to the hypothesis that the overexpression might be the result of a gene duplication at the *Glu-B1* locus (D'Ovidio et al. 1997).

The significant contribution of overexpressing Bx7 subunit towards the genetic variance for dough characteristics has been documented (Radovanovic et al. 2002). Moreover, the association between overexpression of the Bx7 subunit (allele *Glu-B1al*) and quality measurements has been observed in a recent study of a DH population derived from a cross involving a Glenlea-derived breeding line and another genotype with poor bread-making quality (Radovanovic and Cloutier 2003). In this study, sequence divergence in the promoter region of the parental genotype over expressing Bx7 (presence of a 43bp insertion) was exploited to develop dominant and codominant PCR markers used for the identification of *Glu-B1al* allele. These PCR markers were also validated by SDS-PAGE where overproduction of the Bx7 subunit was visualized by a relatively higher intensity band in a profile generated from the total protein fraction of the endosperm. Aside from the 43 bp insertion in the promoter region, the gene encoding the overexpressed Bx7 is also characterized by an 18 bp insertion in the coding region corresponding to an extra hexapeptide motif (Ma et al. 2003). Butow et al. (2004) demonstrated that the two insertions were not always correlated with the overexpression of Bx7 because some accessions displayed one or both of the genotypes but did not over express the Bx7 subunit. They too hypothesized gene duplication as the reason for the overexpression.

Recently, gene duplication has been confirmed by physical mapping and sequencing of this *Glu-B1* locus of *Triticum aestivum* cv Glenlea (Cloutier et al. 2005). The structural organization of the locus revealed the presence of an LTR retrotransposon between the duplicated segments. Further characterization using BLASTn search of the TREP database (http://wheat.pw.usda.gov/ITMI/repeats) revealed that it was somewhat similar to a Tar-1 retrotransposon. The similarity was however lower than 80% defining this element as novel. We named this *copia* LTR retroelement Sasanda_EU157184-1. Nucleotide sequence changes in the two long terminal repeats (LTRs) of individual retrotransposons are used to date their time of insertion (SanMiguel et al. 1998). Absence of transitions or transversions in the LTRs of this new element precluded the calculation of the time of its insertion. However, the presence of three point mutations in the duplicated segments has indicated the time of insertion to be approximately 15000 ±11000 years ago (see chapter 3). Interestingly, studies to establish the timeline of wheat evolution based on gene sequence comparisons revealed that wheat hexaploidization occurred approximately 8000 years ago (Huang et al. 2002). Hence, there was room for speculation that the gene duplication may correspond to a post-hexaploidization event.

2.5 Endosperm texture and its genetic basis

Kernel hardness, defined as a measure of the resistance to deformation, is measured in terms of particle size index (*Psi*) and/or hardness index (HI). The proportion (%) of flour that passes through a standard sieve in a standard time is referred to as particle size index (Turnbull and Rahman 2002). *Psi* values decrease with increasing kernel harness. Hardness index is determined by single kernel characterization system (SKCS) as a measure of the physical force required to crush the kernel, with consideration for other parameters such as the moisture content and the size of the kernel. In SKCS, soft wheat cultivars have lower HI values than hard wheat cultivars.

In hexaploid wheat, endosperm texture is controlled primarily by a single locus on the short arm of chromosome 5D (Mattern et al. 1973; Law et al. 1978; Sourdille et al. 1996). The *Ha* locus comprises *Gsp-1*, *Pina* and *Pinb* genes encoding the grain softness protein (GSP-1), puroindoline-a (PINA) and puroindoline-b (PINB), respectively (Morris 2002). Though this main locus is referred to as the hardness (*Ha*) locus, softness is the dominant trait. Differences in endosperm softness are mainly due to differences in the strength of the adhesion between the starch granule surface and the surrounding protein matrix (Pomeranz and Williams 1990). Investigations provided an indication that a 15 kDa starchy-surface-associated protein complex named friabilin is present at relatively high levels on soft wheat starches (Greenwell and Schofield 1986; Jolly et al. 1993). The expression of friabilin in the endosperm is dependent on the short arm of chromosome 5D harbouring the *Ha* locus, indicating that this protein complex may be related to the *Ha* locus. Friabilin is a complex of related proteins that include puroindoline-a, puroindoline-b, and the grain softness protein. cDNAs encoding these proteins have been characterized and mapped to the distal part of chromosome 5DS (Gautier et al. 1994; Rahman et al. 1994). Genetic complementation studies also showed that *Pina* and *Pinb* genes are the major genes responsible for endosperm hardness/softness texture (Beecher et al. 2002, Martin et al. 2006, Krishnamurthy and Giroux 2001). Deletion of puroindoline gene(s) or substitution mutations in the coding domain results in hard endosperm texture (Morris 2002). Multiple alleles have been characterized for all three of the *Ha* locus genes, each associated with varying degrees of hardness (Bhave and Morris 2008b).

Kernel texture is not an 'all-or nothing' trait, but rather it occurs as a continuum of textural differences, from very hard to very soft. Though the *Ha* locus is the primary genetic determinant of the softness phenotype, other factors contribute towards the substantial differences found among cultivars in the degree of relative softness or hardness. For example, pentosan content, which comprises 2-3% of the wheat flour, positively influences hardness (Hong et al 1989). Similarly, an increase in starch granule associated free lipid is positively correlated with increasing hardness (Morrison et al 1989). Environmental factors can also play a role. For example, the rate at which cytoplasm dries during the grain maturity stage directly determines the extent of damage to the membranes of the endosperm cells, which, in turn, determines the amount of air-spaces in the endosperm. More air-spaces result in softer grain (Dexter et al 1989).

The *Pina and Pinb* sequences have been detected on the homoeologues of more than 200 diploid accessions of wheat and *Aegilops*, but only on the chromosome 5D of hexaploid wheat (*T. aestivum*) indicating its conservation in diploids and deletion in polyploid species (Gautier et al. 2000; Lillemo et al. 2002; Simeone et al. 2006; Li et al. 2008). The only exceptions reported to date include the A-genome of *T. timopheevi* (AAGG) and the A^m genome of *T. zhukovskyi* (A^mA^mAAGG) (Li et al. 2008).

In barley, an orthologue of the *Ha* locus has been identified on chromosome 5H which harbors the hordoindoline genes *Hina* and *Hinb* and *Gsp* (Beecher et al 2001, Darlington et al. 2001). Though hordoindolines are similar to puroindolines of wheat, their role in kernel texture has not been proven unequivocally (Bhave and Morris 2008a). Besides, unlike wheat, ancestors of barley including *Hordeum spontaneum* are hard textured. On the other hand, rye has a very soft kernel texture with lesser variability for the trait compared to wheat (Simeone and Lafiandra 2005). Secaloindoline-a (*Sina*), secaloindoline-b (*Sinb)* and grain softness protein (*Gsp-R1*) genes were mapped to the rye chromosome 5R which is orthologous to the chromosome 5D of wheat (Jolly et al 1993, Gautier et al 2000, Simeone and Lafiandra 2005, Massa and Morris 2006). Rye accessions with harder kernel phenotypes have not been characterized to date (Bhave and Morris 2008). Orthologues of puroindoline genes have also been found in oats (avenoindolines, Tanchak et al. 1998).

Physical mapping of the *Ha* locus from *Ae. tauschii* (D genome) and its orthologue from *T. monococcum* (A^m genome) as well as sequencing of the latter has been carried out (Tranquilli et al. 1999; Turnbull et al. 2003; Chantret et al. 2004). Genomic organization of the *Ha* locus in a hexaploid wheat genotype (Chantret et al. 2005) has been elucidated only recently.

The development of BAC libraries, high efficiency content fingerprinting, wheat cytogenetic stocks, fine genetic maps, advances in sequencing technologies will facilitate understanding of the structural genomic organization in the future.

2.6 Wheat genome organization

In wheat, cytogenetic resources such as aneuploid stocks complemented with genomic tools such as ESTs and genomic DNA sequences have given some insights into its genome organization. Deletion lines (Gill et al. 1996a; Gill et al. 1996b; Endo and Gill 1996; Qi et al. 2003) and localization of ESTs in relation to chromosome bins demarcated by deletion breakpoints revealed an uneven distribution of genes, indicating the possible presence of gene islands (Sidhu and Gill 2004; Qi et al. 2004; Dilbirligi et al. 2004). In a large scale study, physical mapping of 3025 loci including 17 QTLs and 252 characterized genes and using 334 deletion lines identified 48 gene rich regions (GRRs), of which five spanned ~3% of the genome but contained 26% of the genes (Erayman et al. 2004).

Also, studies of sequence organization at the regions harbouring some agronomically important loci such as *Lr, Gli-1*, *Glu-3*, *Glu-1* and *Ha* (Table 2.2) have shown the presence of gene islands with few reported exceptions (Devos et al. 2005). With an estimate of ~5000 Mb per wheat genome and 30,000 genes, an average density of one gene per 166 kb is expected, providing even distribution of the genes (Stein 2007). However, as observed in *Arabidopsis*, a high gene density (one gene per 5-20 kb) was estimated for most of the genomic regions sequenced so far (Table 2.2) with a maximum gene density of one gene per 3.8 kb (Gao et al. 2007). In contrast, when wheat BACs were randomly chosen a gene density as low as one gene per 75 kb (Devos et al. 2005) and one gene per 168 kb have been estimated (Stein 2007). Studies are underway to arrive at an estimate which may accurately represent the gene distribution in *Triticum* genome (Devos, http://www.cropsoil.uga.edu/faculty1/devosprojects.htm). Until then the current model of uneven gene distribution with gene islands (with high gene density) amid vast stretches of repetitive DNA is generally accepted. Also, a body of evidence from sequence organization studies confirmed the conservation of gene order and orientation among orthologous and homoeologous loci, with occasional violations (Table 2.2). However, expansion or contraction and diversification of intergenic regions resulting from the dynamics of transposable elements were common in all investigations. Both mechanisms, namely unequal homologous recombination and illegitimate recombination driven by dispersed repeats such as LTRs and other short sequences, were identified as the main forces of sequence rearrangements. The numerous deletions engineered by the above mechanisms counteracted genome expansion from retroelement amplification indicating a dynamic rather than static genome.

Table 2.2 Studies of genome organization of regions encompassing important genes in wheat and its progenitor species

Loci studied	**Species**	**Genome**	**Accession number**	**Length of the BAC sequence (bp)**	**References**
Receptor-like kinase genes	*T. aestivum* cv ThatcherLr10	$AABB^{\#}DD^{\#}$	AF325196 AF325197 AF325198	35,872 20,754 43,606	Feuillet et al. 2001
Disease resistance genes (LZ-NBS-LRR class)	*Ae. tauschii*	DD	AF446141	106,618	Brooks et al. 2002
Lr21	*Ae. tauschii*	DD	AF532104	27,960	Huang et al. 2003
Lr10	*T. monococcum*	$A^{m}A^{m}$	AF326781	211009	Wicker et al. 2001
Lr1	*T. aestivum* cv Glenlea	$AABBDD^{\#}$	EF567062	137,614	Cloutier et al. 2007
Glu-3 and Pm3	*T. turgidum* ssp *durum* cv Langdon65	$AA^{\#}BB$	AY146587	258,179	Wicker et al. 2003
Glu-3 and Pm3	*T. monococcum*	$A^{m}A^{m}$	AY146588	285,444	Wicker et al. 2003
Pm3	*T. aestivum* cv Chinese Spring	$AA^{\#}BBDD$	DQ251490	8,778	Wicker et al. 2007a
Prolamin genes (Gli-1 and Glu-3)	*T. turgidum* ssp *durum* cv Langdon65	$AA^{\#}BB$	EF426564	139,403	Gao et al. 2007
Prolamin genes (Gli-1 and Glu-3)	*T. turgidum* ssp *durum* cv Langdon65	$AABB^{\#}$	EF426565	157,918	Gao et al. 2007
Glu-1	*Ae. tauschii*	DD	AF497474	102,842	Anderson et al. 2003
Glu-1	*T. turgidum* ssp *durum* cv Langdon65	$AA^{\#}BB$	AY494981	307,015	Gu et al. 2004
Glu-1	*T. turgidum* ssp *durum* cv Langdon65	$AABB^{\#}$	AY368673	285,506	Kong et al. 2004
Glu-1	*T. aestivum* cv Glenlea	$AABB^{\#}DD$	EU_157184	53,205	Cloutier et al. 2005
Glu-1	*T. aestivum* cv Renan	$AA^{\#}BBDD$	DQ537,335	292,102	Gu et al. 2006

Loci studied	Species	Genome	Accession number	Length of the BAC sequence (bp)	References
Glu-1	*T. aestivum* cv Renan	$AABB^{\#}DD$	DQ537336	206,063	Gu et al. 2006
Glu-1	*T. aestivum* cv Renan	$AABBDD^{\#}$	DQ537337	152,010	Gu et al. 2006
VRN1	*T. monococcum*	A^mA^m	AY188331 AY188332 AY188333	133,625 95,541 112,328	Yan et al. 2003
VRN2	*T. monococcum*	A^mA^m	AY485644	438,828	Yan et al. 2004
-	*T. monococcum*	A^mA^m	AF459639	215,241	SanMiguel et al. 2002
Ha	*T. monococcum*	A^mA^m	AY491681	101,101	Chantret et al. 2004
Ha	*Ae. tauschii*	$DD^{\#}$	CR626926	94,421	Chantret et al. 2005
Ha	*T. turgidum* ssp *durum* cv Langdon65	$AA^{\#}BB$	CR626933	25,216	Chantret et al. 2005
Ha	*T. turgidum* ssp *durum* cv Langdon65	$AABB^{\#}$	CR626932	19,229	Chantret et al. 2005
Ha	*T. aestivum* cv Renan	$AA^{\#}BBDD$	CR626929	20,745	Chantret et al. 2005
Ha	*T. aestivum* cv Renan	$AABB^{\#}DD$	CR626930	19,274	Chantret et al. 2005
Ha	*T. aestivum* cv Renan	$AABBDD^{\#}$	CR626934	94,398	Chantret et al. 2005

# Genomic origin of the target locus

2.7 Mobile genetic elements and their contribution to the evolution of genes and genomes

Mobile genetic elements are fragments of DNA with the ability to move from one location to another in a genome and are grouped into two classes based on the mode of transposition (Kumar and Bennetzen 1999). Class I elements spread via RNA intermediaries and class II elements move in their native (DNA) form. They were first characterized in maize as the causal agents of changes in kernel pigmentation (McClintock 1950). Transposable elements have played a crucial role in the evolution of genes and genomes by contributing to the dynamics of size and structure of the genome (Kidwell 2002; Feschotte et al. 2002; Wessler 2006a and 2006b).

Amplification of retroelements in periodic bursts led to the increase in genome sizes of maize, barley and wheat (San Miguel et al. 1998; Kalendar et al. 2000; Wicker et al 2001; Vitte and Panaud 2005). Piegu et al. (2006) reported that the doubling of the genome size in *Oryza australiensis* (965 Mb) compared to its domesticated relative *Oryza sativa* (390 Mb) occurred mainly from the accumulation of thousands of copies of RIRE1, Kangourou and Wallabi retroelements in the last 3 million years.

In eukaryotes, the presence of abundant copies of mobile elements and their distribution across the genome drive structural rearrangements involving segments of varying sizes from a few to hundreds of genes (Coghlan et al. 2005; Bennetzen 2005). The presence of repetitive sequences at recombination breakpoints suggested that homology dependent rearrangements are a very common cause of genomic instability, in addition to transposase mediated rearrangements (Coghlan et al. 2005). Phenotypic differences resulting from retroelement driven chromosomal rearrangements such as deletion (Jiang et al 2004), duplication (Morgante et al. 2005), inversion (Caceres et al. 1999) and translocation (Xiao et al. 2008) have been reported. Integration of reverse transcriptase (RT) sequences of ancient non-LTR retroelements as telomerases, which play a very crucial role in genome stability, has been suggested (Abad et al. 2004; Pardue et al. 2005).

On the gene evolution front, transduplication, a process of capture of fragments of host genes by transposable elements and their potential as a reservoir of novel genes by diversifying coding sequences, was reported (Bureau et al. 1994; Jiang et al. 2004; Juretic et al. 2005; Hoen et al. 2006). Also, origins of chimeric genes by retrotransposition were discovered in rice (Wang et al. 2006). Besides host gene expression changes resulting from the acquired promoter and other regulatory sequences from transposable elements (Bennetzen 2000b; Jordan et al. 2003; Han et al. 2004; Marino-Ramirez et al. 2005), generation of new allelic variants by insertion of elements into introns leading to alternate splicing sites (Varagona et al. 1992), tissue specific RNA processing (Marillonnet and Wessler 1997), exon shuffling (Morgante et al. 2005) and gene disruption (Harberd et al. 1987; Giovanni et al. 2008) were also documented. Recently, creation of intraspecific variation at allelic positions (haplotypes) among inbreds of maize, due to the movement of gene fragments by Helitrons was identified (Lai et al. 2005; Wang and Dooner 2006). Recruitment of coding sequences of transposable elements by the host genome for functions such as gene regulation has been discovered in many eukaryotes including flowering plants (Volff 2006). Also, silencing of adjacent host genes by the read-through transcript originating from the promoter sequences of LTR elements integrated in the antisense strand was documented (Kashkush et al. 2003; Puig et al. 2004). As well, epigenetic regulation of expression of neighboring genes by methylation upon integration of transposable elements was also described (Comai 2000; Slotkin and Martienssen 2007).

Evolutionary origin of the segmental duplication encompassing the wheat *GLU-B1* locus encoding the overexpressed Bx7 (Bx7^{OE}) high molecular weight glutenin subunit.

Raja Ragupathy[1],[2] • Hamid A. Naeem[3] • Elsa Reimer[2] • Odean M. Lukow[2] • Harry D. Sapirstein[3] • Sylvie Cloutier[2]

[1]Department of Plant Science, Faculty of Graduate Studies,
University of Manitoba,
Winnipeg, Manitoba, Canada R3T 2N2

[2]Cereal Research Centre, Agriculture and Agri-Food Canada,
195 Dafoe Road,
Winnipeg, Manitoba, Canada R3T 2M9

[3]Department of Food Science, Faculty of Graduate Studies,
University of Manitoba,
Winnipeg, Manitoba, Canada R3T 2N2

The thesis author, Raja Ragupathy, designed, carried out the experiments, did the data analysis, interpretation and drafted the manuscript. As major advisor, Dr. Sylvie Cloutier guided the direction of the study, participated in data analysis and manuscript review. Drs. Sapirstein and Naeem contributed with their newly developed rapid RP-HPLC methodology and helped in data interpretation. Dr. Odean Lukow and Ms. Elsa Reimer helped in data collection and interpretation of SDS-PAGE and DNA markers (18-bp and 43-bp indels) respectively.

(The following chapter was published in Theoretical and Applied Genetics 2008, 116: 283-296)

CHAPTER 3

Evolutionary Origin of the Segmental Duplication Encompassing the Wheat *GLU-B1* Locus Encoding the Overexpressed Bx7 (Bx7^{OE}) High Molecular Weight Glutenin Subunit

3.1 Abstract

Sequencing of a BAC clone encompassing the *Glu-B1* locus in Glenlea, revealed a 10.3 Kb segmental duplication including the *Bx7* gene and flanking an LTR retroelement. To better understand the evolution of this locus, two collections of wheat were surveyed. The first consisted of 96 diploid and tetraploid species accessions while the second consisted of 316 *Triticum aestivum* cultivars and landraces from 41 countries. The genotypes were first characterized by SDS-PAGE and a total of 40 of the 316 *T. aestivum* accessions were found to display the overexpressed Bx7 phenotype (Bx7^{OE}). Three lines from the 96 diploid/tetraploid collection also displayed the stronger intensity staining characteristic of the Bx7^{OE} subunit. The relative amounts of the Bx7 subunit to total HMW-GS were quantified by RP-HPLC for all Bx7^{OE} accessions and a number of checks. The entire collection was assessed for the presence of four DNA markers namely an 18 bp indel of the coding region of *Bx7* variant alleles, a 43 bp indel of the 5'-region and the left and right junctions of the LTR retrotransposon borders and the duplicated segment. All 43 accessions found to have the Bx7^{OE} subunit by SDS-PAGE and RP-HPLC produced the four diagnostic PCR amplicons. None of the lines without the Bx7^{OE} had the LTR retroelement/duplication genomic structure. However, the 18 bp and 43 bp indel were found in accessions other than Bx7^{OE}. These results indicate that the overexpression of

the Bx7 HMW-GS is likely the result of a single event, i.e., a gene duplication at the *Glu-B1* locus mediated by the insertion of a retroelement. Also, the 18 bp and 43 bp indels pre-date the duplication event. Allelic variants *Bx7**, *Bx7* with and without 43 bp insert and *Bx7*^{OE} were found in both tetraploid and hexaploid collections and shared the same genomic organization. Though the possibility of introgression from *T. aestivum* to *T. turgidum* can not ruled out, the three structural genomic changes of the B-genome taken together support the hypothesis of multiple polyploidization events involving different tetraploid progenitors.

3.2 Introduction

Wheat flour has the unique ability to form dough that exhibits the rheological properties required for the production of leavened bread and for the wider diversity of foods that have been developed taking advantage of this attribute (Weegels et al. 1996). Storage proteins of the endosperm, namely the gluten complex, primarily determine bread making ability (Shewry et al. 1992). Gluten content and composition are critical in providing the rheological properties of flour.

High molecular weight glutenin subunits (HMW-GS) are the most important fractions of the gluten because they form large polymeric structures through disulphide bonds (Wrigley 1996). These polymeric structures are related to the molecular properties of dough and their composition alone may account for 47-60% variation in bread making quality of wheat (Payne 1987; Rakszegi et al. 2005). Both qualitative and quantitative effects of individual subunits on bread making quality and dough functionality are important (Barro et al. 1997). The presence of the Dx5 and Dy10 subunit combination is associated with good quality (Payne 1987; Radovanovic et al. 2002). Similarly, the significant contribution of overexpressing Bx7 subunit towards the genetic variance for mixing characteristics important to dough strength has been reported (Butow et al. 2003; Radovanovic et al. 2002; Vawser and Cornish 2004). The dough strength, in turn, determines the suitability of the flour for bread making and other end uses (Bushuk 1998; Gale 2005; Lukow et al. 1992).

HMW-GSs are encoded at the *Glu-1* loci on the long arms of homoeologous chromosomes 1A, 1B and 1D (Payne et al. 1987) and are designated *Glu-A1, Glu-B1* and *Glu-D1*, respectively. Each *Glu-1* locus contains two tightly linked paralogous genes encoding two different types of HMW-GS, namely the x- and y-type subunits (Payne et al. 1981; Shewry et al. 1992). In bread wheat, most cultivars do not express the expected six HMW-GS but usually three to five subunits due to the silencing of some genes. The gene encoding the Ay subunit in hexaploid wheat is always silent. Each of these complex loci displays extensive allelic variation (Payne and Lawrence 1983).

Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) is used for the separation and identification of HMW-GS in different wheat lines where overproduction of the Bx7 subunit is visualized by a relatively higher intensity staining in a profile generated from the total protein fraction of the endosperm (Lukow et al. 1989). Allelic variation at the locus encoding HMW-GS subunit Bx7 and overexpression was detected quantitatively by reversed phase high performance liquid chromatography (RP-HPLC, Marchylo et al. 1992). In this study, the proportion of subunit Bx7 relative to the total amount of HMW glutenin subunits was significantly higher (41.2 ± 1.5%) in Bx7^{OE} lines than it was in lines not overexpressing the Bx7 subunit (27.1 ± 0.9%). To our knowledge, this is the only HMW glutenin subunit to display the overexpression phenotype.

The presence of two functional copies of the gene encoding the HMW-GS Bx7 subunit and improved transcriptional and/or translational efficiency have been proposed to explain the overexpression of this subunit in certain accessions (Lukow et al. 2002). Southern analysis in TAA 36, a landrace from Israel suggested that the gene encoding subunit Bx7 is present in two copies and their simultaneous expression lead to the overproduction of the subunit (Lukow et al. 1992). In the same study, however, the authors suggested that the mechanism responsible for the Bx7^{OE} phenotype of cultivar Glenlea differed and was attributed to the increased efficiency of transcription caused by differences in the promoter sequence and/or translation. Southern analysis of 'Red River 68', another cultivar overexpressing the Bx7 subunit, revealed the presence of a stronger hybridization signal thereby supporting the gene duplication hypothesis (D'Ovidio et al. 1997). Recently, a BAC clone encompassing the *Glu-B1* locus of cultivar Glenlea was sequenced and a 10.3 Kb duplication including the gene encoding the Bx7 HMW-GS was identified (Cloutier et al. 2005). The structural organization of the locus revealed the presence of a LTR retrotransposon between the duplicated areas.

Transposable elements play an important role in the evolution of the structure, function and regulation of expression of genes and genomes in eukaryotes (Bennetzen 2000b; Grandbastien 1992; Kazarian 2004). They play a key role in evolution by driving structural changes such as duplication and deletion (Jiang et al. 2004; Morgante et al. 2005). Retrotransposons are the most abundant class of transposable elements in plant genomes such as maize and wheat (Kumar and Bennetzen 1999; SanMiguel and

Bennetzen 1998; Vitte and Panaud 2005). They mediate structural chromosomal rearrangements by unequal homologous recombination and illegitimate recombination because of their abundant copy number and distribution across the genome (Bennetzen 2005). This study aimed at understanding the evolutionary origin of the *Glu-B1* locus in accessions overexpressing subunit Bx7. Specifically, whether the origin of the tandem segmental duplication driven by a retroelement leading to two copies of gene encoding Bx7 subunit occurred prior or post advent of hexaploid wheat. An understanding of the structural changes of the *Glu-B1* locus in diploid, tetraploid and hexaploid wheat could provide some insight into the origin(s) of hexaploid wheat.

3.3 Materials and Methods

3.3.1 Plant materials

Germplasm was obtained from the Cereal Research Centre (CRC), the Plant Genetic Resources of Canada (PGRC), the United States Department of Agriculture-National Genetic Resources Program (USDA-NGRP), the International Wheat and Maize Improvement Centre (CIMMYT) and the European Cooperative Programme for Plant Genetic Resources (ECPGR) unit of the Research Institute of Crop Production (RICP). Two collections, one consisting of 96 diploid and tetraploid (*Aegilops* and *Turgidum*) accessions and the other of 316 *T. aestivum* cultivars and landraces from 41 countries were evaluated. The former collection comprised 6 *T. monococcum* (A^mA^m), 2 *Aegilops speltoides* (BB), 11 *Ae. squarrosa* (syn. *Ae. tauschii;* DD), 34 *T. turgidum* (AABB), 14 *T. dicoccoides*, 3 *T. durum*, 8 *T. dicoccum*, 12 *T. carthlicum*, 3 *T. turanicum* and 3 *T. polonicum* accessions.

3.3.2 SDS PAGE analysis

To assess the overexpression of the Bx7 subunit, SDS-PAGE analysis was carried out on single kernels using a HMW glutenin extraction procedure and Coomassie blue staining as previously described (Radovanovic and Cloutier 2003, Appendix I).

3.3.3 RP-HPLC analysis

Five seeds were taken and tested from each of the accessions and checks. Checks were selected for both tetraploid and hexaploid genotypes to have identical numbers of expressed subunits as the accessions tested. The embryo portion of the seeds was removed with a knife and the remaining endosperm and seed coat were crushed with a hammer and ground to a fine powder with a mortar and pestle. Extraction of insoluble glutenins and analysis of HMW-GS were done as described earlier (Naeem and Sapirstein 2007, Appendix II) except that only 30 mg of sample was used for initial extraction. Data were acquired and analyzed using Agilent ChemStation software (version 10.01). The elution profiles were used for the quantification of the Bx7 subunit relative to total HMW-GS.

3.3.4 PCR analyses

Plants were grown at CRC in a growth cabinet or in a greenhouse and genomic DNA was isolated from young leaf tissues using the Plant DNeasy 96 kit following manufacturer's instructions (Qiagen, Maryland, USA). Genomic DNA was quantified by fluorometry and diluted to 100 ng/μl.

Primers designed to amplify an 18 bp indel characteristic of the *Bx7* coding region were from Butow et al. (2004) but the forward primer was modified to include an M13 tail (Schuelke 2000) for subsequent resolution on an ABI 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA). Primers for the dominant marker corresponding to the 43 bp indel of the promoter region were from Radovanovic and Cloutier (2003). Primer pairs flanking the LTR retrotransposon borders and the duplicated region were designed at the left and right junctions of the retroelement. The left junction primers were: TaBAC1215C06-F517, 5'-ACGTGTCCAAGCTTTGGTTC-3' and TaBAC1215C06-R964, 5'-GATTGGTGGGTGGATACAGG-3'. The right junction primers were: TaBAC1215C06-F24671, 5'-CCACTTCCAAGGTGGGACTA and TaBAC1215C06-R25515, 5'-TGCCAACACAAAAGAAGCTG-3'.

PCR reactions for the 18 bp indel were performed in 10 μl using 75 ng of genomic DNA as template and otherwise as previously described in Schuelke (2000). A touchdown program starting with 2 min at 94°C for 20 s, 64°C for 45 s, 72°C for 1 min decreasing the annealing temperature by 1°C every cycle to 54°C, followed by 26 cycles at 54°C annealing and a final 5 min extension at 72°C and cooling to 4°C. The amplification products were resolved on an ABI 3100 Genetic Analyzer (Applied Biosystems, CA, USA). PCR conditions and electrophoresis for the 43 bp indel marker were as previously described with the exception that 10 pmole of primers and 10 μl reaction volumes were used (Radovanovic and Cloutier, 2003). PCR reactions for the retroelement right and left junction markers were performed in 25 μl containing 200 ng of genomic DNA, 10 pmole of each primer, 0.8 mM dNTPs, 1.5 mM $MgCl_2$, and one unit of Taq DNA polymerase. The cycling conditions for amplifying the left junction were 94°C for 5 min followed by 34 cycles of 94°C for 30 s, 63°C for 30 s and 72°C for 1 min followed by final extension and cooling as above. Conditions were identical for the right junction with the exception that the annealing temperature was decreased to 59°C. The PCR products were resolved on 1.5% agarose gels in 0.5X TBE buffer, stained with ethidium bromide and visualized under UV.

3.4 Results

3.4.1 SDS-PAGE analysis of wheat accessions

HMW-GS profiles of the two wheat collections were first obtained by SDS-PAGE (Fig. 3.1). Accessions that displayed either a Bx7, Bx7* or Bx7^{OE} were identified (Table 3.1). The presence of the Bx7^{OE} was evaluated by assessing the staining intensity of the subunit. Accessions without Bx7, Bx7* or Bx7^{OE} were classified as non-Bx7. Among the accessions surveyed, none of the diploids had a subunit Bx7 variant (Table 3.1). Among the 77 tetraploid accessions, 3 previously unreported *T. turgidum* (AABB) accessions (Branco, CN12222 and CN12225) were found to have the Bx7^{OE} subunit. These lines were from Portugal and the Czech Republic. In the hexaploid wheat collection, 219 of the 316 accessions were found to express a Bx7 subunit variant among which 40 lines, all belonging to the subspecies *aestivum,* displayed the Bx7^{OE} phenotype (Appendix III). The largest number of accessions (36) overexpressing the Bx7 subunit were found in the North and South American material. Three lines from Australia/New Zealand and one from Israel also displayed the overexpressing Bx7 phenotype. None of the 55 European and 12 African lines surveyed that had a Bx7 variant displayed the Bx7^{OE} subunit.

Relative overexpression of other HMW-GS was not observed in any of the accessions surveyed.

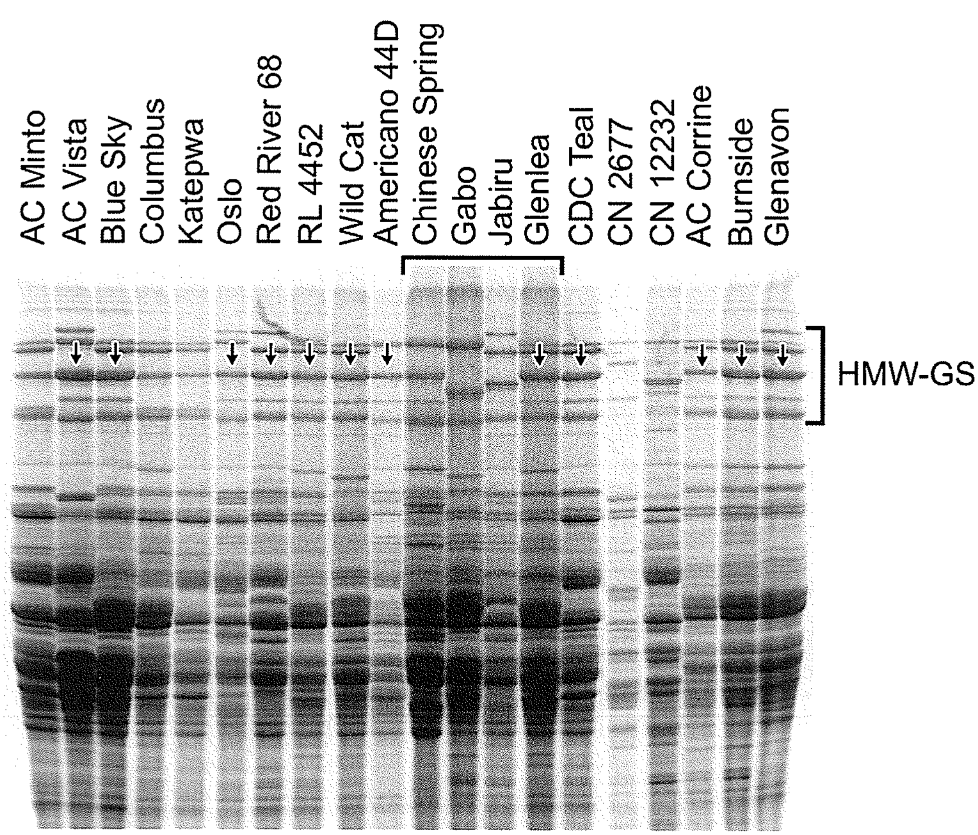


Figure 3.1 SDS-PAGE of HMW-GS of a subset of *Triticum* accessions. Cultivars Chinese Spring, Gabo, Jabiru and Glenlea were used as checks on each gel for the profiling of the HMW-GS. Arrows indicate the Bx7^{OE} subunit.

3.4.2 RP-HPLC analysis

The three *T. turgidum* accessions and the 40 *T. aestivum* accessions that displayed the Bx7^{OE} subunit were analysed by RP-HPLC along with 1 tetraploid check and 12 hexaploid checks (Appendix IV). To ensure that results on the proportion of Bx7 subunits to total HMW subunits were comparable between accessions and checks, these were chosen to have identical numbers of expressed subunits. All of the overexpressing accessions in our study had three and five expressed HMW-GS respectively for tetraploid and hexaploid accessions. The elution profiles were used for the quantification of the proportion of Bx7 subunit (% area) relative to the total amount of HMW-GS (Fig. 3.2). In tetraploid accessions, the proportion of the overexpressing Bx7 subunit relative to the total amount of HMW-GS averaged 66.2 ± 4.6% in the three lines and was higher than the check (57.0%). Similarly, among the hexaploid accessions, the proportion of Bx7^{OE} subunit (41.7 ± 2.5%) was significantly higher ($P < 0.01$) in comparison to the check cultivars with non-overexpressed Bx7 subunit variants (29.6 ± 0.9%) (Table 3.2).

Table 3.1 Survey of wheat accessions for HMW-GS composition at the *Glu-B1* locus as assessed by SDS-PAGE

	Number of accessions			
		HMW-GS at *Glu-B1x*		
Species (Genome)	**Surveyed**	**Non-Bx7**	**Bx7/ Bx7***	**Bx7^{OE}**
Diploid				
T. monococcum (A^mA^m)	6	6	0	0
Aegilops speltoides (SS)	2	2	0	0
Aegilops tauschii (DD)	11	11	0	0
Tetraploid (AABB)				
T. turgidum	31	15	13	3
T. turgidum subsp *turgidum*	3	2	1	0
T. turgidum subsp *dicoccoides*	14	13	1	0
T. turgidum subsp *durum*	3	3	0	0
T. turgidum subsp *dicoccum*	8	8	0	0
T. turgidum subsp *carthlicum*	12	1	11	0
T. turgidum subsp *turanicum*	3	2	1	0
T. turgidum subsp *polonicum*	3	3	0	0
Hexaploid (AABBDD)				
T. aestivum subsp *spelta*	4	2	2	0
T. aestivum subsp *compactum*	4	2	2	0
T. aestivum subsp *macha*	1	1	0	0
T. aestivum subsp *spherococcum*	7	5	2	0
T. aestivum subsp *aestivum*	300	87	173	40
Asia	84	17	66	1
Europe	78	23	55	0
North America	34	4	10	20
South America	51	14	21	16
Africa	29	17	12	0
Australia/ New Zealand	24	12	9	3

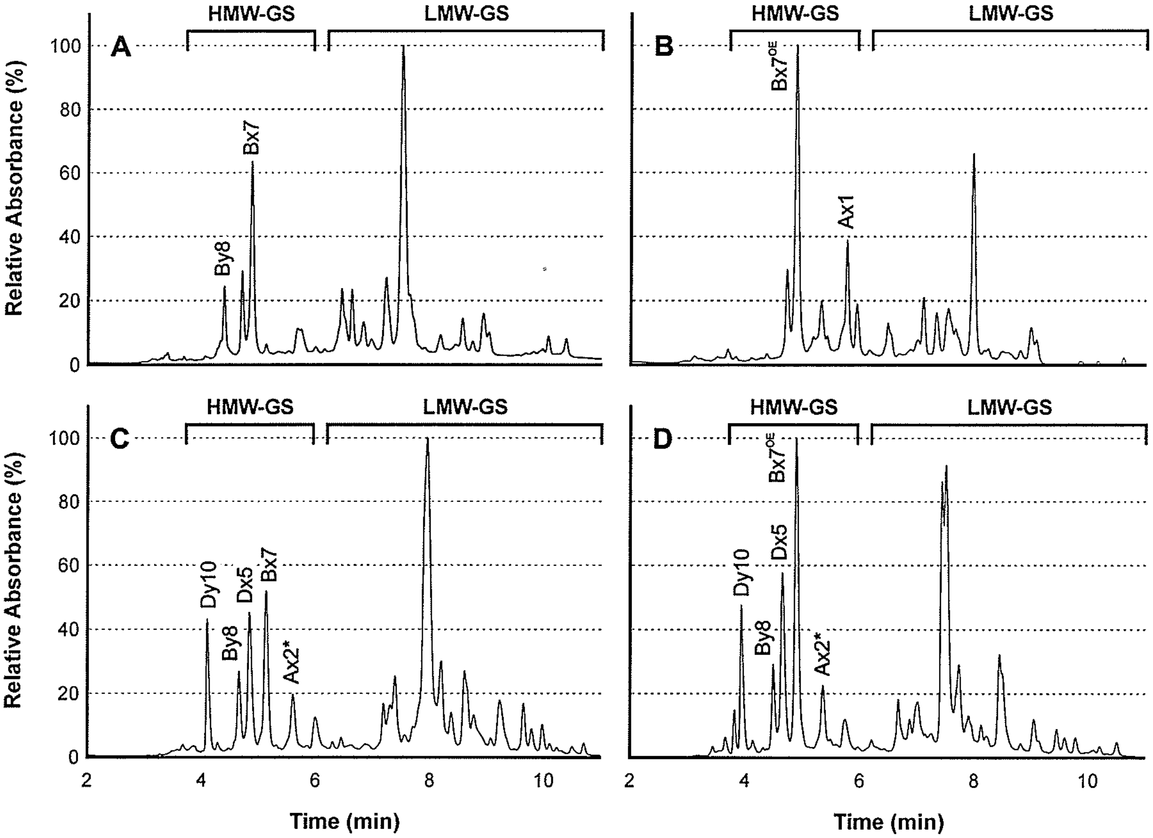


Figure 3.2 RP-HPLC profiles of four wheat accessions. **A.** *T. turgidum* check CN51263 has HMW-GS Bx7 that represented 57% of its total HMW-GS composition. **B.** *T. turgidum* cv Branco displayed HMW-GS Bx7^{OE} that represented 63.3 ± 4.7% of its total HMW-GS composition. **C.** *T. aestivum* check cv Klein Sin Rival has HMW-GS Bx7 that represented 30.4 ± 0.9% of its total HMW-GS composition. **D.** *T. aestivum* cv Glenlea displayed HMW-GS Bx7^{OE} that represented 38.40 ± 2.5 %of its total HMW-GS composition.

Table 3.2 Relative quantification of HMW-GS Bx7^{OE}, Bx7 and Bx7* by RP-HPLC

Species	**Number of accessions**	**HMW-GS at *Glu-B1x***	**Number of expressed HMW-GSs**	**Proportion of Bx7 subunit to total HMW-GS** [a]
T. aestivum subsp *aestivum*	40	Bx7^{OE}	5	41.7 ± 2.5
T. aestivum subsp *aestivum* (check)	12	Bx7/ Bx7*	5	29.6 ± 0.9
T. turgidum	3	Bx7 ^{OE}	3	66.2 ± 4.6
T turgidum (check)	1	Bx7	3	57.0

[a] Mean of duplicate injections ± standard deviation

3.4.3 Sequence characterization of the *Glu-B1* locus

The three main types of Bx7 variants, namely Bx7*, Bx7 and Bx7^{OE} are differentiated by the presence of an 18 bp indel in the repetitive domain corresponding to an extra hexapeptide motif in the Bx7 and Bx7 ^{OE} subunits (Radovanovic and Cloutier, 2003). The Bx7^{OE} subunit is however expressed at a higher level than the Bx7 subunit. A total of 27 of the 30 tetraploid accessions, including the 3 Bx7 ^{OE} accessions, displayed the 18 bp indel marker (Table 3.3). In the hexaploid collection, among the 219 accessions exhibiting a Bx7 variant, 57 Bx7 lines and all 40 lines with the Bx7 ^{OE} subunit were found to have the 18 bp indel marker. The allele encoding subunit *Bx7**, as characterized by the absence of the 18 bp indel, was found in three tetraploid and 120 hexaploid accessions. Hexaploid accession 01C0102607 was mixed (*Bx7/Bx7**) and CN9491 was not determined.

The presence of the 43 bp indel in the promoter region was found in four *T. turgidum* accessions including the three lines overexpressing the Bx7 subunit. Among the *T. aestivum* accessions, all 40 lines exhibiting the Bx7^{OE} subunits were found to have the

43 bp indel. In addition, 18 of the remaining 58 lines with the non-overexpressed Bx7 subunit also harboured this insert (Table 3.3 and Appendix III).

Table 3.3 Presence of the 18 bp indel, 43 bp indel and right and left LTR junction DNA markers at the *Glu-B1* locus of tetraploid and hexaploid accessions having a Bx7 HMW-GS variant

Ploidy (*Glu-B1x*)	**Number of accessions**	**18 bp indel**	**43 bp indel**	**Left and right LTR junction**
Tetraploid (Bx7*)	3	0	0	0
Hexaploid (Bx7*)	120	0	0	0
Tetraploid (Bx7)	24	24	1	0
Hexaploid (Bx7)	57	57	18	0
Tetraploid (Bx7^{OE})	3	3	3	3
Hexaploid (Bx7^{OE})	40	40	40	40

The structural organization of the *Glu-B1* locus of Glenlea wheat is illustrated in Figure 3.3. The tandem duplication of 10.3 kb comprised two open reading frames (ORFs) including the *Bx7* gene, and flanked a complete and a partial LTR retroelement. The complete DNA sequence of TaBAC1215C06 has been deposited in GenBank (EU157184). Several primer pair combinations for the right and left junctions of the retroelement located at the *Glu-B1* locus were designed and tested (data not shown). Two primer pairs were selected for their specificity and robustness. Figures 3.4 A and B illustrate the amplification of a 447 bp and an 884 bp fragment of the left and right junctions between the duplicated segments and the retroelement, respectively. All 43 accessions (three tetraploid and 40 hexaploid lines) identified to have the Bx7^{OE} subunit by SDS-PAGE amplified both markers. Conversely, all other accessions that did not display Bx7^{OE}, including the 203 accessions expressing another Bx7 subunit variant, did not produce the PCR amplicons (Table 3.3). The *Glu-B1* locus specificity of the markers was confirmed by use of the Glenlea *Glu-B1* BAC clone TaBAC1215C06 as template.

The combined results of the three assessment methods for all Bx7^{OE} accessions and checks are given in Table 3.4. References are listed for lines that had previously been reported to have Bx7^{OE}.

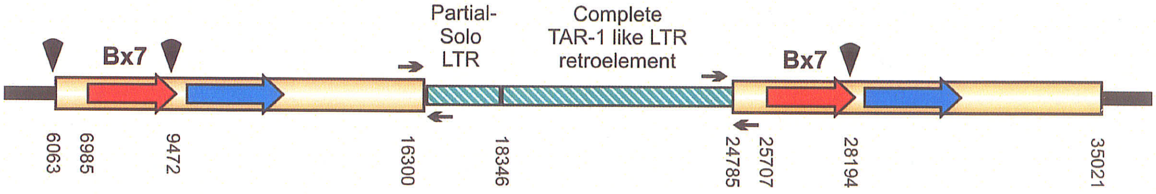


Figure 3.3 Structural organization of the *Glu-B1* locus in cv Glenlea showing a 10.3 kb duplication encompassing the *Bx7* gene flanking a complete and a partial LTR retrotransposon. Nucleotide positions are indicated below. The first ORF of the duplication encodes the HMW-GS Bx7. The downstream predicted ORF encodes a putative protein kinase. The triangles represent indels and the arrows indicate primer binding sites. Primer pair for the left junction generates a 447 bp amplicon and the primer pair for the right junction generates 844 bp amplicon.

PCR analysis results of the three *Glu-B1* markers outlined in Table 3.3 clearly establish a relative timeline for these evolutionary events. Since none of the Bx7* lines have the 43 bp indel and none of the Bx7 lines have the left and right junction markers and since the corollaries are also true, it can be inferred that the 18 bp insertion event pre-dates the 43 bp insertion event which, in turn, pre-dates the LTR retroelement mediated duplication at the *Glu-B1* locus (Fig. 3.5). Further, all these Bx7 variants were found in both tetraploid and hexaploid collections thereby supporting the multiple polyploidization event hypothesis of hexaploid wheat.

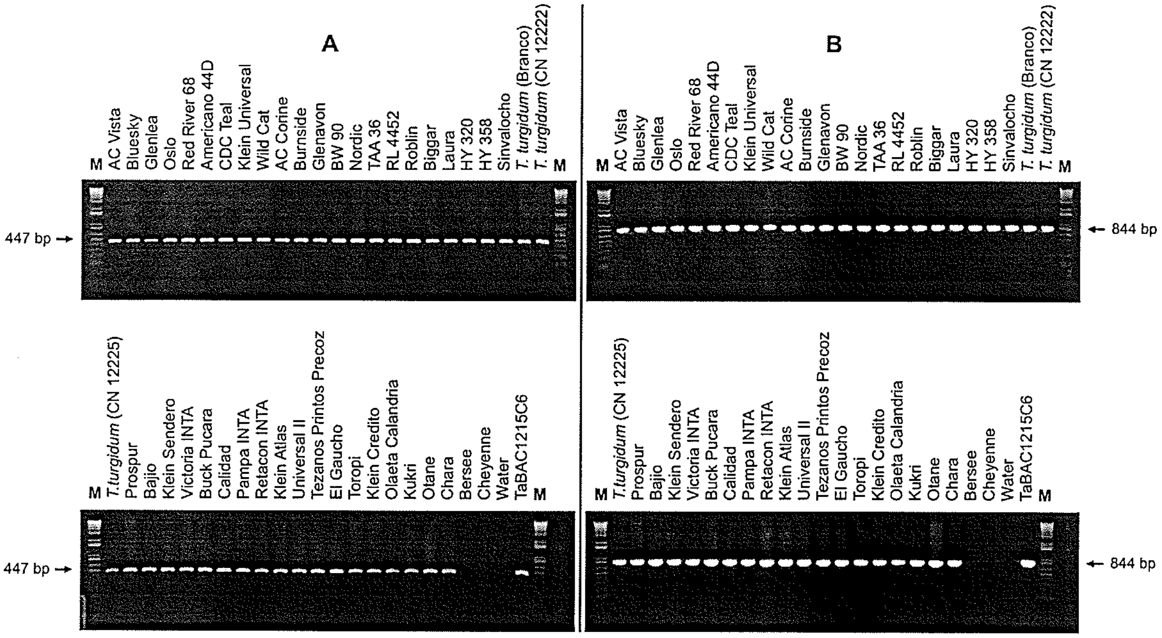


Figure 3.4 PCR amplification of the *Triticum* accessions identified to have the Bx7 ^{OE} HMW-GS by SDS-PAGE. Negative controls cv Bersee (Bx7), cv Cheyenne (Bx7*) and water (no DNA) and positive control BAC clone TaBAC1215C06 from Glenlea were included. **A.** PCR amplification of the left junction of the retroelement and the duplicated region generated a 447 bp amplicon in all Bx7 ^{OE} accessions. **B.** PCR amplification of the right junction of the retroelement and the duplicated region generated an 844 bp amplicon. Marker (*M*) is 1 kb Plus DNA ladder (Invitrogen, Mississauga, Canada).

Table 3.4 SDS-PAGE analysis, RP-HPLC quantification and PCR analyses for the 18 bp indel, the 43 bp indel and the *Glu-B1* retroelement left and right junction markers of tetraploid and hexaploid accessions with the Bx7^{OE} phenotype and the check lines

Accession Number	**Name**	**origin**	**HMW-GS Bx7^{OE} (SDS-PAGE)**	**Relative proportion of Bx7subunit to total HMW-GS (RP-HPLC)**	**18 bp indel**	**43 bp indel**	**Left and right LTR junction**	**Reference(s)**
T. aestivum accessions with Bx7^{OE} subunit								
-	AC Vista	Canada	+	42.49	+	+	+	-
-	Bluesky	Canada	+	41.47	+	+	+	Butow et al. 2004, Marchylo et al. 1992, and Vawser and Cornish 2004
CN 44438	Oslo	Canada	+	42.43	+	+	+	Marchylo et al.1992, and Vawser and Cornish, 2004
CItr 14193	Red River 68	USA	+	40.67	+	+	+	Butow et al. 2004, D'Ovidio et al. 1997, and Vawser and Cornish 2004
-	RL 4452	Canada	+	36.94	+	+	+	-
-	Roblin	Canada	+	43.25	+	+	+	Butow et al. 2004, Marchylo et al. 1992, and Vawser and Cornish 2004

Accession Number	Name	origin	HMW-GS $Bx7^{OE}$ (SDS-PAGE)	Relative proportion of Bx7subunit to total HMW-GS (RP-HPLC)	18 bp indel	43 bp indel	Left and right LTR junction	Reference(s)
CN 51820	Wild Cat	Canada	+	38.26	+	+	+	Butow et al. 2004, Marchylo et al. 1992, and Vawser and Cornish 2004
PI 191937	Americano 44D	Uruguay	+	40.32	+	+	+	Butow et al. 2004
-	CDC Teal	Canada	+	38.36	+	+	+	-
-	AC Corinne	Canada	+	41.73	+	+	+	-
-	Burnside	Canada	+	40.89	+	+	+	-
-	Glenavon	Canada	+	41.06	+	+	+	-
CN 43694	BW90	Canada	+	44.24	+	+	+	Marchylo et al. 1992
-	Nordic	USA	+	44.24	+	+	+	Marchylo et al. 1992
-	TAA 36	Israel	+	43.57	+	+	+	Lukow et al. 1989
CItr 12606	Klein Universal	Argentina	+	39.76	+	+	+	Gianibelli et al. 2002
CN 51812	Bigger	Canada	+	40.77	+	+	+	Butow et al. 2004, Marchylo et al. 1992, and Vawser and Cornish 2004
CN 44167	Laura	Canada	+	42.93	+	+	+	Butow et al. 2004,Marchylo et al.1992, and Vawser and Cornish 2004

Accession Number	**Name**	**origin**	**HMW-GS Bx7^{OE} (SDS-PAGE)**	**Relative proportion of Bx7subunit to total HMW-GS (RP-HPLC)**	**18 bp indel**	**43 bp indel**	**Left and right LTR junction**	**Reference(s)**
CN 42929	HY320	Canada	+	41.4	+	+	+	Marchylo et al. 1992
CN 44146	HY358	Canada	+	39.22	+	+	+	Marchylo et al. 1992
CN 11969	Sinvalocho	Uruguay	+	44.07	+	+	+	Butow et al. 2004, and Vawser and Cornish 2004
-	Glenlea	Canada	+	38.4	+	+	+	Butow et al. 2004, Lukow et al. 1989, and Vawser and Cornish 2004
CWI 16281	Prospur	USA	+	37.16	+	+	+	Vawser and Cornish 2004
BW 386	Bajio	Mexico	+	37.35	+	+	+	Vawser and Cornish 2004
CWI 77253	Klein Sendero	Argentina	+	47.15	+	+	+	Butow et al. 2004
BW 12005	Victoria INTA	Argentina	+	40.69	+	+	+	Butow et al. 2004
BWI 1255	Buck Pucara	Argentina	+	41.95	+	+	+	-
BW 464	Calidad	Argentina	+	44.33	+	+	+	Vawser and Cornish 2004
BW 152416	Pampa INTA	Argentina	+	43.99	+	+	+	Butow et al. 2004
CWI 33350	Retacon INTA	Argentina	+	41.06	+	+	+	Butow et al. 2004

Accession Number	Name	origin	HMW-GS Bx7^{OE} (SDS-PAGE)	Relative proportion of Bx7subunit to total HMW-GS (RP-HPLC)	18 bp indel	43 bp indel	Left and right LTR junction	Reference(s)
BW 4689	Klein Atlas	Argentina	+	47.24	+	+	+	
CWI 14048	Universal II	Argentina	+	41.36	+	+	+	Vawser and Cornish 2004
BW 779	Tezanos Printos Precoz	Argentina	+	42.79	+	+	+	Butow et al. 2004 ,and Vawser and Cornish 2004
CN 10020	El Gaucho	Argentina	+	45.06	+	+	+	-
CN 44011	Toropi	Brazil	+	40	+	+	+	-
CN 10856	Klein Credito	Uruguay	+	44.4	+	+	+	-
CN 11243	Olaeta Calandria	Uruguay	+	40.41	+	+	+	-
Aus 29472	Kukri	Australia	+	44.24	+	+	+	Butow et al. (2002, 2004), and Vawser and Cornish 2004
Aus 30426	Otane	New Zealand	+	42.94	+	+	+	Butow et al. 2004, Sutton 1991, and Vawser and Cornish 2004
Aus 30031	Chara	Australia	+	41.01	+	+	+	Butow et al. (2002 ,2004), and Vawser and Cornish 2004

Accession Number	Name	origin	HMW-GS Bx7^{OE} (SDS-PAGE)	Relative proportion of Bx7subunit to total HMW-GS (RP-HPLC)	18 bp indel	43 bp indel	Left and right LTR junction	Reference(s)
T. turgidum accessions with Bx7^{OE} subunit								
CN 2644	Branco	Portugal	+	63.29	+	+	+	-
CN 12222	CN 12222	Czech	+	63.66	+	+	+	-
CN 12225	CN 12225	Czech	+	71.6	+	+	+	-
Checks (*T. aestivum* cultivars)								
PI 447404	Yang Mai No.1	China	-	29.6	-	-	-	-
CItr 6731	Benefactor	UK	-	30.26	-	-	-	-
01C0100613	Bankuti	Hungary	-	30.65	-	-	-	-
CWI 14942	Klein Sin Rival	Argentina	-	30.4	-	-	-	-
CN 10719	Kenya Farmer	Kenya	-	29.8	-	-	-	-
01C0200129	Maja	Czech Republic	-	30.81	+	-	-	-
CItr 8885	Cheyenne	USA	-	27.58	-	-	-	-
CN 38927	Katepwa	Canada	-	29	-	-	-	-
CN 11189	Neepawa	Canada	-	29.56	-	-	-	-
PI 520297	Stoa	USA	-	28.93	-	-	-	-
-	AC Minto	Canada	-	30.09	-	-	-	-
-	Columbus	Canada	-	28.95	-	-	-	-

Accession Number	Name	origin	HMW-GS $Bx7^{OE}$ (SDS-PAGE)	Relative proportion of Bx7subunit to total HMW-GS (RP-HPLC)	18 bp indel	43 bp indel	Left and right LTR junction	Reference(s)
Check (*T. turgidum* accession)								
CN 51263	CN 51263		-	57.04	-	-	-	-

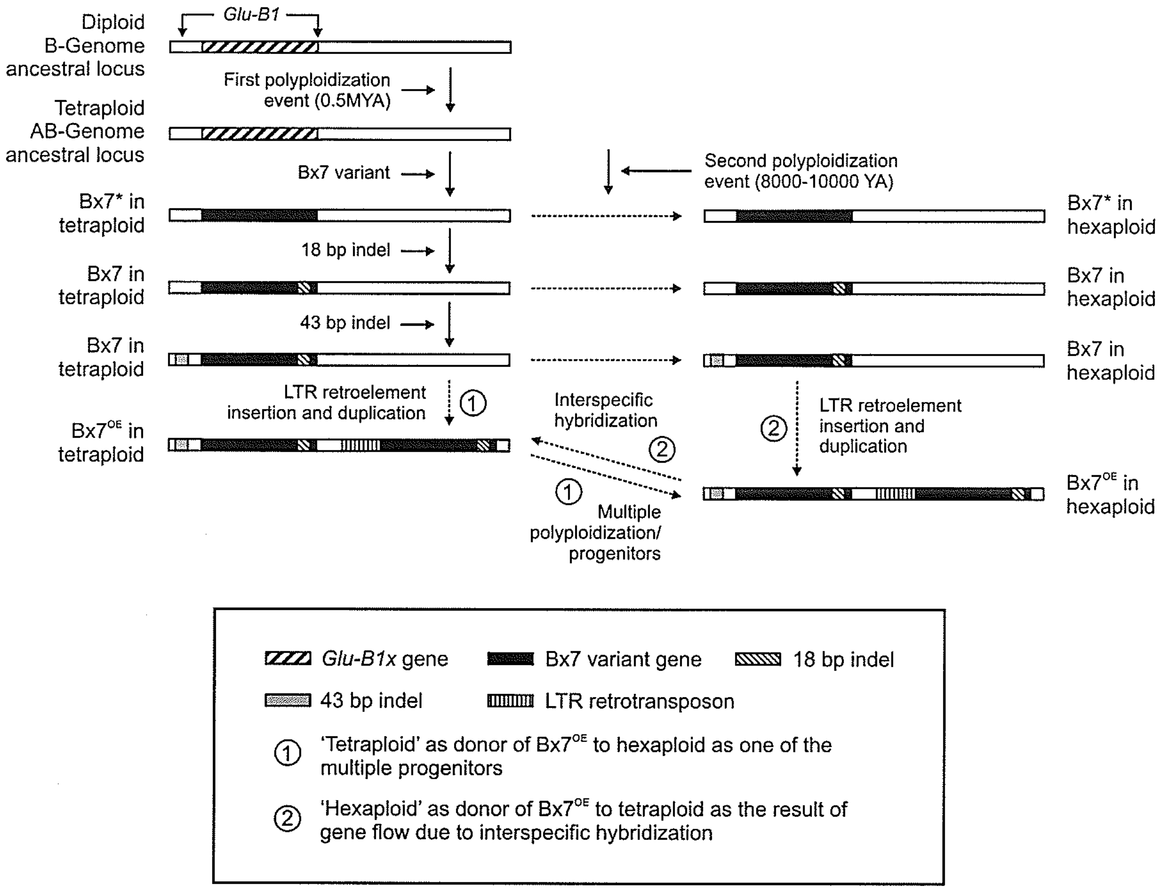

Diploid B-Genome ancestral locus
Glu-B1
First polyploidization event (0.5MYA)
Tetraploid AB-Genome ancestral locus
Bx7 variant
Second polyploidization event (8000-10000 YA)
Bx7* in tetraploid
Bx7* in hexaploid
18 bp indel
Bx7 in tetraploid
Bx7 in hexaploid
43 bp indel
Bx7 in tetraploid
Bx7 in hexaploid
LTR retroelement insertion and duplication
1
Interspecific hybridization
2
LTR retroelement insertion and duplication
2
Bx7^{OE} in tetraploid
1
Multiple polyploidization/ progenitors
Bx7^{OE} in hexaploid
Glu-B1x gene
Bx7 variant gene
18 bp indel
43 bp indel
LTR retrotransposon
1 'Tetraploid' as donor of Bx7^{OE} to hexaploid as one of the multiple progenitors
2 'Hexaploid' as donor of Bx7^{OE} to tetraploid as the result of gene flow due to interspecific hybridization

Figure 3.5 Diagram of the chronological events that occurred at the *Glu-B1* locus of wheat lines with a *Bx7* allele variant illustrating the possibilities of several independent polyploidization events involving different tetraploid ancestors. Elements not drawn to scale to better illustrate the insertion and deletion events.

3.5 Discussion

Gene duplication generates the raw materials for evolutionary novelties such as new functions and expression patterns (Lynch and Conery 2000). Genome-wide dispersed duplicated regions originate from ancient polyploidization followed by chromosome fusions and translocations as observed in *Arabidopsis* (Vision et al. 2000; Wolfe 2001). In contrast, tandem segmental duplications result from unequal crossing over mediated by repetitive DNA such as retroelements (Zhang 2003; Dubcovsky and Dvorak 2007).

Duplication of the ancestral *Glu-1* sequence leading to paralogous gene copies encoding x- and y- type subunits was dated at 7.2-10.0 million years ago (MYA) before the divergence of the wheat genomes 5.0-6.9 MYA (Allaby et al. 1999). In the absence of any direct evidence for the molecular mechanism involved in the duplication of the ancestral *Glu-1* gene, the predominance of repetitive elements (>80% of the genome, Smith and Flavell 1975) can indicate their possible role in duplication of this locus as previously suggested (Dubcovsky and Dvorak 2007). Presence of retroelements in the intergenic regions of genes encoding x- and y- type subunits of HMW glutenin (Anderson et al. 2003; Gu et al. 2006) suggest that the mechanism of inter-element ectopic recombination could have lead to these paralogous genes. Similarly, presence of the LTR retroelements in the interval of the 10.3 Kb duplicated segments encompassing the Bx7 gene at the *Glu-B1* locus indicate their possible involvement in the origin of the duplication in the cultivar Glenlea (Fig 3.3, Cloutier et al. 2005).

The duplication of the gene encoding the HMW-GS Bx7 has been proposed as the cause for the overexpression of this subunit in cultivar Glenlea (Cloutier et al. 2005). While most cultivars and accessions of hexaploid wheat express 3-5 subunits, lines with the Bx7^{OE} phenotype can also express up to five different subunits but likely have six functional genes encoding HMW-GS. A 43 bp indel located 572 bp upstream from the transcription initiation site had previously been described and used to develop a marker associated with the Bx7^{OE} subunit (Radovanovic and Cloutier 2003; Butow et al. 2004). This marker was used in two independent studies to characterize a Canadian and an Australian segregating population and to establish the significant genetic contribution of the *Bx7^{OE}* allele to dough strength characteristics (Butow et al. 2003; Radovanovic et al.

2002). While this 43 bp indel was always present in Bx7^{OE} lines, the reverse was not always true i.e., this indel was found in some non-Bx7^{OE} lines (Butow et al. 2004). Our extensive study corroborated these findings. Butow et al. (2004) also used RP-HPLC and the presence of an 18 bp indel (one hexapeptide motif) in the coding region of the allele to characterize lines from eight different *Glu-B1* allelic groups but found that neither the 43 bp nor the 18 bp indel were perfectly linked with the overexpression phenotype. These short tandem insertions were also found in the accessions of *T. turgidum* indicating that these indels occurred prior to the polyploidization event that led to the formation of hexaploid wheat 8,000 - 10,000 years ago. Indeed, the duplication of the gene encoding the Bx7 subunit hypothesized by Lukow et al. (1992) and D'Ovidio et al. (1997) based on Southern hybridization and recently demonstrated by BAC sequencing (Cloutier et al. 2005) was confirmed in the present study. Without exception, all 43 accessions overexpressing HMW-GS Bx7 shared the same locus structure with the LTR retroelement flanked by the duplication encompassing the *Bx7* gene, and, all 369 accessions that were non-Bx7 ^{OE} lacked this genomic structure. The 43 bp insert was found in four *T. turgidum* accessions including the three Bx7^{OE} lines indicating that this short tandem insertion occurred prior to the retrotransposon mediated duplication. Similarly, the 18 bp indel in the coding region also pre-dates the duplication corroborating the finding of Butow et al. (2004) who reported these two indels in accessions of *T. turgidum.*

DNA markers are being employed as tools to improve the efficiency of selection in the pre-breeding and breeding activities in wheat (Eagles et al. 2001; Gale 2005). Marker assisted selection for Bx7^{OE} subunit will be useful since this subunit contributes to dough strength (Butow et al. 2003; Lukow et al. 1992; Marchylo et al. 1992; Radovanovic et al. 2002). DNA markers were developed based on both the coding region of the gene (Butow et al. 2003, 2004; Ma et al. 2003) and the promoter region of the gene encoding HMW-GS Bx7 (Butow et al. 2004; Juhasz et al. 2003; Radovanovic and Cloutier 2003). However, they were employed in the selection strategies along with SDS-PAGE and/or RP-HPLC profiling because of their imperfect association with the Bx7$^{\mathrm{OE}}$ phenotype. The dominant left and right junction markers developed in this study are perfectly linked to the Bx7$^{\mathrm{OE}}$ phenotype and can be applied in the selection schemes of wheat breeding programs.

Pedigree analysis indicated that an Argentinean landrace, namely Klein Universal II released in 1922, was the source of worldwide dissemination of the *Glu-B1al* allele (Bx7$^{\mathrm{OE}}$ + By8) through the CIMMYT germplasm used in modern wheat breeding programs (Butow et al. 2004). Further, it was suggested that Americano 44D, a Uruguayan landrace of unknown origin could be the donor of the allele found in Klein Universal II. However, the origin of the *Glu-B1al* allele found in the Israeli landrace (TAA36) and the Hungarian landrace Bankuti 1201 could not be explained by the Argentinean ancestor. Historical information indicated that Eastern European landraces introduced by immigrants in the nineteenth century could be the possible source of the *Glu-B1al* allele found in Americano 44D, TAA36 and Bankuti 1201(Butow et al. 2004). The presence of the gene duplication in the tetraploid accessions of European origin as reported in the present study reinforces the possible European origin of the *Bx7*^{OE} allele found in these landraces.

The survey also found previously unreported lines with the HMW-GS Bx7^{OE} in the hexaploid cultivars namely, Klein Credito and Olaeta Calandria from Uruguay and Toropi from Brazil. The high frequency of the Bx7^{OE} subunit reported in the Argentinean wheat cultivars was also confirmed (Gianibelli et al. 2002; Vawser and Cornish 2004). However, the Hungarian landrace Bankuti 1201 previously reported to have HMW-GS Bx7^{OE} (Juhasz et al. 2003), did not exhibit the genome organization corresponding to the duplication nor did it show this subunit on SDS-PAGE. The presence of different biotypes in this accession was reported earlier and is presumed to be the reason for the discrepancies between the various reports to date (Butow et al. 2004; Juhasz et al. 2003).

Evolutionary mechanisms in the form of retroelement mediated segmental chromosomal duplication can have significant phenotypic impacts on agriculturally important loci such as the *Glu-B1* locus. Gene duplication by helitron-like transposons in maize and gene fragment acquisition by pack-MULES in rice were reported (Jiang et al. 2004; Morgante et al. 2005). Unlike the *cut and paste* mode of transposition of these mobile elements, the *copy and paste* mode of propagation of retroelements distribute their homologous sequences across the genome, which in turn, increases the probability of ectopic recombination leading to deletions and duplications. Deletions as an evolutionary force are implicated in drastic reductions in genome sizes (Bennetzen 2002; Ma et al. 2004; Vitte and Panaud 2005). However, the phenotypic impacts of duplications have not been frequently discovered in plant genomes of agricultural importance. The observed segmental duplication resulting in two functional copies of the gene encoding HMW-GS Bx7 signifies that the evolutionary dynamics of the genome driven by retroelements may have played a role in shaping the structural organization of not only the biologically important loci discovered earlier but also agriculturally important loci (Gaut et al. 2007).

Comparative sequence analyses of the coding regions of the *Glu-1* alleles and their immediate upstream regions were carried out extensively (Halford et al. 1987; Anderson and Greene 1989; Mackie et al. 1996; Allaby et al. 1999; Shewry et al. 2003). However, the genome organization of the orthologous *Glu-1* loci with its distal flanking regions covering a few 100 Kb emerged only after the construction of large insert BAC libraries. The observed intergenic distances between genes encoding x- and y-type subunits of HMW glutenin are 140 Kb, 168 Kb and 51 Kb respectively in the A, B and D genomes (Gu et al. 2004; Kong et al. 2004; Anderson et al. 2003). Though there is conservation of gene order and orientation at the orthologous *Glu-1* loci, micro colinearity at the intergenic region is disrupted mainly due to the insertion of retrotransposons as observed in the maize genome (SanMiguel et al. 1996). The segmental duplication encompassing the gene coding for the Bx7 subunit described herein, originated as a consequence of the mechanism of unequal homologous crossing over driven by the LTR retroelement inserted into the locus (Cloutier et al. 2005). The models of the retroelement mediated origin of this tandem segmental duplication at the *Glu-B1* locus are presented in Appendix V.A and V.B. It is possible to estimate the time of insertion of LTR retroelements by comparing observed nucleotide substitutions between right and left LTRs because point mutations accumulate over time (Gaut et al. 1996; San Miguel et al. 1998). Cloutier et al. (2005) identified no base substitutions between the LTRs of the retroelement at the *Glu-B1* locus. However, nucleotide substitution rate of the duplicated region yielded an estimated time of the duplication of

15,000 years ago ± 11,000 years, which overlaps with the polyploidization event of hexaploid wheat estimated to be 8,000-10,000 years ago (Huang et al. 2002). Stress can activate transposons and lead to their retrotransposition into new sites (Grandbastien 1992). Polyploidization per se could have caused the transposition of the LTR retroelement at the *Glu-B1* locus followed by inter-element recombination leading to the segmental duplication. In this case, the *Glu-B1al* allele would not be found in diploid or tetraploid progenitors but would be restricted to hexaploid wheat accessions. Our results clearly showed that this was not the case because three tetraploid accessions displayed the *Bx7^OE^* allele and were structurally identical at the genomic level to the Bx7 $^{\mathrm{OE}}$ hexaploid lines i.e., they had the duplicated region flanking the LTR retroelement, the 18 bp and the 43 bp indels. Moreover, they showed high staining intensity on SDS-PAGE and had a higher proportion of Bx7 HMW-GS when compared to the check. Butow et al. (2004) had hypothesized the presence of the Bx7 $^{\mathrm{OE}}$ phenotype in *T. turgidum* var. Portugal 170 but could not confirm the overexpression phenotype.

Many polyploid species were hypothesized to have formed recurrently from several crosses involving different gene pools of their progenitor species (Soltis and Soltis 1999). Polyploid wheat exists in both tetraploid and hexaploid forms which could have originated independently from hybridizations between distinct diploid or tetraploid ancestors (Feldman 2001). Comparative analysis of orthologous regions for shared genome organization between the D-genome of hexaploid wheat and *Ae. tauschii,* its diploid D genome donor, indicated the existence of more than one shared allele (Caldwell et al. 2004; Dvorak et al. 1998; Giles and Brown 2006; Talbert et al. 1998). These shared alleles suggested that the hexaploid wheats were formed by recurrent hybridizations involving more than one genotype of *Ae. tauschii.* Gu et al. (2006) compared nucleotide substitution rates for the A and B genomes of tetraploid and hexaploid wheats at the *Glu-1* loci and found that they differed significantly despite co-evolving in the same nuclei in their respective species thereby supporting the hypothesis of more than one tetraploid ancestor with distinct A genome lineage in the origin of hexaploid wheat. The present study provides additional evidence for the multiple tetraploid ancestor hypothesis for hexaploid wheat, however based on evidence from the B-genome. The existence of two different shared genome organizations at the *Glu-B1* locus of *T. turgidum* and *T. aestivum* indicates that at least one *T. turgidum* line with the *Bx7* duplication and LTR retrotransposon and one without could have served as progenitors in the formation of hexaploid wheat. Findings supporting the same hypothesis were reported at the orthologous *Lr10* loci where two conserved deletion point haplotypes were described in the A genome at the three ploidy level i.e. *T. monococcum*, *T turgidum* and *T. aestivum* (Isidore et al. 2005).The three *T. turgidum* lines described to have a Bx7^{OE} phenotype and the retroelement could have acquired it through inter-specific hybridization with hexaploid lines in either natural habitats or in classical plant breeding efforts aimed at introgression of desirable traits. These scenarios are however unlikely because two of the lines are landraces and the third one, cultivar Branco, was described as a released landrace not improved by breeders. Aside from the $Bx7^{OE}$ versus *Bx7* allele presence in both tetraploid and hexaploid wheat collections, the multiple ploidization hypothesis involving different tetraploid ancestors is further supported by the findings of *Bx7** and *Bx7* alleles with and without the 43 bp indel in both collections as well.

3.6 Conclusion

In this study, the genomic organization of the *Glu-B1* locus and the expression level of the Bx7 subunit were simultaneously assessed in a number of diploid, tetraploid and hexaploid accessions. The perfect correlation between the presence of the gene duplication and the overexpression of the Bx7 HMW-GS reinforce the causal link between genotype and phenotype. The duplication described herein occurred as a consequence of the transposition of an LTR retroelement.

The structural organization associated with the segmental duplication encompassing the *Glu-B1* locus in three tetraploid accessions indicated that the retroelement mediated recombination event occurred prior to the polyploidization event resulting in hexaploid wheat speciation. Our data also supports the proposal of multiple polyploidization events in the origin of the hexaploid wheat genome, primarily based on the presence of two independent genome organizations at the *Glu-B1* locus between tetraploid and hexaploid accessions. However, the possibility of gene flow as the result of interspecific hybridization between *T. aestivum* and *T. turgidum* in the natural habitats was not excluded. The result also serves as an evidence for the role of retroelements on the evolution of agriculturally important loci. Finally, the DNA markers identified in the present study can be used as perfectly linked markers for the *Glu-B1al* allele encoding the $Bx7^{OE}$ subunit in wheat breeding programs.

Molecular phylogeny of the Sasanda LTR *copia* retrotransposon family reveals recent amplification activity in *Triticum aestivum* (L.)

Raja Ragupathy [1, 2] and Sylvie Cloutier [2]

[1]Department of Plant Science, Faculty of Graduate Studies,
University of Manitoba, Winnipeg, Canada R3T 2N2

[2]Cereal Research Centre, Agriculture and Agri-Food Canada,
195 Dafoe Road, Winnipeg, Canada R3T 2M9

The thesis author, Raja Ragupathy, designed, carried out the experiments, did the data analysis, interpretation and drafted the manuscript. As major advisor, Dr. Sylvie Cloutier guided the direction of the study, participated in data analysis and manuscript review.

CHAPTER 4

Molecular Phylogeny of the Sasanda LTR *Copia* Retrotransposon Family Reveals Recent Amplification Activity in *Triticum aestivum* (L.)

4.1 Abstract

Retrotransposons constitute a major proportion of the *Triticeae* genomes. Genome-scale studies have revealed their role in evolution affecting both genome structure and function. In this study, family members of an LTR *copia* retrotransposon which mediated the duplication of the gene encoding the high molecular weight glutenin subunit Bx7 in cultivar Glenlea were characterized. This novel element was named Sasanda_EU157184-1. High density filters of the Glenlea hexaploid wheat BAC library were screened with a Sasanda long terminal repeat (LTR) specific probe and ~1075 positive clones representing an estimated copy number of 347 elements per haploid genome were identified. The 242 BAC clones with the strongest hybridization signal were selected. To maximize isolation of complete elements, this subset of clones was screened with a reverse transcriptase (RT) domain probe and DNA was isolated from the 133 clones that produced a strong hybridization signal. Fingerprinting confirmed that 12 clones represented the same locus as other clones in the subset and these were then removed. Left (5') and right (3') LTRs as well as the RT domains were PCR amplified and sequencing was carried out on the final subset of 121 clones. Phylogenetic inference was obtained from a data set consisting of 100 RT, 89 left LTR and 89 right LTR sequences representing 233, 460 and 501 active sites for comparison, respectively. Neighbour-joining tree constructed using the Kimura II parameter method with a mutation rate of

$2x10^{-8}$ substitutions per synonymous site per year indicated that the element is at least 1.2 to 1.8 million years old and has evolved into a minimum of five sub-families. The insertion times of the 89 complete elements were estimated based on the divergence between their LTRs. Corroborating the inference from the RT domain, analysis of the LTR domains also indicated bursts of amplification from 1.7 million years ago (MYA) to 0 MYA except for one member which dated 2.9 ± 0.4 MYA coinciding with the divergence of *Triticum* and *Aegilops,* 3 MYA. In 49 elements, the left and right LTRs were identical indicating recent transposition activity. The element can be used to develop retrotransposon based markers such as sequence specific amplification polymorphism (SSAP), retrotransposon-microsatellite amplification polymorphism (REMAP) and inter retrotransposon amplification polymorphism (IRAP), all of which are well suited for genotyping studies.

4.2 Introduction

Plant genomes vary greatly in their C-values, *i.e.,* the DNA amount of the unreplicated haploid genome (Bennet and Leitch 1995). The genome size of rice (430 Mb), sorghum (748 Mb), maize (2292 Mb-2716 Mb), barley (4873 Mb), *Triticum monococcum* (5751 Mb), *Aegilops squarrosa* (4024 Mb) and bread wheat (15966 Mb) represent nearly a 40 fold variation even among the plants belonging to the same family of *Poaceae* (Arumuganathan and Earle 1991). Polyploidization is the single most important phenomenon contributing to genome size variations. At the same ploidy level, however, repetitive fractions of DNA namely, transposable elements, tandem repeats, microsatellites and rDNA genes account for genome size differences (Kubis et al. 1998).

Nearly 3600 annotated sequences representing different repeat types have been characterized so far from diverse eukaryotic genomes (Jurka et al. 2005). Among them, the contribution of transposable elements to the evolution of the genome has been studied at the genome-scale level because of their abundance and diversity (Havecker et al. 2004; Wessler 2006a and 2006b; Morgante 2006). They constitute ~70% of the repeat fraction, which occupies 90% of *Triticeae* genomes (Li et al. 2004). Copy number increase of a few families of elements alone contributed to the increase in the genome size of some species such as barley and rice (Kalendar et al. 2000; Piegu et al. 2006). Additionally, they play a central role in the evolution of structure and organization of the genome (Bennetzen 2000b; Lönnig and Saedler 2002; Vitte and Bennetzen 2006). They also contribute to gene evolution by donating regulatory sequences, driving gene duplications and exon shuffling, altering structure as well as expression and causing gene disruptions (Jordan et al. 2003; Morgante et al. 2005; White et al.1994; Juretic et al. 2005; Giovanni et al. 2008). As well, instances of recruitment of transposase by the host organisms for their normal cellular functions were discovered (Muehlbauer et al. 2006). They also activate and/or silence adjacent genes by epigenetic mechanisms such as methylation (Kashkush et al. 2003).

Transposable elements are classified into class I (retrotransposons) and class II (DNA transposons) based on their mechanism of transposition. The former transposes via an RNA intermediary and the latter transpose directly as DNA. DNA transposons were first described in maize by McClintock as controlling elements causing variations in kernel pigmentation by jumping across the genome by cut-and-paste mechanism (McClintock 1987). The copy number of DNA transposons will increase only when the transposition occurs during DNA replication whereas the copy number of retrotransposons can increase anytime during the life cycle of an organism because of the copy-and-paste mechanism of propagation (Wicker et al. 2007b). Transposons are further classified into subclasses, orders, superfamilies and families (Wicker et al. 2007b). Superfamilies are defined as a group of elements with limited homology at the protein level, high heterogeneity at the nucleotide level, having specific target site duplications (presence and size) and a distinct order of polyprotein domain arrangement. A total of 29 superfamilies have been identified so far (Wicker et al. 2007b). *Ty1-copia* and *Ty3-gypsy* are two prominent superfamilies studied in eukaryotes so far. The order of arrangement of the integrase (IN) and the reverse transcriptase (RT) domains distinguishes these two superfamilies. In *copia* elements, the IN domain is located upstream of the RT domain while the order is reversed in *gypsy* elements. Groups of elements with conserved nucleotides and mobilized by the same set of proteins belong to a family and usually have at least 80% sequence identity. Below this threshold, elements are classified in separate families. In addition, LTR retrotransposon families are also defined by unique non-cross hybridizing long terminal repeats (LTRs) because they are rapidly evolving components of the genome (SanMiguel et al. 1998; Wicker and Keller 2007). Subfamilies are identified by phylogenetic analysis of the members of a family represented by the individual insertions in the genome. RT, IN and LTR domains have been used to gain understanding of the dynamics of retroelement families in wheat, rice, maize, barley, pigeon pea, mungbean and cotton (Matsuoka and Tsunewaki 1996; Matsuoka and Tsunewaki 1999; Ma et al. 2004; Marillonnet and Wessler 1998; Vicient, et al. 2005; Lall et al. 2002; Dixit et al. 2006; Hawkins et al. 2008).

Most retroelements are fragmented with high copy numbers of solo LTRs due to the removal of the internal domain along with an LTR by inter-element or intra-element unequal homologous recombination as well as illegitimate recombination (Devos et al. 2002; Vitte and Panaud 2003). Also, nesting with multiple tiers of different elements is a common feature of large genomes like maize, wheat and barley (SanMiguel et al. 1996; Wicker et al. 2003; Wicker et al. 2005). Complete retroelements are defined as insertions that have intact ends but this does not imply that the elements are actually functional or that they even contain internal domains (Wicker and Keller 2007). Indeed, complete elements with non-degenerated internal domains are rare. Time of insertion of retroelements is estimated by comparing the sequence identity of the LTRs based on the fact that they were identical at the time of insertion and that accumulated mutations can be correlated to the amount of time since insertion (SanMiguel et al. 1998). Rates of nucleotide substitutions have been estimated from comparative studies of orthologous genes and retroelements and can be used to estimate retroelement insertion time.

There are thousands of transposable element families in plants (Wicker et al. 2007b). Nearly 32 families of *copia* elements (Wicker and Keller 2007) were represented in the Triticeae REPetitive element database (TREP, http://wheat.pw.usda.gov/ITMI/Repeats). Though a few families of elements have been well characterized in rice, barley, wheat, legumes and *Brassica* species (McCarthy et al. 2002; Schulman and Kalendar 2005; Wicker and Keller 2007; Holligan et al. 2006; Alix et al. 2005), studies focussing on the family of an element are only recently emerging. Such studies may be useful in understanding the retroelement family evolution *per se*, their role in genome dynamics, and unravel their applications in genotyping and molecular mapping studies (Schulman 2007).

In wheat, a *copia* retroelement named Sasanda_EU157184-1 (6437 bp) was discovered at the *Glu-B1* locus of cultivar Glenlea where it was implicated in the segmental duplication of the gene encoding high molecular weight glutenin subunit Bx7 (Cloutier et al. 2005; Ragupathy et al. 2008). The locus contained a solo Sasanda element adjacent to a complete element which is in turn flanked by a 10.3 kb duplication comprising the HMW-GS *Bx7* gene. The three Sasanda LTRs of the *Glu-B1* locus were identical, indicative of their recent transposition. Furthermore the complete element displayed a non-degenerated coding region. Point mutations in the duplicated sequences permitted us to identify the time of insertion to 15,000 ± 11,000 years ago. Haplotype study of a large collection of diploid, tetraploid and hexaploid accessions identified the event in tetraploid wheat, prior to the advent of hexaploid wheat approximately 8,000 years ago (Ragupathy et al. 2008). In this study, we characterized the evolutionary dynamics of the members of the Sasanda retroelement family in the genome of *T. aestivum* cv Glenlea, and more specifically copy number estimation, subfamily structure and time of insertion.

4.3 Materials and Methods

4.3.1 Hybridization based screening of high density filters of the BAC library

A 759 bp LTR sequence specific to Sasanda_EU157184-1 was amplified by PCR using BAC clone TaBAC1215C6 as template and primer Sasanda_EU157184_49F and Sasanda_EU157184_808R (Table 4.1). Approximately 100 ng of template DNA was used in 25μl reaction volume containing 1X PCR buffer, 0.2 mM each dNTPs, 1.5 mM $MgCl_2$, 10 pmole each primer, 1U *Taq* polymerase and 80 ng BSA. Cycling conditions were 94°C for 5 minutes followed by 35 cycles of 94°C for 15 seconds, 65°C for 30 seconds, and 72 °C for 45 seconds. Final extension at 72°C for 10 minutes was followed by incubation at 4° C. Following electrophoresis in 1% agarose gel, the amplicon was extracted using the QiaexII kit following manufacturer's instructions (Qiagen, Mississauga, Canada). This LTR probe was labelled with [32P]dCTP using Ready-To-Go DNA labelling beads following the manufacturer's instructions (Amersham Biosciences, Quebec, Canada). The 24 high density filters of the Glenlea BAC library (656,640 clones) were hybridized following the procedure described in Nilmalgoda et al. (2003). BAC addresses were deduced from unique double spotted hybridization signals (each clone was printed in duplicate on the filters) and 242 positive BAC clones were rearrayed using a QBOT robotic workstation. Southern blots of rearrayed clones were made following the same procedure as the high density filters (Nilmalgoda et al. 2003).

4.3.2 Estimation of copy number

The number of double spotted hybridization signals was counted for all 24 high density filters and copy number was estimated based on the genome coverage of the library taking into account the number of positive hits, number of BAC clones screened, average insert size of the clones, the number of empty clones and the number of clones with chloroplast DNA (Jiang et al. 2002).

Table 4.1 List of primers used for PCR amplification of LTR and RT probes and sequencing of LTR and RT domains

Name	Sequence (5'→3')
PCR and sequencing of RT domain	
Sasanda_3718F	GACGGCTTTTCTAAATGGAGAG
Sasanda_3982R	CAGTATGTCATCAACATACAAGCAC
PCR amplification of left LTR	
Sasanda_49F/5510F	AATCTCTAAGGGCCCATGTG
Sasanda_1096R	CGTGAGCCATAAGGTGGTTT
PCR amplification of right LTR	
Sasanda_5446F	GAGATCTGGTGGGGGATTG
Sasanda_808R/6269R	CACGAGAGGGATAAGCGATG
Sequencing LTRs	
Sasanda_1096R	CGTGAGCCATAAGGTGGTTT
Sasanda_49F/5510F	AATCTCTAAGGGCCCATGTG
Sasanda_77F/5538F	GGCACTAGGTGTGTGGGAA
Sasanda_445F/5906F	ACGAAACAGAACGCATCTCC
Sasanda_546R/6007R	TCTCCCGTCAACCGTGTA
Sasanda_667R/6128R	GATGATGAAGGTGATGCGG
Sasanda_808R/6269R	CACGAGAGGGATAAGCGATG
Sasanda_806R/6267R	CGAGAGGGATAAGCGATGTA
Sasanda_5446F	GAGATCTGGTGGGGGATTG

4.3.3 Isolation of complete elements

To identify elements with a coding domain, the rearrayed clone subset was hybridized with a reverse transcriptase (RT) probe (RT primers, Table 4.1). BAC clones hybridizing to the RT probe were selected for fingerprinting and sequencing.

4.3.4 BAC plasmid extraction, fingerprinting and contig assembly

BAC plasmid DNA was extracted from the subset of clones that produced a strong hybridization signal with both the Sasanda LTR and RT probes using the Eppendorf Perfectprep BAC96 kit following manufacturer's instructions (Eppendorf, Hamburg, Germany) adapted for a liquid handling robot (QIAGEN 3000, Mississauga, Canada). SNaPshot fingerprinting of BAC clones was performed following Luo et al. (2003). Contig assemblies were built using the software FPC (fingerprinted contigs; Soderlund et al. 1997) at high stringency level with tolerance of 4 and a 1×10^{-18} Sulston score.

4.3.5 Amplification of RT and LTR domains and purification of the PCR product

Using a 1:100 dilution of the BAC clone subset plasmid extract as template, PCR amplification of the RT as well as the left and right LTR domains was performed with primers listed in Table 4.1. The primer pair Sasanda_49F and Sasanda_1096R was used for the left LTR domain of Sasanda elements. Similarly, target specific primers Sasanda_5446F and Sasanda_6269R were used to amplify the right LTRs. The left and right LTR regions were differentially targeted by designing one of the PCR primers in the conserved internal domain. PCR conditions were as described earlier for the probe preparation. PCR products were purified using MultiScreen384-PCR filter plates as previously described (Huang and Cloutier 2007) with a few modifications. Briefly, ~50 μl PCR product from two 25 μl reactions were combined and vacuum was applied until the wells were completely empty. The products were washed with 65 μl of water. After drying the wells, the samples were resuspended in 46 μl of water and transferred to storage plates kept at 4°C. Quality assessment and quantification of the purified products were performed by resolution of 2 μl aliquots on agarose gels.

4.3.6 Sequencing of PCR products

Approximately 50 ng of purified PCR product were lyophilized using the Savant speedvac system (Global Medical Instrumentation, Inc., Minnesota, USA). Sequencing reactions were performed in 384 well plates (Applied Biosystems, CA, USA). A total of 6 μl of sequencing reaction mix containing 1 μl 5X sequencing buffer (Applied Biosystems, CA, USA), 1 μl primer (5.2 pmoles/μl) , and 0.4 μl BigDye reaction mix (Applied Biosystems, CA, USA) completed with water was added per well . The cycling conditions implemented in a PTC-200 thermocycler (MJ Research, MA, USA) were 92°C for 5 minutes followed by 60 cycles of 92°C for 10 seconds, 55°C for 5 seconds, 60°C for 4 minutes and a final extension step at 60°C for 4 minutes, followed by incubation at 4°C. Primers used for sequencing the RT and the left and right LTRs are listed (Table 4.1). The sequencing reactions were ethanol precipitated (Huang and Cloutier 2007), air-dried and denatured in 10 μl of Hi-Di formamide per well (Applied Biosystems, CA, USA) prior to resolution on an ABI3100 Genetic Analyser (Applied Biosystems, CA, USA).

4.3.7 Sequence analysis

The PHRED (Ewing et al. 1998) and CAP3 software (Huang and Madan, 1999) implemented in an in-house developed data pipeline called SOOMOS v0.6 (Banks, personal communication) were used for base calling and assembly of independent forward and reverse reads. A minimum PHRED quality score of 20 was used for base calling and 80% identity over a minimum of 40 bp overlap between reads was used for assembly with parameters of 6, 2 and -5 for gap penalty, match value and mismatch penalty, respectively. Also, the maximum expected size of contigs was used as a constraint during the assembly process. Finally, the contigs were manually inspected for potential mis-assemblies and for the inclusion of mate-pair reads specific to each clone. Using assembled RT and left and right LTR sequences, alignments were made using ClustalW (Higgins et al. 1994) and phylogenetic trees were generated by neighbour-joining algorithm (Saitou and Nei 1987) and Kimura II parameter model of base substitution (Kimura 1980) as implemented in MEGA4 (Tamura et al. 2007). The trees were validated using Bootstrap tests performed with 1000 replicates.

4.3.8 Calculation of insertion time

Insertion times were estimated following the method of SanMiguel et al. (1998). Briefly, estimates of evolutionary divergence between each aligned LTR pair in units of number of base substitutions/site/year and its standard error were calculated using the Kimura II parameter model implemented in MEGA4 (Tamura et al. 2007). A substitution rate of $2x10^{-8}$ sites per year, based on studies of rice retrotransposons was used (Vitte et al. 2004).

4.4 Results

The retroelement sequence Sasanda_EU157184-1 was obtained from sequencing of the Glenlea hexaploid wheat BAC clone TaBAC1215C06. It is 6443 bp long and is comprised of two identical 976 bp LTRs (Fig. 4.1). The non-degenerated internal region of 4485 bp includes a 3819 bp domain encoding a predicted 1272 amino acid polyprotein.

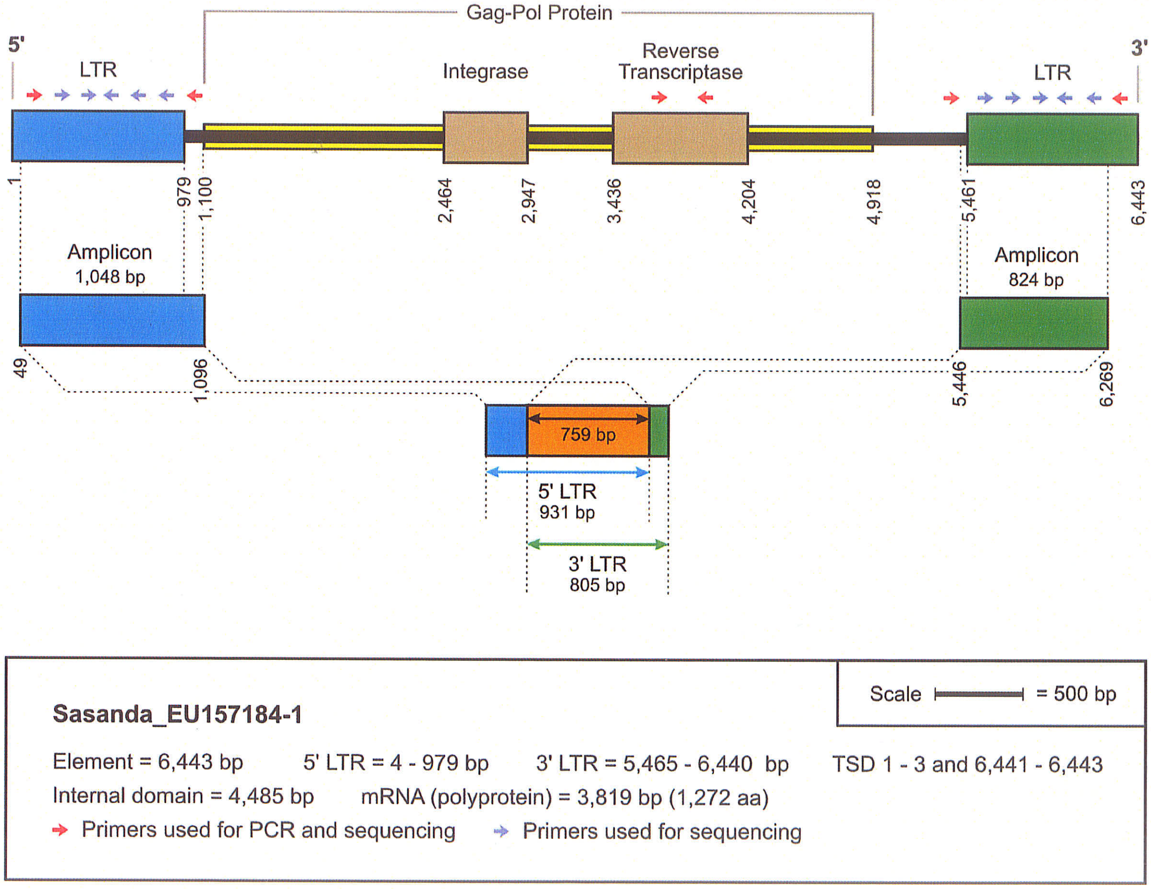


Figure 4.1 Structural features of the retroelement Sasanda_EU157184-1. Left and right LTR amplicons and their overlapping sequences are identified.

4.4.1 Estimation of copy number and isolation of complete elements

Screening of the 24 high density filters of the Glenlea BAC library with the LTR probe yielded ~1075 positive double hybridization signals (Fig. 4.2). Based on the estimated haploid genome coverage of this library of 3.1X (Nilmalgoda et al. 2003), an estimated copy number of 347 Sasanda elements would be present per bread wheat haploid genome. These could be either complete or solo elements. From the original 1075 BAC clone set, a total of 242 clones showing strong hybridization signals with the LTR probe were rearrayed. A reduced subset of 133 BAC clones hybridized to the RT probe and was selected for further fingerprinting and sequencing (Table 4.2).

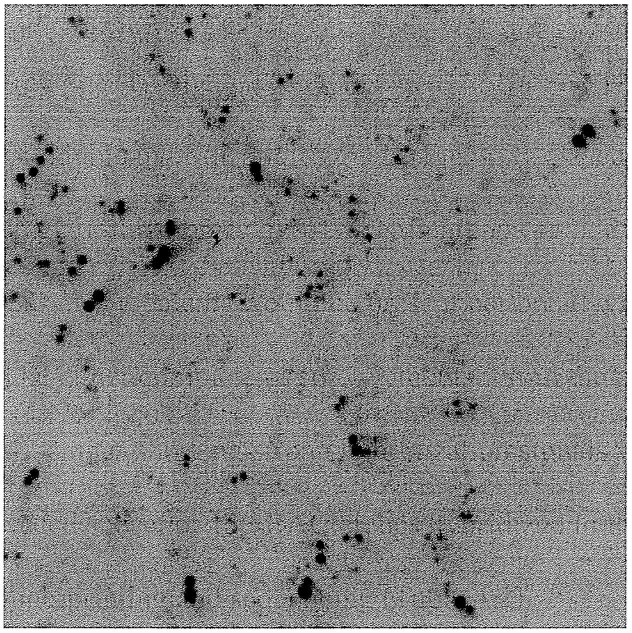


Figure 4.2 High density filter of the Glenlea BAC library hybridized with a Sasanda retroelement LTR probe shows multiple double hybridization signals. Each of the 24 filters contained 27,648 BAC clones printed in 12 distinct duplicate patterns.

4.4.2 Fingerprinting

To identify overlapping clones originating from the same locus, fingerprinting and contig assembly was performed on the 133 BAC clone subset yielding six contigs comprising 18 clones and 115 singletons (Fig. 4.3). Sequence comparison between the LTRs and/or RT sequences across clones from a same contig confirmed that they represented the same locus (data not shown). From each of the six contigs, only one representative clone (the largest one) was retained for analysis of the LTR and RT domains and the remaining 12 clones namely TaBAC181N16, TaBAC588C14, TaBAC617N18, TaBAC640G24, TaBAC891G5, TaBAC951E2, TaBAC989I1, TaBAC1004A19, TaBAC1048A3, TaBAC1081K7, TaBAC1556I7 and TaBAC1651C17 were not considered for the clustering analysis. From here onward, results will be restricted to the 121 BAC subset that hybridized to both Sasanda-LTR and Sasanda-RT probes and represented different loci.

Table 4.2 BAC clones that hybridized to LTR probe and subsets that hybridized to RT probe, were fingerprinted and from which LTR and/or RT sequences were obtained

SI.No	Hybridized to LTR probe	Hybridized to RT probe	Redundant after finger-printing and assembly	RT sequences generated	Left and right LTR sequences generated	Left LTR sequences generated	Right LTR sequences generated
1	TaBAC6F23						
2	TaBAC16G16						
3	TaBAC36D16						
4	TaBAC59H24						
5	TaBAC71N18						
6	TaBAC78A4						
7	TaBAC109M19						
8	TaBAC125L13						
9	TaBAC138B12						
10	TaBAC140E22						
11	TaBAC181L6	+					
12	TaBAC181N16	+	+				
13	TaBAC191L9	+				+	
14	TaBAC204J7	+		+	+		
15	TaBAC205J24						
16	TaBAC208K22						
17	TaBAC212D1	+		+			
18	TaBAC214C1						
19	TaBAC274C23	+		+			
20	TaBAC275E2	+		+			+
21	TaBAC287G10						
22	TaBAC304E8						
23	TaBAC306B4						

SI.No	Hybridized to LTR probe	Hybridized to RT probe	Redundant after finger-printing and assembly	RT sequences generated	Left and right LTR sequences generated	Left LTR sequences generated	Right LTR sequences generated
24	TaBAC330O17						
25	TaBAC336O13	+		+			+
26	TaBAC346P4	+					+
27	TaBAC348P21						
28	TaBAC352A9	+					
29	TaBAC357F24	+		+			
30	TaBAC359L5	+		+	+		
31	TaBAC442O8						
32	TaBAC451J13						
33	TaBAC461F9						
34	TaBAC464A20						
35	TaBAC468A21						
36	TaBAC480I13						
37	TaBAC472D17						
38	TaBAC487N11						
39	TaBAC502P8						
40	TaBAC503I5						
41	TaBAC504A1						
42	TaBAC507O3						
43	TaBAC519D14						
44	TaBAC519B24						
45	TaBAC538O22						
46	TaBAC530C17						
47	TaBAC561O5						
48	TaBAC559D10						
49	TaBAC559J22						

SI.No	Hybridized to LTR probe	Hybridized to RT probe	Redundant after finger-printing and assembly	RT sequences generated	Left and right LTR sequences generated	Left LTR sequences generated	Right LTR sequences generated
50	TaBAC576J12						
51	TaBAC588C14	+	+				
52	TaBAC600N2	+		+	+		
53	TaBAC612M14	+			+		
54	TaBAC612H6	+		+	+		
55	TaBAC602I12	+		+	+		
56	TaBAC612E21	+		+	+		
57	TaBAC608F22	+					+
58	TaBAC623M2	+		+	+		
59	TaBAC617N18	+	+				
60	TaBAC625M9						
61	TaBAC628E7						
62	TaBAC628L21						
63	TaBAC644P5	+			+		
64	TaBAC640G24	+	+				
65	TaBAC640B23	+			+		
66	TaBAC678L11	+			+		
67	TaBAC702A7	+		+		+	
68	TaBAC709C2	+		+	+		
69	TaBAC659E16	+		+	+		
70	TaBAC730N12	+		+	+		
71	TaBAC726O24	+		+	+		
72	TaBAC780I4	+		+	+		
73	TaBAC772M17	+		+	+		
74	TaBAC770D8						
75	TaBAC772A9	+		+		+	

SI.No	Hybridized to LTR probe	Hybridized to RT probe	Redundant after finger-printing and assembly	RT sequences generated	Left and right LTR sequences generated	Left LTR sequences generated	Right LTR sequences generated
76	TaBAC782E3	+		+	+		
77	TaBAC792D8						
78	TaBAC784N22	+		+			+
79	TaBAC785O24	+		+	+		
80	TaBAC819A2						
81	TaBAC832M5						
82	TaBAC834F9						
83	TaBAC833L19						
84	TaBAC832D22						
85	TaBAC840A2	+		+	+		
86	TaBAC852O1						
87	TaBAC850C21						
88	TaBAC855P2	+			+		
89	TaBAC864P7	+		+	+		
90	TaBAC867B11						
91	TaBAC872B17	+		+			+
92	TaBAC886C2	+		+	+		
93	TaBAC881I17	+		+	+		
94	TaBAC882J20	+		+	+		
95	TaBAC910O9	+		+	+		
96	TaBAC903M11						
97	TaBAC903N16						
98	TaBAC891G5	+	+				
99	TaBAC919I15	+		+		+	
100	TaBAC921D8	+		+	+		
101	TaBAC926H11	+		+			

SI.No	Hybridized to LTR probe	Hybridized to RT probe	Redundant after finger-printing and assembly	RT sequences generated	Left and right LTR sequences generated	Left LTR sequences generated	Right LTR sequences generated
102	TaBAC929P15						
103	TaBAC933C16	+		+	+		
104	TaBAC926L16	+		+			+
105	TaBAC932H16	+		+	+		
106	TaBAC932N2	+				+	
107	TaBAC895G18						
108	TaBAC942P8	+		+		+	
109	TaBAC940K21						
110	TaBAC951O2	+		+	+		
111	TaBAC951E2	+	+				
112	TaBAC972H2	+		+	+		
113	TaBAC983L9	+		+		+	
114	TaBAC989M2	+		+	+		
115	TaBAC989I1	+	+				
116	TaBAC985C2	+		+	+		
117	TaBAC991C5	+		+	+		
118	TaBAC994D11	+			+		
119	TaBAC992C21	+		+	+		
120	TaBAC989D24	+		+	+		
121	TaBAC997P14	+		+	+		
122	TaBAC1007L14	+		+	+		
123	TaBAC999A15	+		+	+		
124	TaBAC1004A19	+	+				
125	TaBAC1002C20						
126	TaBAC1001O24						
127	TaBAC1018I4						

SI.No	Hybridized to LTR probe	Hybridized to RT probe	Redundant after finger-printing and assembly	RT sequences generated	Left and right LTR sequences generated	Left LTR sequences generated	Right LTR sequences generated
128	TaBAC1012B24						
129	TaBAC1029D2						
130	TaBAC1022J6						
131	TaBAC1038K3						
132	TaBAC1038K6						
133	TaBAC1038M9						
134	TaBAC1038F5						
135	TaBAC1041C4						
136	TaBAC1041D7						
137	TaBAC1039E18						
138	TaBAC1048A3	+	+				
139	TaBAC1046G15	+		+	+		
140	TaBAC1047G22						
141	TaBAC1046E20						
142	TaBAC1065D4						
143	TaBAC1072H10	+			+		
144	TaBAC1081P11	+		+	+		
145	TaBAC1090A6	+		+	+		
146	TaBAC1081K7	+	+				
147	TaBAC1085H3	+		+	+		
148	TaBAC1104J1	+		+	+		
149	TaBAC1101J4						
150	TaBAC1102H1	+		+	+		
151	TaBAC1090F24						
152	TaBAC1103E14	+		+			
153	TaBAC1102G21						

SI.No	Hybridized to LTR probe	Hybridized to RT probe	Redundant after finger-printing and assembly	RT sequences generated	Left and right LTR sequences generated	Left LTR sequences generated	Right LTR sequences generated
154	TaBAC1120P16	+		+	+		
155	TaBAC1114N8	+		+	+		
156	TaBAC1107J10						
157	TaBAC1122H12	+		+	+		
158	TaBAC1114B6	+		+	+		
159	TaBAC1109D11						
160	TaBAC1114A10	+		+	+		
161	TaBAC1151P5	+		+		+	
162	TaBAC1147P12	+			+		
163	TaBAC1117B19	+		+	+		
164	TaBAC1126A22						
165	TaBAC1139H17						
166	TaBAC1133C17						
167	TaBAC1150L3	+		+	+		
168	TaBAC1144J3						
169	TaBAC1151D2	+				+	
170	TaBAC1145I7						
171	TaBAC1147I22						
172	TaBAC1147D24	+				+	
173	TaBAC1196E2	+		+	+		
174	TaBAC1190G14						
175	TaBAC1222P19	+		+	+		
176	TaBAC1204G8	+		+	+		
177	TaBAC1205G11						
178	TaBAC1215C6	+		+	+		
179	TaBAC1223D13	+		+		+	

SI.No	Hybridized to LTR probe	Hybridized to RT probe	Redundant after finger-printing and assembly	RT sequences generated	Left and right LTR sequences generated	Left LTR sequences generated	Right LTR sequences generated
180	TaBAC1223D17	+					
181	TaBAC1216B16	+		+	+		
182	TaBAC1264I14	+		+			
183	TaBAC1264F15	+			+		
184	TaBAC1273F12	+		+	+		
185	TaBAC1309O22						
186	TaBAC1322B13						
187	TaBAC1327B23						
188	TaBAC1358K3						
189	TaBAC1398D5	+		+	+		
190	TaBAC1403I2	+		+	+		
191	TaBAC1401J6	+		+	+		
192	TaBAC1428B22	+		+	+		
193	TaBAC1438J6						
194	TaBAC1452G16	+		+			+
195	TaBAC1455B6	+		+	+		
196	TaBAC1455B10	+		+	+		
197	TaBAC1462A11	+		+	+		
198	TaBAC1488L9	+			+		
199	TaBAC1491P3	+		+	+		
200	TaBAC1489B10						
201	TaBAC1495E19	+		+	+		
202	TaBAC1490B22	+		+	+		
203	TaBAC1504B2	+		+	+		
204	TaBAC1516I1						
205	TaBAC1522P18						

SI.No	Hybridized to LTR probe	Hybridized to RT probe	Redundant after finger-printing and assembly	RT sequences generated	Left and right LTR sequences generated	Left LTR sequences generated	Right LTR sequences generated
206	TaBAC1531I3	+		+	+		
207	TaBAC1525B3	+		+	+		
208	TaBAC1542N10	+		+	+		
209	TaBAC1537B1						
210	TaBAC1543J17	+			+		
211	TaBAC1546A17						
212	TaBAC1560M5	+		+	+		
213	TaBAC1556I7	+	+				
214	TaBAC1553P17	+		+	+		
215	TaBAC1551N22	+		+		+	
216	TaBAC1558B22	+		+		+	
217	TaBAC1563P23						
218	TaBAC1571J8						
219	TaBAC1568I16						
220	TaBAC1570J20						
221	TaBAC1570E14						
222	TaBAC1583B2	+		+	+		
223	TaBAC1576J13	+		+	+		
224	TaBAC1576K24	+		+	+		
225	TaBAC1579B23	+			+		
226	TaBAC1588H2	+		+	+		
227	TaBAC1598H3	+		+	+		
228	TaBAC1598I7	+		+	+		
229	TaBAC1610P5	+		+	+		
230	TaBAC1618I22						
231	TaBAC1639J5						

SI.No	Hybridized to LTR probe	Hybridized to RT probe	Redundant after finger-printing and assembly	RT sequences generated	Left and right LTR sequences generated	Left LTR sequences generated	Right LTR sequences generated
232	TaBAC1641G23	+		+			+
233	TaBAC1645M12	+		+	+		
234	TaBAC1647G17	+		+			+
235	TaBAC1651C17	+	+				
236	TaBAC1678D2						
237	TaBAC1678H4						
238	TaBAC1678L14						
239	TaBAC1679B20						
240	TaBAC1698H12						
241	TaBAC1703J18						
242	TaBAC1708J19						

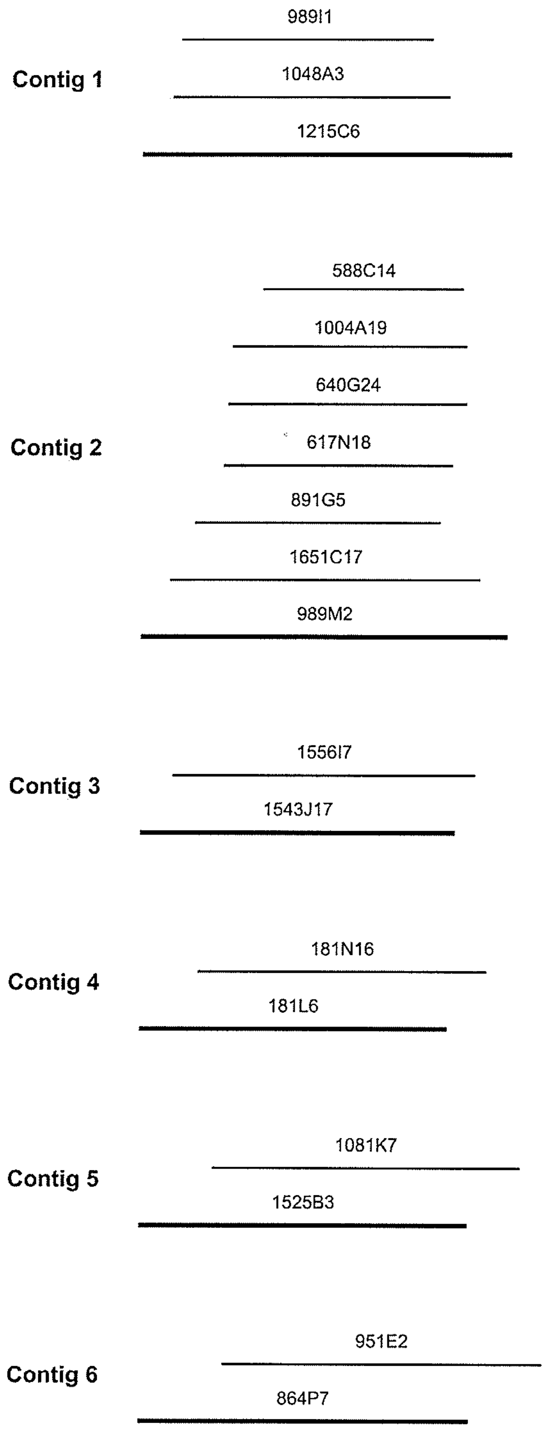


Figure. 4.3 Contig assemblies of BAC clones representing the same genomic location. The thicker line clones were included in the clustering analysis while the remaining 12 were removed because they represented the same locus/element.

4.4.3 Amplification and sequencing of RT domains

The primer pair Sasanda_3718F and Sasanda_3982R was used to amplify a 265 bp region of the RT domain. The purified amplicons were sequenced with the same two primers and high quality overlapping sequences of 240 bp were obtained from 100 of the 121 BAC clone subset (Table 4.2). Sequences from the remaining 21 clones were either poor quality and/or short (from 159 bp to 214bp) and were not included in the analysis. Overall percent identity of DNA sequences from the 100 clone subset was estimated at 97.65%. The alignment of the deduced amino acid sequences also indicated high sequence homogeneity with an overall identity of 94.50% (Fig. 4.4).

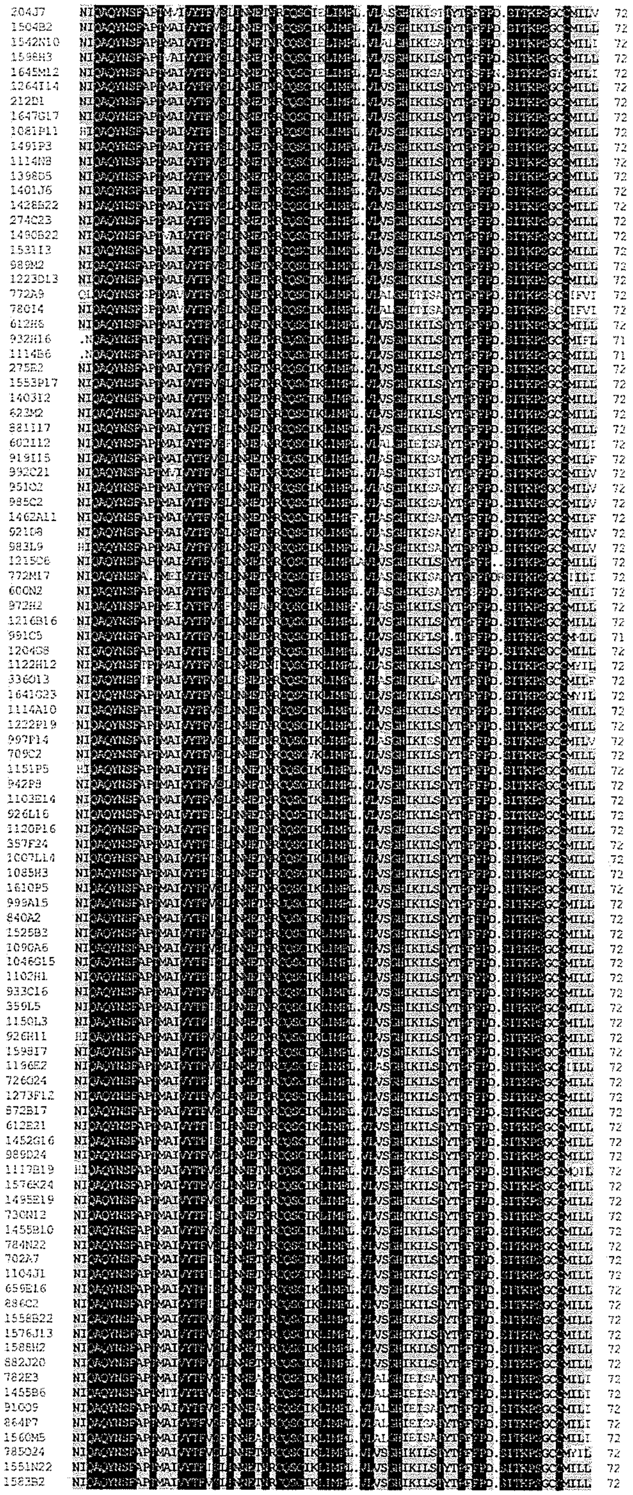


Figure 4.4 Alignment of the deduced amino acid sequences of the partial RT domain of Sasanda elements from 100 BAC clones shows an overall identity of 94.5%.

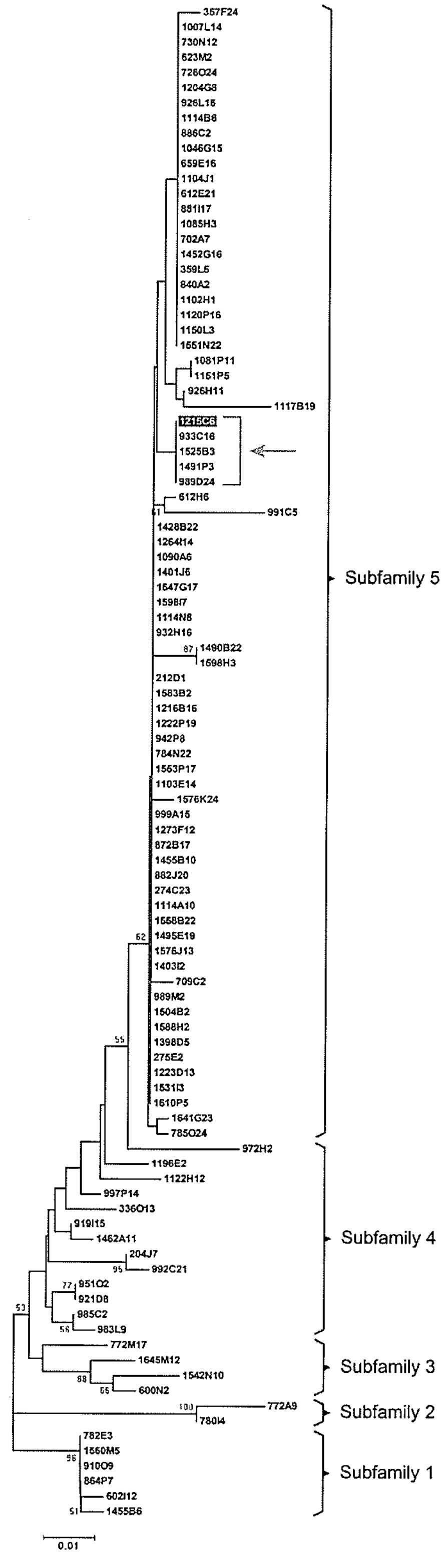
4.4.4 Amplification and sequencing of left and right LTR domains

Left and right LTRs were amplified using an internal LTR primer and a primer designed in the core region of the retroelement (Table 4.1). The 1048 bp amplicon included 931 bp of the left LTR while the 824 bp amplicon covered 805 bp of the right LTR. Because LTRs are direct repeats, the common overlapping sequence between the two amplicons was 757 bp (Fig. 4.1). Left and right LTRs were successfully amplified from all BAC clone templates although amplification yield varied across samples. Up to nine primers were used to obtain multiple overlapping sequences for each clone with ~4X coverage for each nucleotide. This coverage level was deemed necessary to ensure accurate base-calling and identify site mutations with high level of confidence. High quality deep coverage assemblies were obtained for 102 left LTRs and 99 right LTRs. A total of 89 BAC clones generated high quality left and right LTRs, 13 only a high quality left LTR, 10 only a right LTR and sequences of the remaining 9 clones could not be assembled for either LTR because of poor quality of sequence reads, insufficient length and/or lack of overlap between reads. The common high quality 567 bp regions of the left and right LTRs corresponding to the coordinates 223 to 790 and 220 to 786 were extracted from the left and right LTRs, respectively. Left LTR sequences shared an identity of 96.0% among themselves and the right LTR sequences were 96.7% identical. Between LTR pairs of individual elements, the identity percentage varied from 87.10% (in Sasanda_1531I3) to 100% in 31 elements (Table 4.3). Among 89 elements having both LTRs, the only other element with <90% identity between left and right LTR was Sasanda_612H6 (88.7%). In 35 elements the LTR pairs shared an identity of > 99%. An overall average of 95.84% identity was observed among the LTR pairs.

4.4.5 Phylogenetic inference based on the RT domain

Consensus DNA sequences of 240 bp each representing the RT domain of Sasanda elements from 100 BAC clones were aligned using ClustalW implemented in MEGA4 and used for the construction of a neighbour-joining tree. To deduce the evolutionary relationships among family members, transition, transversion and multiple substitutions at the same site as estimated by the Kimura II parameter model and the complete deletion option of MEGA4 were taken into account. The optimal tree with a branch length sum of 0.3345 was constructed based on the 233 common base positions of the final dataset (Fig. 4.5). An overall mean distance of 0.020 substitutions among sequences was observed. Based on the observed topology supported by a bootstrap value of >50%, five sub-families were identified indicating multiple bursts of transpositions (Fig. 4.5). The branch length is an indicator of evolutionary distance and therefore, clustering of several members with a branch length of zero indicated that the sequences were nearly identical to the sequences originating from the same node, representing very recent amplification. Subfamily-1 consists of six members with a shared identity of 99.86%. Subfamily-2 consists of only two members which are 98.72% identical. Subfamilies 3 and 4 with four and 13 members respectively both had 98.63% identities. Subfamily-5 was the largest with 75 members sharing 98.85% identity. When consensus sequences representing each subfamily were compared, the percent identities ranged from 92.77% (between subfamilies 2 and 5) to 98.30 % (between subfamilies 3 and 4). At least four more paralogous copies (Sasanda_933C16, Sasanda_1525B3, Sasanda_1491P3, Sasanda_989D24) of the insertion found at the *Glu-B1* locus (Sasanda_1215C6) were identified from the clusters.

Figure 4.5 Phylogenetic tree based on 100 RT sequences of Sasanda retroelements inferred using the neighbour-joining method implemented in MEGA4. Evolutionary distances are computed using the Kimura-2 parameter method. The percentage of replicate trees in which the associated elements clustered together in the bootstrap test (1000 replicates) is shown next to the branches only when the value is >50. The tree is drawn to scale with branch lengths indicating sequence divergence. Sasanda EU157184-1 is highlighted and the arrow indicates clusters of paralogous elements.


357F24
1007L14
730N12
623M2
726O24
1204G8
926L16
1114B8
886C2
1046G15
659E16
1104J1
612E21
881I17
1085H3
702A7
1452G16
359L6
840A2
1102H1
1120P16
1150L3
1651N22
1081P11
1151P5
926H11
1117B19
1215C6
933C16
1525B3
1491P3
989D24
612H6
61
991C5
1428B22
1264I14
1090A6
1401J6
1647G17
1598I7
1114N8
932H16
87
1490B22
1598H3
212D1
1583B2
1216B16
1222P19
942P8
784N22
1553P17
1103E14
1576K24
999A15
1273F12
872B17
1455B10
882J20
274C23
1114A10
1558B22
62
1495E19
1576J13
1403I2
709C2
989M2
1504B2
1588H2
1398D5
275E2
1223D13
1531I3
1610P5
1641G23
785O24
55
972H2
1196E2
1122H12
997P14
336O13
919I15
1462A11
204J7
95
992C21
77
951O2
921D8
985C2
58
983L9
53
772M17
1645M12
1542N10
88
66
600N2
100
772A9
780I4
782E3
1560M5
96
910O9
864P7
602I12
51
1455B6
0.01
Subfamily 5
Subfamily 4
Subfamily 3
Subfamily 2
Subfamily 1

4.4.6 Phylogenetic inference based on the LTR domains

Phylogenetic trees based on the LTR domains may provide better insight into the evolutionary relationships among the closely related members of a family than RT domain based phylogenetic inference. From the original 567 bp long sequence representing the left LTR, only 460 positions were informative for phylogenetic inference based on the Kimura II parameter substitution model obtained by the neighbour joining method implemented in MEGA4. The optimal tree with the sum of branch length of 0.3032, generated with 460 positions in the final dataset representing left LTR, is shown in the Fig. 4.6. Similarly, the optimal tree with a sum of branch length of 0.3157, generated with 501 sites in the final dataset from the right LTR, is shown in Fig. 4.7. From 567 bp of the original dataset, only 501 sites were informative for phylogenetic inference because all positions containing gaps were eliminated from the dataset with the complete delete option of MEGA4. Broadly, nine subfamilies can be inferred from each of the two phylogenetic trees generated from the left and right LTR domains of the 89 members though the clusters did not contain exactly the same elements (Fig. 4.6 and 4.7). These 89 sequences represent the same region of the left and right LTRs in the same set of clones. All five paralogues identified based on the RT domain including Sasanda_1215C6 present at the *Glu-B1* locus also clustered together in the phylogenetic trees constructed independently based on left and right LTR sequences, though five more copies (Sasanda_1553P17, Sasanda_1403I2, Sasanda_951O2, Sasanda_1542N10, and Sasanda_122P19) were added to the cluster. A few of the sequences could not be clustered into subfamilies.

Figure 4.6 Phylogenetic tree based on 89 left LTR sequences of Sasanda retroelements inferred using the neighbour-joining method implemented in MEGA4. Evolutionary distances are computed using the Kimura-2 parameter method. The percentage of replicate trees in which the associated elements clustered together in the bootstrap test (1000 replicates) is shown next to the branches only when the value is >50. The tree is drawn to scale with branch lengths indicating sequence divergence. Sasanda EU157184-1 is highlighted and the arrow indicates clusters of paralogous elements..

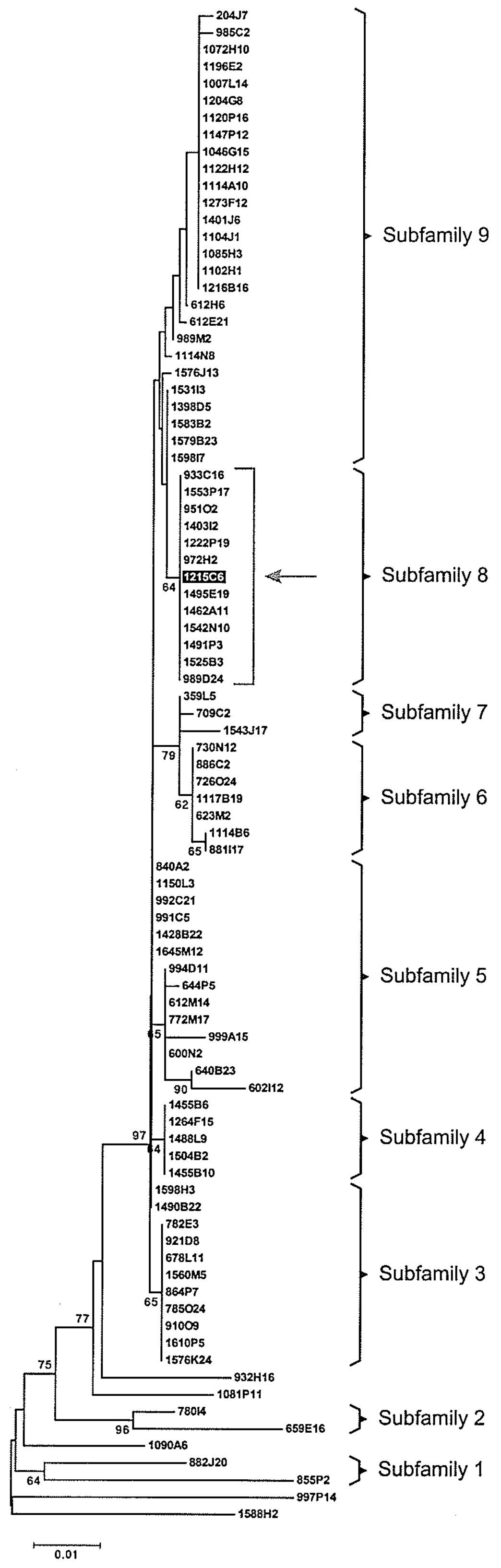
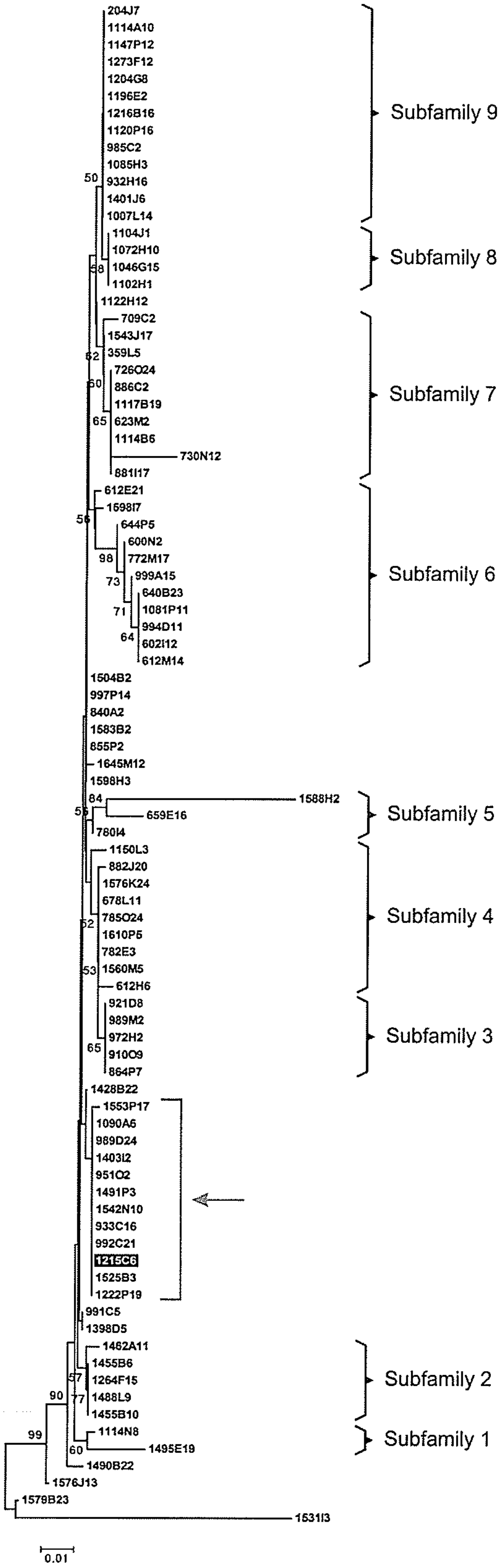

204J7
985C2
1072H10
1196E2
1007L14
1204G8
1120P16
1147P12
1046G15
1122H12
1114A10
1273F12
1401J6
1104J1
1085H3
1102H1
1216B16
612H6
612E21
989M2
1114N8
1576J13
1531I3
1398D5
1583B2
1579B23
1598I7
Subfamily 9
933C16
1553P17
951O2
1403I2
1222P19
972H2
1215C6
64
1495E19
1462A11
1542N10
1491P3
1525B3
989D24
Subfamily 8
359L5
709C2
1543J17
Subfamily 7
730N12
79
886C2
726O24
1117B19
62
623M2
1114B6
65
881I17
Subfamily 6
840A2
1150L3
992C21
991C5
1428B22
1645M12
994D11
644P5
612M14
772M17
65
999A15
600N2
640B23
90
602I12
Subfamily 5
1455B6
1264F15
97
1488L9
64
1504B2
1455B10
Subfamily 4
1598H3
1490B22
782E3
921D8
678L11
1560M5
864P7
65
785O24
910O9
1610P5
1576K24
Subfamily 3
932H16
77
1081P11
75
780I4
96
659E16
Subfamily 2
1090A6
882J20
64
855P2
Subfamily 1
997P14
1588H2
0.01

Figure 4.7 Phylogenetic tree based on 89 right LTR sequences.of Sasanda retroelements inferred using the neighbour-joining method implemented in MEGA4. Evolutionary distances are computed using the Kimura-2 parameter method. The percentage of replicate trees in which the associated elements clustered together in the bootstrap test (1000 replicates) is shown next to the branches only when the value is >50. The tree is drawn to scale with branch lengths indicating sequence divergence. Sasanda EU157184-1 is highlighted and the arrow indicates clusters of paralogous elements.


204J7
1114A10
1147P12
1273F12
1204G8
1196E2
1216B16
1120P16
985C2
1085H3
50
932H16
1401J6
1007L14
1104J1
1072H10
58
1046G15
1102H1
1122H12
709C2
1543J17
62
359L5
726O24
60
886C2
1117B19
65
623M2
1114B6
730N12
881I17
612E21
1598I7
56
644P5
600N2
98
772M17
73
999A15
640B23
71
1081P11
994D11
64
602I12
612M14
1504B2
997P14
840A2
1583B2
855P2
1645M12
1598H3
84
1588H2
56
659E16
780I4
1150L3
882J20
1576K24
678L11
785O24
52
1610P5
782E3
53
1560M5
612H6
921D8
989M2
972H2
65
910O9
864P7
1428B22
1553P17
1090A6
989D24
1403I2
951O2
1491P3
1542N10
933C16
992C21
1215C6
1525B3
1222P19
991C5
1398D5
1462A11
1455B6
57
1264F15
90
77
1488L9
1455B10
99
1114N8
60
1495E19
1490B22
1576J13
1579B23
1531I3
0.01
Subfamily 9
Subfamily 8
Subfamily 7
Subfamily 6
Subfamily 5
Subfamily 4
Subfamily 3
Subfamily 2
Subfamily 1

4.4.7 Calibration of the RT based phylogenetic tree

To estimate the time of retrotransposition activities represented by the tree nodes, the phylogenetic tree was linearised and calibrated using the molecular clock of 2×10^{-8} substitutions/site/year observed in rice retrotransposons (Vitte et al. 2004). Our data suggests that at least eight bursts of retrotransposition leading to the amplification of Sasanda elements occurred in the last 1.2 million years indicating that the element precedes the advent of tetraploid wheat (Fig. 4.8). Also, the preponderance of elements with identical LTR sequence indicates recent transposition activity.

4.4.8 Insertion times

Evolutionary divergence of the aligned LTR pairs of 89 members having both LTRs were estimated based on the observed differences in the nucleotide sites corrected for multiple substitutions at same sites using the Kimura II parameter model (Table 4.3). These divergence estimates were used for dating the insertions which indicated periodic waves of amplifications from 1.7 million years ago (MYA) to now and also corroborated the independent inference obtained from the molecular clock based calibration of the phylogenetic tree generated based on the RT domain. The oldest element characterized was Sasanda_1531I3. Its LTRs shared 87.10% identity and it was dated at 2.9 ± 0.4 MYA coinciding with the divergence of *Triticum* and *Aegilops* approximately 3 MYA (Table 4.3). In 49 elements the left and right LTRs were identical indicating that the elements were active very recently (Fig. 4.9). As well, an additional 18 members also displayed high identity LTRs suggesting their comparatively recent amplification i.e., within 0.2 MYA. The distribution of elements (Fig. 4.9) also indicated the occurrence of waves of retrotransposition in punctuated intervals rather than continuous amplification pattern of elements.

Figure 4.8 Linearized phylogenetic tree of 100 RT sequences of Sasanda retroelements. The calibration to convert the evolutionary distance into time was based on the molecular clock of $2x10^{-8}$ substitutions per site per year, as described by Vitte et al. 2004. Arrows indicate major bursts of retrotranspositions. Time scale is indicated at the bottom.

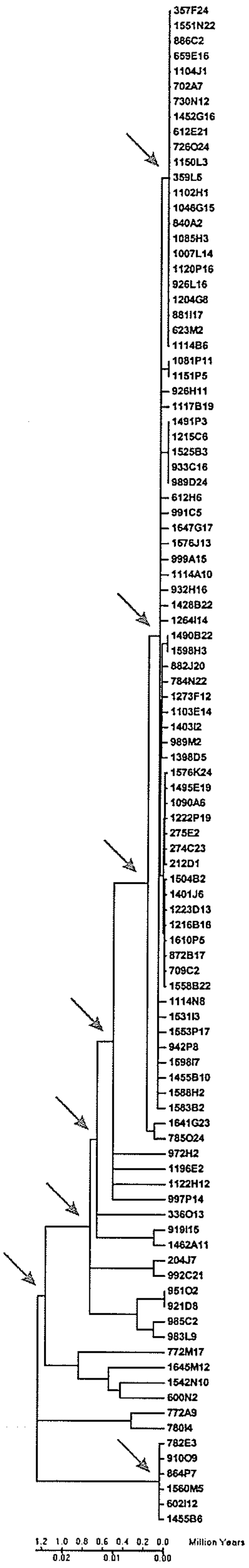

357F24
1551N22
886C2
659E16
1104J1
702A7
730N12
1452G16
612E21
726O24
1150L3
359L5
1102H1
1046G15
840A2
1085H3
1007L14
1120P16
926L16
1204G8
881I17
623M2
1114B6
1081P11
1151P5
926H11
1117B19
1491P3
1215C6
1525B3
933C16
989D24
612H6
991C5
1647G17
1576J13
999A15
1114A10
932H16
1428B22
1264I14
1490B22
1598H3
882J20
784N22
1273F12
1103E14
1403I2
989M2
1398D5
1576K24
1495E19
1090A6
1222P19
275E2
274C23
212D1
1504B2
1401J6
1223D13
1216B16
1610P5
872B17
709C2
1558B22
1114N8
1531I3
1553P17
942P8
1598I7
1455B10
1588H2
1583B2
1641G23
785O24
972H2
1196E2
1122H12
997P14
336O13
919I15
1462A11
204J7
992C21
951O2
921D8
985C2
983L9
772M17
1645M12
1542N10
600N2
772A9
780I4
782E3
910O9
864P7
1560M5
602I12
1455B6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Million Years
0.02
0.01
0.00

Table 4.3 Identity percentage, estimated LTR divergence and dating of insertion times of Sasanda elements of individual clones

SI.No	Clone Name	Left LTR sequence length (bp)	Right LTR sequence length (bp)	% identity between the two LTRs	LTR divergence[1]	Std Error for LTR divergence[1]	Insertion time[2] (MYA)	Std Error for Insertion time[2] (MYA)
1	TaBAC204J7	566	566	99.65	0.002	0.002	0.05	0.05
2	TaBAC359L5	572	572	100	0	0	0	0
3	TaBAC600N2	566	566	99.29	0.009	0.005	0.23	0.13
4	TaBAC602I12	566	566	100	0	0	0	0
5	TaBAC612E21	566	566	99.47	0.007	0.004	0.18	0.1
6	TaBAC612H6	528	566	88.67	0.009	0.005	0.23	0.13
7	TaBAC612M14	566	566	98.94	0.014	0.006	0.35	0.15
8	TaBAC623M2	577	577	100	0	0	0	0
9	TaBAC640B23	566	566	99.29	0.009	0.005	0.23	0.13
10	TaBAC644P5	566	565	99.12	0.007	0.004	0.18	0.1
11	TaBAC659E16	557	557	94.82	0.026	0.008	0.65	0.2
12	TaBAC678L11	566	566	100	0	0	0	0
13	TaBAC709C2	572	572	99.47	0.005	0.003	0.13	0.08
14	TaBAC726O24	577	577	100	0	0	0	0
15	TaBAC730N12	577	577	98.44	0.021	0.007	0.53	0.18
16	TaBAC772M17	566	566	99.29	0.009	0.005	0.23	0.13
17	TaBAC780I4	560	551	94.32	0.026	0.008	0.65	0.2
18	TaBAC782E3	566	566	100	0	0	0	0
19	TaBAC785O24	566	566	100	0	0	0	0
20	TaBAC840A2	566	571	99.12	0	0	0	0
21	TaBAC855P2	566	565	93.81	0.068	0.013	1.7	0.33
22	TaBAC864P7	566	566	100	0	0	0	0
23	TaBAC881I17	577	577	99.83	0	0	0	0

SI.No	Clone Name	Left LTR sequence length (bp)	Right LTR sequence length (bp)	% identity between the two LTRs	LTR divergence[1]	Std Error for LTR divergence[1]	Insertion time[2] (MYA)	Std Error for Insertion time[2] (MYA)
24	TaBAC882J20	566	570	94.21	0.048	0.011	1.2	0.28
25	TaBAC886C2	577	577	100	0	0	0	0
26	TaBAC910O9	566	566	100	0	0	0	0
27	TaBAC921D8	566	566	100	0	0	0	0
28	TaBAC932H16	566	553	94.51	0.033	0.009	0.83	0.23
29	TaBAC933C16	566	566	100	0	0	0	0
30	TaBAC951O2	566	565	99.82	0	0	0	0
31	TaBAC972H2	566	566	99.12	0.009	0.005	0.23	0.13
32	TaBAC985C2	566	565	99.65	0.002	0.002	0.05	0.05
33	TaBAC989D24	556	566	98.23	0	0	0	0
34	TaBAC989M2	566	566	99.65	0.005	0.003	0.13	0.08
35	TaBAC991C5	566	566	99.82	0.002	0.002	0.05	0.05
36	TaBAC992C21	556	566	97.88	0.005	0.003	0.13	0.08
37	TaBAC994D11	566	566	98.94	0.014	0.006	0.35	0.15
38	TaBAC997P14	566	550	91.68	0.055	0.012	1.38	0.3
39	TaBAC999A15	566	566	98.58	0.016	0.006	0.4	0.15
40	TaBAC1007L14	566	566	100	0	0	0	0
41	TaBAC1046G15	552	552	99.64	0	0	0	0
42	TaBAC1072H10	552	552	99.64	0	0	0	0
43	TaBAC1081P11	566	546	92.57	0.038	0.009	0.95	0.23
44	TaBAC1085H3	566	567	99.65	0	0	0	0
45	TaBAC1090A6	566	563	95.4	0.045	0.01	1.13	0.25
46	TaBAC1102H1	552	552	99.64	0	0	0	0
47	TaBAC1104J1	552	552	99.64	0	0	0	0
48	TaBAC1114A10	566	566	99.29	0	0	0	0

SI.No	Clone Name	Left LTR sequence length (bp)	Right LTR sequence length (bp)	% identity between the two LTRs	LTR divergence[1]	Std Error for LTR divergence[1]	Insertion time[2] (MYA)	Std Error for Insertion time[2] (MYA)
49	TaBAC1114B6	577	574	99.48	0	0	0	0
50	TaBAC1114N8	566	566	99.29	0.009	0.005	0.23	0.13
51	TaBAC1117B19	577	577	100	0	0	0	0
52	TaBAC1120P16	566	566	100	0	0	0	0
53	TaBAC1122H12	567	566	99.65	0.002	0.002	0.05	0.05
54	TaBAC1147P12	566	566	99.82	0	0	0	0
55	TaBAC1150L3	528	566	92.92	0	0	0	0
56	TaBAC1196E2	566	566	99.65	0	0	0	0
57	TaBAC1204G8	566	566	99.82	0	0	0	0
58	TaBAC1215C6	566	566	100	0	0	0	0
59	TaBAC1216B16	566	566	99.82	0	0	0	0
60	TaBAC1222P19	566	566	100	0	0	0	0
61	TaBAC1264F15	572	572	100	0	0	0	0
62	TaBAC1273F12	566	566	99.12	0	0	0	0
63	TaBAC1398D5	566	566	100	0	0	0	0
64	TaBAC1401J6	566	566	100	0	0	0	0
65	TaBAC1403I2	566	566	100	0	0	0	0
66	TaBAC1428B22	566	566	99.82	0.002	0.002	0.05	0.05
67	TaBAC1455B6	565	565	100	0	0	0	0
68	TaBAC1455B10	565	565	100	0	0	0	0
69	TaBAC1462A11	565	566	97.52	0.007	0.004	0.18	0.1
70	TaBAC1488L9	565	565	100	0	0	0	0
71	TaBAC1490B22	565	565	99.29	0.002	0.002	0.05	0.05
72	TaBAC1491P3	566	566	100	0	0	0	0
73	TaBAC1495E19	565	566	94.16	0.021	0.007	0.53	0.18

SI.No	Clone Name	Left LTR sequence length (bp)	Right LTR sequence length (bp)	% identity between the two LTRs	LTR divergence[1]	Std Error for LTR divergence[1]	Insertion time[2] (MYA)	Std Error for Insertion time[2] (MYA)
74	TaBAC1504B2	566	565	99.47	0.005	0.003	0.13	0.08
75	TaBAC1525B3	566	566	100	0	0	0	0
76	TaBAC1531I3	565	566	87.1	0.116	0.017	2.9	0.43
77	TaBAC1542N10	566	566	100	0	0	0	0
78	TaBAC1543J17	572	572	99.47	0.007	0.004	0.18	0.1
79	TaBAC1553P17	566	566	99.82	0.002	0.002	0.05	0.05
80	TaBAC1560M5	566	566	100	0	0	0	0
81	TaBAC1576J13	566	566	97.35	0.007	0.004	0.18	0.1
82	TaBAC1576K24	566	566	100	0	0	0	0
83	TaBAC1579B23	566	566	95.22	0.021	0.007	0.53	0.18
84	TaBAC1583B2	566	566	99.82	0	0	0	0
85	TaBAC1588H2	567	564	98.06	0.012	0.005	0.3	0.13
86	TaBAC1598H3	566	566	100	0	0	0	0
87	TaBAC1598I7	566	566	99.29	0.005	0.003	0.13	0.08
88	TaBAC1610P5	566	566	100	0	0	0	0
89	TaBAC1645M12	566	566	99.82	0.002	0.002	0.05	0.05

[1] Corrected for multiple substitutions in a site by the Kimura II parameter model

[2] Using $2x10^{-8}$ substitutions/site/year, Vitte et al. 2004

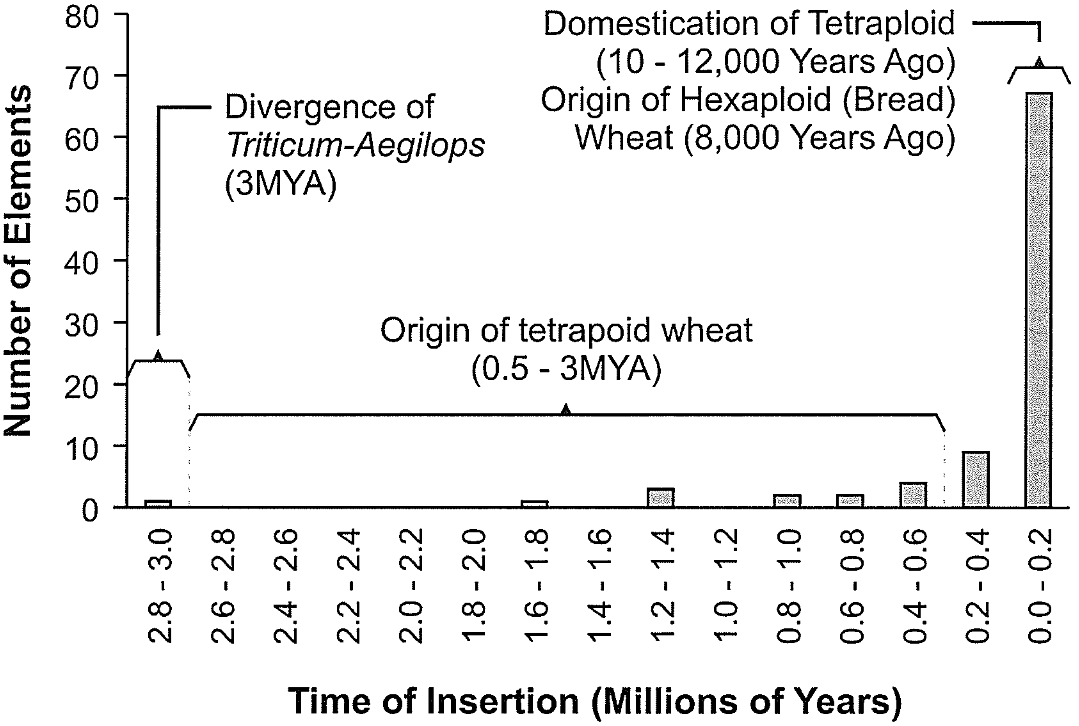


Figure 4.9 Insertion times of the Sasanda retroelement family members.These estimates based on the observed divergence of LTR pairs of 89 elements indicate its presence in the genome at the time of divergence of *Triticum* and *Aegilops* (~3 MYA) and a major burst of recent transposition corresponding to the domestication of tetraploid wheat and the advent of hexaploid wheat.

4.5 Discussion

Mobile genetic elements play a major role in the evolution of structure and function of eukaryotic genomes (Biemont and Vieira 2006). Members of *copia* retroelements are found in diverse group of plants (Voytas et al. 1992). We have previously isolated a *copia* type retrotransposon named Sasanda_EU157184-1 from the *Glu-B1* locus of *T. aestivum* cv Glenlea (Cloutier et al. 2005; Ragupathy et al. 2008). The structural organization of the locus revealed the presence of this element between the tandem segmental duplication harbouring the HMW-GS *Bx7* gene. Absence of nucleotide changes in the LTRs of this novel element indicated its recent transposition. As well, the presence of complete non-degenerated coding region also pointed at its recent activity. In this study, we characterized the Sasanda *copia* element family and the evolutionary relationships among its members.

Among the characterized 121 copies of the Sasanda elements, 89 elements were found to be complete with both the LTRs. In addition, 13 members generated only left LTR sequences, 10 members only right LTR sequences and none from 9 BAC clones. Sequence divergence to the sequencing or amplification primers could be responsible for the failure to generate an assembly for these targets. Template quality and quantity or the presence of more than one element per clone could also account for failed sequencing reactions. Similarly, for the RT domain, 100 members produced good quality sequences, among which 77 members were found to be complete with the presence of both the RT domain and LTR pairs. Some RT sequences were rejected because they were poor or short.

BLASTn searches (Altschul et al. 1990) with the Sasanda element indicated the presence of orthologous elements in maize suggesting that the element's origin precedes the wheat ancestor-maize speciation 15-20 MYA and that vertical transmission from the ancestral genome occurred (Gill et al. 2004). Our estimated 347 copies of the Sasanda element per haploid genome represent a moderate number of copies when compared with some families in wheat (Angela and WIS), barley (BARE1) and rice (RIRE), where thousands of copies were reported (Wicker and Keller 2007; Suoniemi et al. 1996; Piegu et al. 2006). However, most *copia* elements in rice, *Triticeae* and Arabidopsis were present in low copy number, though families such as *Houba* and *Osr8* were identified to have 151 and 77 full length copies in the rice genome, respectively (Vitte and Panaud 2005; Wicker and Keller 2007). An average of ~20 copies was estimated for families such as *Tara, Adena, Osr1, Osr10, Ostonor1 and SC3* in rice. In Arabidopsis, *copia* elements have only one or two intact elements with the exception of *Atcopia78, Atcopia58* and *Anika* elements which have 8, 5 and 3 copies, respectively (Wicker and Keller 2007).

Sequence characterization of RT as well as LTR domains indicated that the Sasanda elements isolated are highly homogeneous suggesting recent amplification. Morgane, another element family recently characterized in wheat, also exhibited minor sequence variation with 95.5% identity among its members (Sabot et al. 2006). Low variability in the RT and the LTR domains indicated that the lineages inferred were found to be closely related. On the basis of branching pattern of neighbour-joining trees, we have inferred at least 5 to 9 subfamilies, among the characterized members of this *copia* family. However, the observed sequence divergence within and between subfamilies is very low. The phylogenetic tree based on RT domain is different from the phylogenetic trees based on left and right LTRs because mutations in LTRs are better tolerated than in coding sequences (Wicker and Keller 2007). In other words, the RT domain of the retroelements is more conserved than the LTR domains because of its functional importance. Hence, the rate of accumulation of mutations in the LTR domains exceeds that of the RT domain. Also, the divergence between the left and right LTR sequences (maximum 13% as in the case of Sasanda_1531I3) resulted in the differences observed in the clustering pattern of the elements, generated from the common region of LTRs obtained from the common set of elements. SanMiguel et al. (1998) suggested that the LTRs in a family can vary from 0 to 50% in their sequence identity. Bootstrap values provide an indication of the confidence in the groupings (Felsenstein 1985). However, only sufficient divergence among the sequences will provide family structure supported by very high bootstrap values (Wicker and Keller, 2007). High levels of identity among a large number of Sasanda sequences resulted in comparatively low bootstrap values and consequently lower nodal confidence in some cases.

To convert the divergence estimates into insertion times, we chose to use the mutation rate of $2x10^{-8}$ substitutions/site/year (Vitte et al. 2004) instead of $6.13x10^{-9}$ substitutions/site/year (Gaut et al. 1996) because the former, obtained from retroelements, is likely to be more accurate in this study than the latter which was derived from the *Adh* gene. Mutations in genes are less tolerated because of selection pressure hence yielding a more conservative mutation rate. Retrotransposons, not subjected to the same selective pressure, can accumulate mutations more readily because of their comparatively more neutral status (SanMiguel et al. 1998). This hypothesis was corroborated by our results where the mutation rate was higher in LTR domains (non coding) than in RT domain (coding).

The phylogenetic analyses of the RT domain and the LTR domains revealed that the majority of the Sasanda elements studied originated recently. Calibration of the RT based phylogenetic tree, to date the radiation of lineages was also supported by LTR pair based divergence estimates (Fig. 4.8 and Table 4.3). In 49 elements, LTR pairs were identical indicating recent transposition activity. To our knowledge, this is the first report of a single element family in the wheat genome with such a high number of complete elements that have not accumulated mutations in the LTRs. In maize, most retroelements were found to be amplified within the last three million years (SanMiguel et al. 1998). In rice, recent bursts of transpositions of three families of retroelements with insertion estimates from 2.9 MYA to now, were reported (Piegu et al. 2006). Wicker and Keller (2007) suggested that many families in wheat have LTRs with an identity of 93-100% corresponding to insertion times of less than 3 million years. Holligan et al. (2006) reported that in *Lotus japonicus*, ~58% of 82 *copia* families studied have LTRs, which were 98% identical, indicating highly similar members with the exception of the lineages of LINES.

Active transposable elements are a very common characteristic of *Poaceae* genomes (Vicient et al. 2001). Transposon activation may occur as a response of the genome upon exposure to both external and internal environmental stimuli causing stress (McClintock 1984). Active elements quickly affect genome restructuring thereby creating genetic variations that can counteract the negative impacts of stress even over a short evolutionary time (Wessler 1996). Abiotic stresses such as drought (Kalender et al. 2000), allopolyploidization (Levy and Feldman 2004), tissue culture (Kukuchi et al. 2003, Tang et al. 2005), introgressive hybridization with wild species (Liu and Wendel 2000), wounding and methyl jasmonate (Takeda et al. 1998), UV light (Ramallo et al. 2008) and biotic stresses such as *Fusarium* infection (Ansari et al. 2007) were correlated with the activation of transposable elements. Also, changes in the global and local methylation status and heterochromatinization of the genome may alter the repression of elements' amplification by the host machinery (Madlung and Comai 2004; Ammiraju et al. 2007).

Recent activity of the elements results in the presence of full sized copies leading to insertion site polymorphism (Flavell et al. 1998; Holligan et al. 2006). Though retroelements were estimated to have short turnover periods because they decay over time due to deletions driven by illegitimate recombination and unequal non homologous recombination (Ma et al. 2004), these recombination mechanisms leave footprints of solo LTRs leading to sequence divergence at a given locus. Retrotransposon based marker types namely, sequence specific amplification polymorphism (SSAP), retrotransposon microsatellite amplification polymorphism (REMAP), inter retroelement amplification polymorphism (IRAP) were developed for linkage and diversity analysis in crops including wheat, exploiting the ubiquitous nature of these elements (Syed and Flavell 2006; Kalendar et al. 1999; Queen et al. 2004). The presence of ~347 copies of the Sasanda element in the haploid genome with at least 89 distinct intact members can provide novel marker types to further saturate existing genetic maps (Röder et al. 1998; Somers et al. 2004). In wheat, the genetic to physical distance can be considerable mainly because the wheat genome is so large. Hence, saturation of the linkage map of wheat would be highly desirable. Besides, the presence of orthologous elements across grass genomes enhances the potential for their transferability (Sabot et al. 2004).

Genome Organization and Retrotransposon Driven Molecular Evolution of the Endosperm *Hardness* (*Ha*) Locus in *Triticum aestivum* cv Glenlea

Raja Ragupathy [1, 2] and Sylvie Cloutier [2]

[1]Department of Plant Science, Faculty of Graduate Studies,
University of Manitoba, Winnipeg, Canada R3T 2N2

[2]Cereal Research Centre, Agriculture and Agri-Food Canada,
195 Dafoe Road, Winnipeg, Canada R3T 2M9

The thesis author, Raja Ragupathy, designed, carried out the experiments, did the data analysis, interpretation and drafted the manuscript. As major advisor, Dr. Sylvie Cloutier guided the direction of the study, participated in data analysis and manuscript review.

CHAPTER 5

Genome Organization and Retrotransposon Driven Molecular Evolution of the Endosperm *Hardness* (*Ha*) Locus in *Triticum aestivum* cv Glenlea

5.1 Abstract

Wheat endosperm texture, which determines many of its end-use properties, is controlled primarily by a locus (*Ha*) present on chromosome 5 homoeologues. The *Ha* locus comprises paralogous *Gsp-1*, *Pina,* and *Pinb* genes encoding the so-called grain softness protein, puroindoline-a and puroindoline-b respectively. *Pina* and *Pinb* sequences were detected only on the D-genome of hexaploid wheat and its diploid progenitors while *Gsp-1* was on all three homoeologous chromosome 5 loci. Shotgun sequencing of three BAC clones from the hexaploid wheat cultivar Glenlea was performed and sequences of 172 kb, 168 kb and 70 kb were obtained for the homoeologous regions of 5A, 5B and 5D respectively. Annotation and analysis of the sequences revealed the presence of genes amidst the mosaic organization of fragments of retroelements and DNA transposons. The previously reported colinearity of 5' and 3' boundaries of the *Ha* loci across genomes, helped to delimit the region, which spanned 3,925 bp, 5,330 bp and 31,607 bp in the A-, B- and D-genomes respectively. Glenlea is null for *Pina* because the gene was almost entirely deleted with the exception of its promoter region and a short 5' coding region. The truncated *Pina* and the *Pinb* genes were followed by the conserved genes of the 3' boundary of the locus. A solo LTR of Angela retroelement, 1.9 kb downstream to *Gsp-A1* and a fragment of Sabrina retroelement, 2.8 kb downstream of *Gsp-B1* were discovered. We propose that the insertion of these elements into the intergenic regions of the locus

have driven the independent deletions of genomic segments harbouring *Pina* and *Pinb* genes in the A- and B-genomes of hexaploid wheat, by the mechanism(s) of unequal homologous/illegitimate recombination. Structural differences between Glenlea and Renan sequences from the *Ha* locus region of the A-genome suggested the advent of more than one tetraploid ancestor in the origin of hexaploid wheat. Based on the size of the region in the D-genome of *T. aestivum* cv Renan (66,103 bp, CR626934), deletions of ~62.5 kb and ~61 kb, respectively in the A- and B-genomes of cv Glenlea were estimated. Such retroelement activity was also likely responsible for the *Pina* deletion in Glenlea. Presence of fragments of Romani and Vagabond retroelements between *Pina* promoter and *Pinb* gene indicated their role in the deletion of ~29 kb fragment from the 32 kb interval observed in *T. monococcum* (AY491681). Similarly, compared to the corresponding interval in the D-genome of *Ae. tauschii* (58,298 bp, CR626926) and *T. aestivum* cv Renan (17,701bp, CR626934), segmental deletions of ~55 kb and ~15 kb, including the coding region of *Pina* can be inferred. In total, 14 genes with a density of one gene per 12 kb, 18 genes with a density of one gene per 9 kb and 10 genes with a density of one gene per 7 kb were found in the *Ha* loci of A-, B- and D-homoeologues, respectively. Comparative analysis of these clones with the orthologous regions from *T. monococcum* (A^mA^m), *Ae. tauschii* (DD), *T. turgidum* (AABB) and *T. aestivum* (AABBDD) are presented.

5.2 Introduction

Wheat endosperm texture is an important physical characteristic that determines end use properties of wheat. Three major classes of wheat are identified based on the endosperm texture, namely, soft, hard and very hard, each having its specific end uses (Morris, 2002). The *Hardness* locus (*Ha*) present on the short arm of chromosome 5 homoeologues harbors the major genes for this trait (Mattern et al. 1973; Law et al. 1978; Sourdille et al. 1996). The *Ha* locus comprises *Gsp-1*, *Pina* and *Pinb* genes encoding the grain softness protein (GSP-1), puroindoline-a (PINA) and puroindoline-b (PINB), respectively (Morris 2002). Both *Pina* and *Pinb* genes have a single exon of 447 bp, encoding a polypeptide of 148 aa while *Gsp-1* is 495 bp long and encodes a polypeptide of 164 aa (Chantret et al. 2004).

Endosperm texture of diploid, tetraploid and hexaploid wheats can be soft to very hard with varying degrees of hardness in between the two extremes. Soft wheats have an abundance of friabilins, a 15 kDa polypeptide complex on the surface of water-washed starch granules. Relatively low levels of friabilins are found in hard wheat and they are absent in very hard wheat such as *T. durum* (Greenwell and Schofield 1986, Morrison et al. 1992, Rahman et al. 1994). Puroindoline-a and puroindoline-b were characterised as the main components of the friabilin complex (Gautier et al. 1994, Hogg et al. 2004). Direct evidence that *Pina* and *Pinb* are the major genetic factors contributing to endosperm hardness was provided through functional complementation of *Pina* and *Pinb* (Beecher et al. 2002, Martin et al. 2006). Also, enhancement of the softness of transgenic rice grains by wheat *Pina* and *Pinb* sequences indicated their predominant role in grain texture (Krishnamurthy and Giroux 2001). The current model of endosperm organization associated with texture is that the tryptophan-rich domain of puroindolines interacts with phospholipids and bind to the surface of the starch granules preventing its direct interaction with matrix proteins composed of glutenins and gliadins (Gautier et al 1994, Ikeda et al. 2005, Bhave and Morris 2008a). Though, GSP-1 has 57-58 % homology with PINA and PINB, a direct role of grain softness protein on texture has not been proven so far (Bhave and Morris 2008a). There are additional genetic, biochemical and environmental factors which modify the degree of hardness in a quantitative fashion leading to the occurrence of continuous variation for grain hardness, especially in hexaploid wheat (Turnbull and Rahman 2002; Bhave and Morris 2008b).

Hard endosperm texture is the result of deletion, frame-shift mutation or substitution mutation in the coding region of one of the puroindoline genes (Morris 2002). There are multiple alleles characterized for each of the three genes of the *Ha* locus from the diverse gene pools of wheat from around the world, including its diploid progenitors (Bhave and Morris 2008b). In *Pinb*, a single nucleotide polymorphism (SNP) resulting in the substitution of glycine by serine at position 46, leads to a hard phenotype because it occurs in the lipid-binding domain (Giroux and Morris, 1997). This mutation, designated as *Pinb-D1b* allele was found to be the most prevalent *Pinb* mutation among North American and European wheat cultivars (Morris et al. 2001, Huang and Roder 2005). All other described mutations in *Pinb* are also single nucleotide substitutions except in some cultivars where frame shift mutations and null *Pinb* were identified (Ikeda et al. 2005). For *Pina*, the *Pina-D1a* (wild type) and *Pina-D1b* (null) alleles have been described in *Triticum* species (Bhave and Morris 2008b).

Wheat and rice diverged from a common ancestor 46 million years ago (MYA), oats 25 MYA, barley 11-13 MYA and rye 7 MYA (Huang et al. 2002). Puroindoline orthologues were found in oats (avenoindolines, Tanchak et al. 1998), barley (hordoindolines, Darlington et al. 2001) and rye (secaloindolines, Massa and Morris 2006). However, they were not found in rice, sorghum and maize suggesting that they arose from the common ancestor after rice but before oat speciation (Gautier et al. 2000). Puroindoline sequences were found in all diploid *Triticum* and *Aegilops* species and in the D-genome of hexaploid wheat but were absent in the tetraploid wheat species (Tranquilli et al. 1999; Gautier et al. 2000). However, *Gsp-1* was present in all three homoeologues of chromosome 5 (Jolly et al. 1996; Sourdille et al. 1996). Diploid wheats radiated <4.5 MYA followed by the formation of tetraploid wheats 0.5-3 MYA, with hexaploid wheat originating from hybridization between tetraploid wheat (AABB) and *Ae. tauschii* (DD), approximately 6000-8000 years ago (Huang et al. 2002). Puroindolines were deleted in tetraploid wheats and were restored in hexaploid wheats by the addition of the diploid *Ae. tauschii* (DD). Extensive analysis of haplotype structure of the *Ha* loci among 300 polyploid and 90 diploid accessions of wheat, including *Aegilops,* revealed the conservation of the puroindoline genes in diploid ancestors and their independent deletions in polyploid species (Li et al. 2008). The exceptions are *T. timopheevi* (AAGG) where puroindolines are present on the A genome and the *T. timopheevi* derived hexploid *T. zhukovskyi* (A^mA^mAAGG), where they are deleted from the A genome but retained in the A^m genome (Li et al. 2008).

Sequencing of the *T. monococcum Ha* locus revealed that *Gsp-*A^m*1*, *Pina-*A^m*1* and *Pinb-*A^m*1* were located 37 and 32 kb apart, respectively (Chantret et al. 2004). However, sequencing of the *Ae. tauschii Ha* locus revealed that *Gsp-D1* and *Pina-D1* were only 17.9 kb apart while *Pina-D1 and Pinb-D1* were separated by 58.3 kb (Chantret et al. 2005). The intergenic intervals were found to have retrotransposons which played a crucial role in the evolution of genome structural organization (Bennetzen 2005).

Triticum aestivum cv Glenlea is wild type for *Pinb* (*Pinb-D1a*), but has a *Pina* null genotype (*Pina-D1b*) and therefore displays a hard endosperm texture (Dubreil et al. 1998, Morris et al. 1998). Here, we report on the sequencing and comparative analyses of the homoeologous *Ha* loci from the A-, B- and D-genomes, unravelling the genetic mechanisms that have driven the evolution of the *Ha* loci, specifically the absence of *Pina* and *Pinb* genes in the A- and B-genomes and the *Pina* gene in the D-genome of hexaploid wheat in general and of cv Glenlea in particular.

5.3 Materials and Methods

5.3.1 BAC selection and physical mapping

The multi-dimensional pools generated from the entire BAC library of *Triticum aestivum* cv Glenlea (Nilmalgoda et al. 2003) were screened by PCR using genome specific *Gsp-1* primers and puroindoline-b (*Pinb-D1*) primers. The A-genome specific primers were: *Gsp*F371G, 5'-GATATGCCGCTCTCTTGGG-3' and *Gsp*R600A, 5'-GGATCAATGTTGCACTTGGA-3'. The B-genome specific primers were: *Gsp*F370A, 5'-GGATATGCCGCTCTCTTGGA-3' and *Gsp*R600A, as above. The D-genome specific primers were: *Gsp*F370T, 5'-GGATATGCCGCTCTCTTGGT-3' and *Gsp*R600T, 5'-GGATCAATGTTGCACTGGGT-3'. The *Pinb* primers were: 5'-ATGAAGACCTTATTCCTCCTA-3' and 5'-TCACCAGTAATAGCCACTAGGGAA-

3'. Two clones for *Pinb,* two clones for *Gsp-1A* and seven clones for *Gsp-1B* were identified. No clone with *Gsp-1D* was recovered from the BAC library. The clones were fingerprinted and assembled using the high information content method (Luo et al. 2003). Software packages GenoProfiler (http://wheat.pw.usda.gov/Physical Mapping/ tools/ genoprofiler/genoprofiler.html) and FPC (fingerprinted contigs; Soderlund et al. 1997) were employed to build contig assemblies.

5.3.2 Sequencing, assembly and gap closing

Three BAC clones representing the *Ha* loci of the homoeologous A-, B- and D-genomes were selected (TaBAC502E9, TaBAC1551N13 and TaBAC1067B03, respectively). Construction of shotgun libraries and sequencing were carried out by Genome Express (38044 Grenoble Cedex 9, France) and DNA Landmarks (St-Jean-sur-Richelieu, Canada) as follows. The BAC clones were sheared with a Hydroshear and shotgun libraries ranging from 2.5 kb to 3.5 kb and 3.5 to 5 kb were constructed and sequenced using Big-Dye V3.1 Terminator chemistry and resolved on an ABI 3730 and 3730XL DNA Analyzer (Applied Biosystems, CA, USA). Base calling was carried out using PHRED (Ewing et al. 1998) with a minimum quality score of 40 and contig assembly was carried out using PHRAP (www.phrap.org) with a genome coverage of >10X. The initial assembly was manually curated using CONSED (Gordon et al. 1998). The orientation and order of the contigs/scaffolds in the assembly were confirmed by restriction mapping of the BAC clones and supported by mate-pair reads flanking gaps. Gaps were closed by primer walking from the end regions of the contigs or scaffolds using subclones spanning the gap and BAC clones as templates. The assembly was checked for its consistency by PCR analyses.

5.3.3 Sequence analyses

Structural and functional annotation of the genes and transposable elements were carried out by homology as well as structural feature based methods, using the bioinformatics tools and pipelines available in the public domain as described below.

5.3.4 Annotation of transposable elements

LTR-Finder was used to identify the full length retrotransposons with or without target site duplications using the default parameters (Xu and Wang 2007). The potential locations of truncated and fragmented DNA transposons and retroelements were identified using TENest, the *Triticeae* repetitive sequence database (TREP) and the default parameters (Kronmiller and Wise 2008). Finally, the coordinates of predicted elements were extracted and BLASTn and BLASTx (Altschul et al. 1990) and were used for searches against TREP (Wicker et al. 2002) and RepBase (Jurka 2000) to identify LTR and internal domains, as per the guidelines of Wicker et al. (2007b). In some cases, exact positions of predicted elements were identified using BLAST2 (Tatusova and Madden 1999). In the end, various software results were manually curated.

5.3.5 Annotation of genes

RiceGAAS was used to locate and identify putative genes (Sakata et al. 2002). The coordinates of genes predicted by FGENESH (www.softberry.com), GENSCAN (http://genes.mit.edu/GENSCAN.html) and RiceHMM (http://rgp.dna.affrc.go.jp/RiceHMM) were identified. Genes that were part of mobile elements were eliminated. BLASTx searches of remaining putative genes against the non-redundant protein sequence database nr (NCBI) were carried out to annotate genes and pseudogenes encoding known proteins. Homology at an E-value of $< e^{-10}$ was considered as the threshold to identify real matches. Homology searches using BLASTn against dbEST was also carried out to validate the BLASTx results. Candidate genes predicted only by the *ab initio* programs and/or homologous to ESTs without any protein matches were designated as hypothetical genes.

5.3.6 Comparative analyses

Finally, comparative analyses of genome organization of the *Ha* loci from three BAC clones were done by comparing them to one another and their orthologues from *T. monococcum* (A^mA^m), *Ae. tauschii* (DD), *T. turgidum* (AABB) and *T. aestivum* (AABBDD) using the software DOTTER (Sonnhammer and Durbin 1996), BLAST2 (Tatusova and Madden 1999 (http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi,), DNAMAN (Version 3.2, Lynnon Biosoft, USA) and JDOTTER (Brodie et al. 2004, http://athena.bioc.uvic.ca/workbench.php?tool=jdotter&db=).

5.4 Results

5.4.1 Sequence assembly of the clones harbouring the *Ha* loci

For TaBAC502E9 from the A-genome of hexaploid wheat cultivar Glenlea, a 172,643 bp contiguous sequence was obtained with ~12 X genome coverage. Two gaps of approximately 98 bp and 240 bp starting at the coordinates 98,747 and 160,238, respectively could not be resolved. For TaBAC1551N13 from the B-genome, a 168,467 bp contiguous sequence was obtained with a genome coverage of ~10X. The assembly included a single unresolved gap estimated at 273 bp starting at the position 128,030. For TaBAC1067B3 representing the D-genome, a 69,916 bp contiguous sequence was obtained with a genome coverage of ~15X. Two gaps estimated at 716 bp and 300 bp and starting at the coordinates 19,047 and 53,209 respectively remained. The GC content of all three BAC clones was similar at 45, 46 and 47% respectively.The three BAC clone sequences were submitted to GenBank under accession number EU835980, EU835981 and EU835982.

5.4.2 Annotation of transposable elements

The structural organization of the mobile genetic elements and genes identified in BAC clones TaBAC502E9, TaBAC1551N13 and TaBAC1067B3 representing the A-, B- and D-genomes, respectively, are depicted in Fig. 5.1. Approximately 47% of the TaBAC502E9 sequence was characterized as mobile genetic elements. Similarly, transposable elements comprise 44% and 31% of the sequences of TaBAC1551N13 and TaBAC1067B03, respectively.

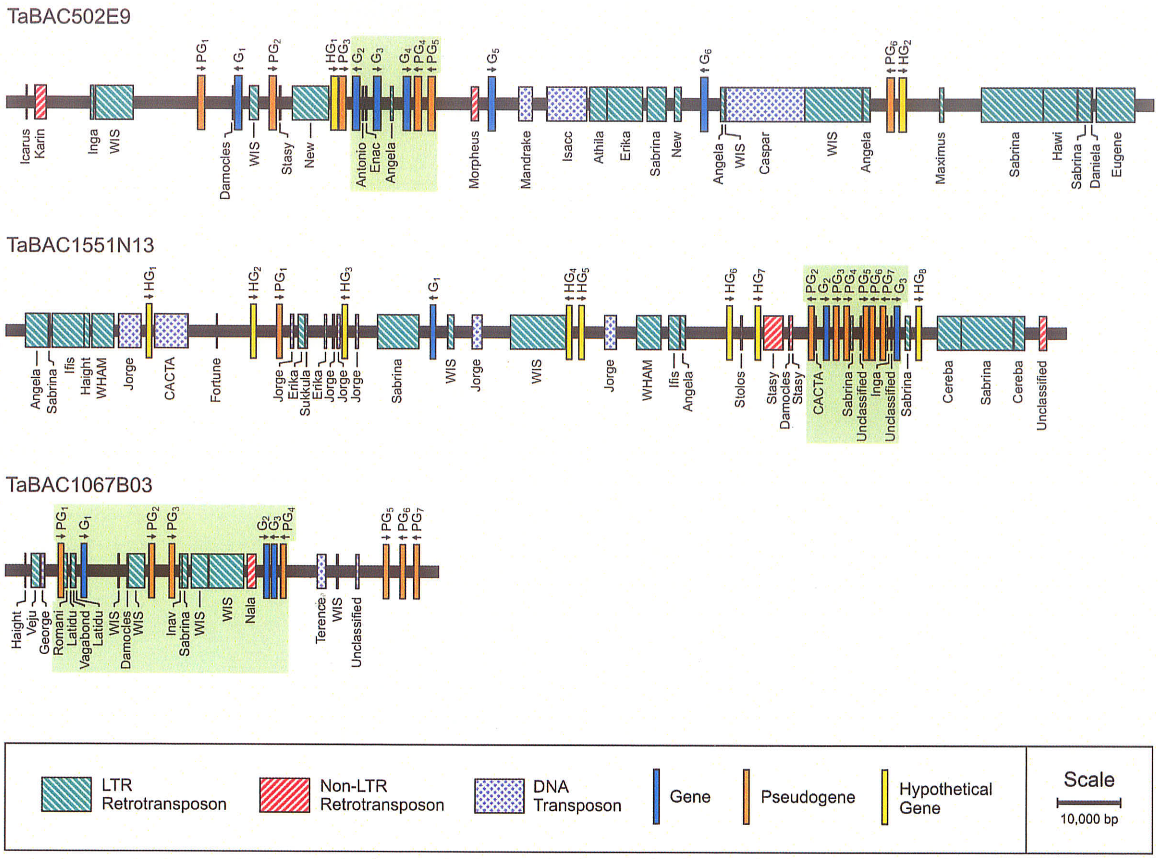


Figure 5.1 Schematic representation of the annotation of BAC clone TaBAC502E9 (A-genome, ~172 kb), TaBAC1551N13 (B-genome, ~168 kb), TaBAC1067B3 (D-genome, ~70 kb) from *T. aestivum* cv Glenlea. The *Ha* loci regions are indicated by shaded boxes. Arrows indicate the transcriptional orientation of the genes. Genes (G), pseudogenes (PG) and hypothetical genes (HG) are numbered as per their order and as described in Tables 5.1.

In TaBAC502E9, 19 insertions of LTR retrotransposons belonging to the families Inga, WIS, Angela, Athila, Erika, Sabrina, Maximus, Hawi, Daniela and Eugene were identified. Among them, four WIS elements and three insertions each of Angela and Sabrina were found. Most of them were either solo LTRs, truncated or degenerated. One complete retroelement (spanning from 39,484 to 44,927-bp; size-5444 bp) with 128 bp LTR pair, 5180 bp internal domain and 5 bp target site duplication (TSD) on either side, belonging to the *copia* superfamily had not been previously described and was named Raja_TaBAC502E9-1. Three non-LTR retroelements namely, Karin, Stasy and

Morpheus were also identified. As well, seven DNA transposons were also found: Icarus, Damocles, Antonio, Enac, Mandrake, Isaac and Caspar.

In TaBAC1551N13, 16 insertions of LTR retroelements, four insertions of non-LTR retroelements and 13 insertions of DNA transposons were identified. The LTR retroelements belong to the Angela, Sabrina, Ifis, Haight, WHAM, Erika, WIS, Inga and Cereba families. Among them, four Sabrina elements and two Angela, WIS and WHAM elements were found. Four non-LTR retroelements belonging to Stasy and unclassified families were also identified. Among DNA transposons, seven elements belonging to the Jorge family were found. The other families were CACTA, Fortune, Stolos and Damocles. Three elements were unclassified.

In TaBAC1067B3, 13 insertions of LTR retroelements belonging to the Haight, Veju, Romani, Latidu, Vagabond, WIS, Inav and Sabrina families were identified. A non-LTR retroelement (Nala) and four DNA transposons (George, Damocles, Terence and an unclassified element) were also found.

5.4.3 Annotation of genes

In TaBAC502E9, a total of 14 genes including six pseudogenes and two hypothetical genes were identified (Table 5.1, Fig. 5.1). The gene density was estimated at one gene per 12 kb. A total of 18 genes including seven pseudogenes and eight hypothetical genes with a gene density of one gene per 9 kb were identified in TaBAC1551N13 (Table 5.1, Fig. 5.1). In TaBAC1067B3, a total of 10 genes including seven pseudogenes were found (Table 5.1 and Fig. 5.1). The gene density of this region was estimated to be one gene per 7 kb.

Table 5.1 Annotation of coding sequences of TaBAC502E9, TaBAC1551N13 and TaBAC1067B3 of the *Ha* locus homoeologues of the A-, B- and D-genome of hexaploid wheat cv Glenlea

Annotation designation	**Start position**	**End position**	**Strand**	**Length (bp)**	**Exon**	**Predicted protein size (#aa)**[a]	**BLAST result**[b]
TaBAC502E9_PG1	26,269	26,401	+	133	single	-	99% identity with the Pseudo_kinase from *T. aestivum* BAC clone CT009735 (E=1e^{-47})[c]
TaBAC502E9_G1	30,969	32,318	-	1,269	2	422	97% identity with Chalcone synthase from *T. turgidum,* emb-CAJ13548.1 (E=0), EST-97% identity over 564 bp of gb-CA600207.1 (E=0)
TaBAC502E9_PG2	36,426	37,073	+	648	single	215	98% identity with an unnamed protein from *T. aestivum*, emb-CAJ13537.1(E=9e-73)
TaBAC502E9_HG1	45,263	46,844	+	1,083	2	360	Predicted by FGENESH and RiceHMM, EST- 98% identity over 94 bp of gb-CA666607.1 (E=5e^{-04})
TaBAC502E9_PG3	46,975	47,349	+	375	single	124	100% identity with the Vesicle associated membrane protein from *T. turgidum,* emb-CAJ13552.1 (E=2e^{-29})
TaBAC502E9_G2	48,289	51,479	-	1,278	4	425	100% identity with the Beta 1-3-galactosyl-o-glycosyl glycoprotein (BGGP) from *T. turgidum,* emb-CAH10066.1 (E=0); EST-96% identity over 599 bp of gb-BU100707.1 (E=0)
TaBAC502E9_G3	53,528	54,022	+	495	single	164	100% identity with the *GSP-A1* from *T. aestivum,* gb-AA09276 (E=1e^{-88}); EST-92% identity over 462 bp of gb-CJ528780.1 (E=0)

Annotation designation	Start position	End position	Strand	Length (bp)	Exon	Predicted protein size (#aa)[a]	BLAST result[b]
TaBAC502E9_G4	57,452	59,755	+	1,605	2	534	99% identity with an unknown protein similar to *Arabidopsis* hypothetical protein yhjx (F25A4.25), emb-CAH10068.1 (E=0), EST-92% identity over 584 bp of gb-CD919995.1 (E=0)
TaBAC502E9_PG4	60,350	61,691	-	975	3	325	77% identity with the Cell division protein AAA ATPase family from *T. aestivum,* gb-CAH10057 (E=5e^{-126})
TaBAC502E9_PG5	62,694	64,063	-	1,314	2	438	93% identity with the Cell division protein AAA ATPase family from *T. aestivum*, emb-CAH10048.1 (E=0)
TaBAC502E9_G5	71,947	73,518	+	1,572	single	523	100% identity with the Cell division protein AAA ATPase family from *T. turgidum,* emb-CAH10071.1 (E=0), EST-95% identity over 621 bp of gb-CD875876.1 (E=0)
TaBAC502E9_G6	104,307	105,872	-	1,566	single	521	100% identity with the Cell division protein AAA ATPase family from *T. turgidum,* emb-CAJ13559.1 (E=0), EST-98% identity over 625 bp of gb-CD875876.1 (E=0)
TaBAC502E9_PG6	132,730	136,527	-	2,040	7	679	92% identity with Glucose/Sorboson dehydrogenase from *H. vulgare,* gb-AAV49993 (E=0)
TaBAC502E9_HG2	137,269	138,270	+	1,002	single	333	Predicted by FGENESH, RiceHMM, GENSCAN-Maize and GENSCAN-Arabidopsis, EST-85% identity over 224 bp of gb-BQ240962.1 (E=2e^{-043})

Annotation designation	**Start position**	**End position**	**Strand**	**Length (bp)**	**Exon**	**Predicted protein size (#aa)**[a]	**BLAST result**[b]
TaBAC1551N13_HG1	19,165	20,512	+	552	2	183	Predicted by FGENESH, GENSCAN-Maize and RiceHMM, EST-94% identity over 312 bp of gb-CA681862.1 (E=e^{-129})
TaBAC1551N13_HG2	36,195	36,785	+	396	2	131	Predicted by FGENESH, GENSCAN-Maize, GENSCAN-Arabidpsis and RiceHMM
TaBAC1551N13_PG1	39,999	41,750	+	654	5	217	27% identity with the putative uncharacterized protein-Os08g0514800 from *Oryza sativa* (E=2e^{-23})
TaBAC1551N13_HG3	51,273	53,963	-	1,176	7	391	Predicted by FGENESH, GENSCAN-Arabidpsis and RiceHMM, EST-93% identity over 576 bp of gb-CJ522422.1 (E=0)
TaBAC1551N13_G1	67,118	67,702	-	585	single	194	100% identity with an unnamed protein from *T. aestivum,gb*-CAJ13527 (E=5e^{-74}), EST-78% identity over 215 bp of gb-AL815828.1 (E=2E^{-008})
TaBAC1551N13_HG4	88,431	89,440	-	771	3	256	Predicted by FGENESH, GENSCAN-Arabidpsis and RiceHMM, EST-91% identity over 123 bp of gb-CK161792.1 (E=2e^{-039})
TaBAC1551N13_HG5	90,585	91,365	+	471	2	156	Predicted by FGENESH, GENSCAN-Arabidpsis and RiceHMM, EST-91% identity over 268 bp of gb-CJ633543.1(E=2e-^{097})
TaBAC1551N13_HG6	113,969	114,493	+	525	single	174	Predicted by FGENESH and GENSCAN-Maize

Annotation designation	Start position	End position	Strand	Length (bp)	Exon	Predicted protein size (#aa)[a]	BLAST result[b]
TaBAC1551N13_HG7	118,102	118,634	+	354	2	117	Predicted by FGENESH, GENSCAN-Maize, GENSCAN-Arabidpsis and RiceHMM, EST-100% identity over 351 bp of local EST assembly-Contig 3434 (E=0)
TaBAC1551N13_PG2	126,247	128,456	-	678	3	225	99% identity with the Beta 1-3-galactosyl-o-glycosyl glycoprotein (BGGP) from *T. aestivum,* emb-CAH10050.1 (E=5e^{-125})
TaBAC1551N13_G2	130,003	130,497	+	495	single	164	100% identity with the *GSP-B1* from *T. aestivum,* gb-CAA56596 (E=2e^{-89}), EST-97% identity over 463 bp of gb-CJ528780.1 (E=0)
TaBAC1551N13_PG3	131,203	131,835	+	341	2	114	97% identity with the Cell division protein AAA ATPase family from *T. aestivum,* emb-CAH10203.1 (E=1e^{-58})
TaBAC1551N13_PG4	132,644	133,099	+	456	single	151	82% identity with the *H. vulgare* ATPase2 (Cell division protein AAA ATPase family), gb-AAV499983.1 (E=1e^{-68})
TaBAC1551N13_PG5	135,332	137,709	+	1,479	3	492	100% identity with an unknown protein product similar to Arabidopsis hypothetical protein yhjx (F25A4.25), emb-CAH10054.1 (E=0)
TaBAC1551N13_PG6	137,876	139,627	-	1,359	3	452	90% identity with the *H. vulgare* ATPase3 (Cell division protein AAA ATPase family), gb-AAV49988.1 (E=2e^{-111})

Annotation designation	Start position	End position	Strand	Length (bp)	Exon	Predicted protein size (#aa)[a]	BLAST result[b]
TaBAC1551N13_PG7	140,493	142,439	-	1,125	4	374	85% identity with the Cell division protein AAA ATPase family from *T. aestivum,* emb-CAH10048.1 (E=$1e^{-128}$)
TaBAC1551N13_G3	143,791	145,353	+	1,563	single	520	100% identity with the Cell division protein AAA ATPase family from *T. aestivum,* emb-CAH10057.1 (E=0); EST-98% identity over 716 bp of gb-CK204059.1 (E=0)
TaBAC1551N13_HG8	147,870	148,139	+	270	single	89	Predicted by FGENESH, GENSCAN-Maize, GENSCAN-Arabidpsis and RiceHMM
TaBAC1067B3_PG1	5,824	6,723	+	21	single	7	Puroindoline a protein (PinA) from *T. aestivum,* gb-AJ302091
TaBAC1067B3_G1	9,467	9,913	+	447	single	148	100% identity with the puroindoline b protein (PinB) from *T. aestivum*, gb-AAT40245.1 (E=$9e^{-68}$), EST-98% identity over 448 bp of gb-CD887650.1 (E=0)
TaBAC1067B3_PG2	19,728	20,417	+	690	single	229	93% identity with the Cell division protein AAA ATPase family from *T. aestivum*, emb-CAH10048.1 (E=$4e^{-118}$)
TaBAC1067B3_PG3	22,740	24,286	+	1,488	2	495	76% identity with the *H. vulgare* ATPase2 (Cell division protein AAA ATPase family), gb-AAV49983.1 (E=0)
TaBAC1067B3_G2	38,307	40,645	+	1,602	2	533	99% identity with an unknown protein similar to Arabidopsis hypothetical protein yhjx (F25A4.25), emb-CAH10204.1 (E=0),EST-92% identity over 584bp of gb-CD919995.1 (E=0)

Annotation designation	**Start position**	**End position**	**Strand**	**Length (bp)**	**Exon**	**Predicted protein size (#aa)**[a]	**BLAST result**[b]
TaBAC1067B3_G3	40,945	42,435	-	1,491	single	496	100% identity with the Cell division protein AAA ATPase family from *T. aestivum,* emb- CAH10203.1 (E=0), EST-99% identity over 234 bp of gb-CA691639.1 ($E=e^{-127}$)
TaBAC1067B3_PG4	43,028	45,196	-	1,458	4	486	92% identity with the Cell division protein AAA ATPase family from *T. aestivum,* emb- CAH10201.1 (E=0)
TaBAC1067B3_PG5	60,308	61,756	+	1,212	3	403	29% identity with the Conserved hypothetical protein from *Oryza sativa,*OsJ_024475, gb-EAZ40992.1
TaBAC1067B3_PG6	63,566	64,892	-	1,211	4	403	27% identity with the Conserved hypothetical protein from *Oryza, sativa,* OsJ_031453, gb-EAZ17244.1
TaBAC1067B3_PG7	66,191	66,714	-	397	2	132	26% identity with the Conserved hypothetical protein from *Oryza, sativa,*OsJ_019969, gb-EAZ36486.1

[a] #aa: number of amino acid residues

[b] Best BLASTx match against the NCBI non-redundant (nr) database and/or best BLASTn match against dbEST

[c] Best BLASTn match against the NCBI nucleotide (nt) database because no hit with BLASTx-nr and BLASTn-dbEST

5.4.4 Structural organization of the homoeologous *Ha* loci in the A-, B- and D-genomes of cv Glenlea

Through comparative genomic sequence analysis, the boundaries of the homoeologous *Ha* loci had previously been defined as gene *BGGP* (encoding β-1, 3-galactosyl-o-glycosyl-glycoprotein) at the 5'end and hypothetical gene, similar to an *Arabidopsis thaliana* gene, named *gene 8* at the 3'end (Chantret et al. 2005). While the rice genome is devoid of puroindoline gene, sequence similar to *Gsp-1* and orthologous to the wheat *Ha* locus sequence was found in a region of chromosome 12 (Chantret et al. 2004). The defined *Ha* loci spanned 3,925 bp, 5,330 bp and 31,607 bp in TaBAC502E9, TaBAC1551N139 and TaBAC1067B3, respectively (Fig. 5.2). Because our sequences are longer at the 3'-end, we can observe that the colinearity between the 3' homoeologous loci extends beyond the *A. thaliana* hypothetical gene to include tandemly repeated *ATPase* genes intercalated by various mosaics of repetitive elements.

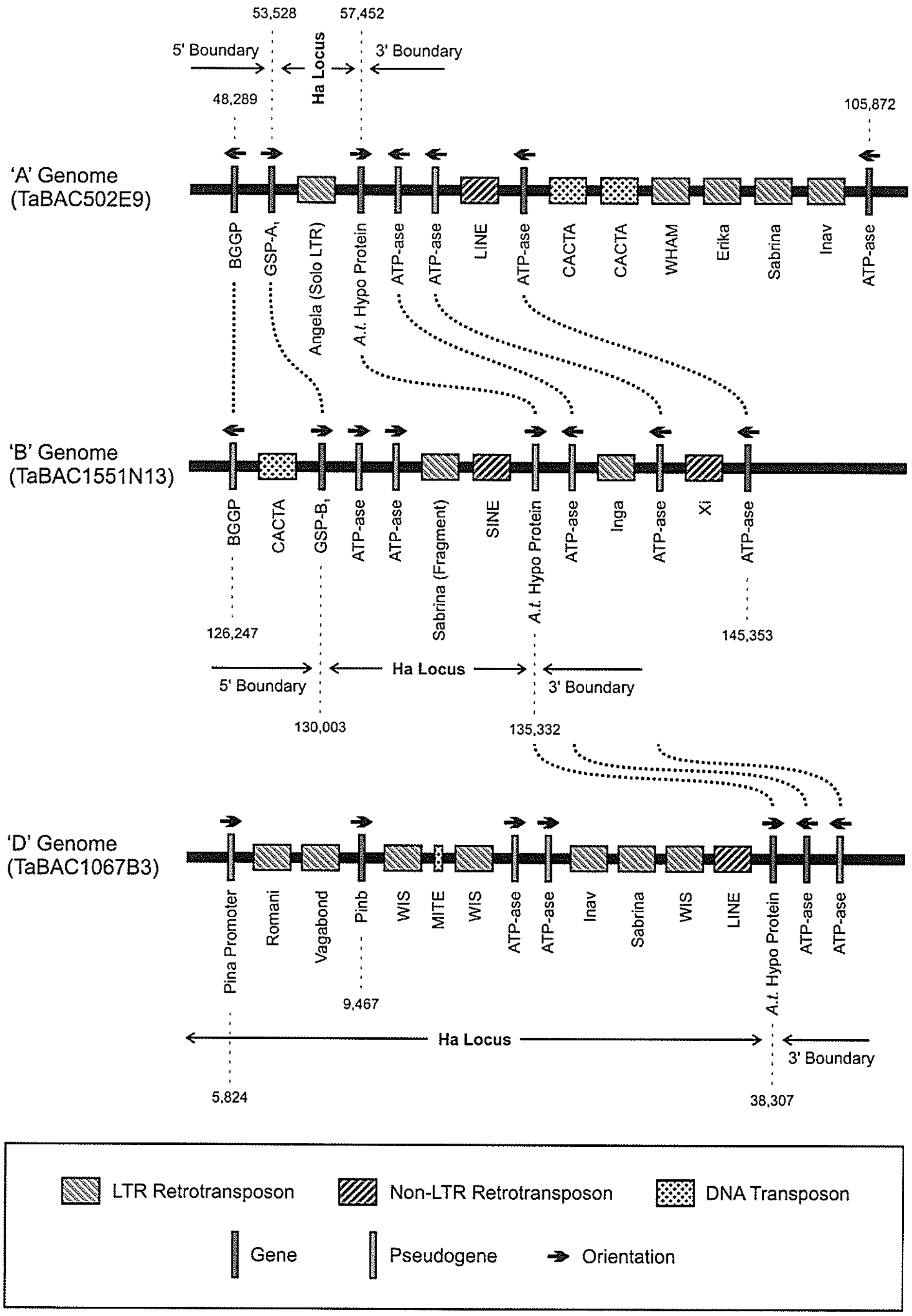


Figure 5.2 Genome organization of the *Ha* loci from the A-, B- and D-genomes of *T. aestivum* cv Glenlea, demarcated by its 5' and 3' boundaries (see text for details). The conserved genes at the possible orthologous positions are connected with dotted lines. Positions of the important genes of the loci in whole BAC sequences are indicated. Transcriptional orientations of the genes are indicated by arrows. For clarity purposes elements are not drawn to scale.

In TaBAC502E9 sequence, the *BGGP* was found from 48,289 to 51,479-bp, upstream of *Gsp-A1* (495 bp) encoding the grain softness protein. Approximately 1.5 kb downstream of the *Gsp-A1*, a 441 bp solo LTR of Angela belonging to the *copia* superfamily was found. Further downstream, the 3' boundary specific hypothetical gene followed by two copies of *ATPase* was present (Fig. 5.2). Two more copies of *ATPase* were also present in the immediate vicinity of the 3' boundary, separated by two DNA transposons and six retroelement insertions.

In TaBAC1551N13, the *BGGP,* found from 126,247 to 128,456-bp was followed by a CACTA element. The *Gsp-B1* (495 bp) was followed by two copies of *ATPase,* a 627 bp fragment of Sabrina retroelement belonging to the *athila* superfamily and a SINE insertion (Fig. 5.2). The hypothetical gene delimiting the 3' boundary was upstream of three copies of the *ATPase* gene.

In TaBAC1067B3, a 879 bp promoter region along with a 21 bp truncated coding region of the *Pina-D1* encoding puroindolineA (Glenlea is null for *Pina*) was located from 5824 to 6723-bp (Fig. 5.2). The *Pinb-D1* (447 bp) coding for puroindoline-b (148 aa) was located at the coordinates 9467 to 9913-bp and followed by the conserved genes of the 3' boundary of the locus. Fragments of Romani and Vagabond retroelements were identified between the truncated *Pina* (except the initial 21 bp coding region) and *Pinb* genes. In the interval between the *Pinb* and the hypothetical gene marking the 3' boundary (9,913-38,307-bp), two more copies of the *ATPase* gene were found as well as transposable elements WIS, Inav, Sabrina, LINE and MITE.

5.4.5 Comparison of homoeologous *Ha* locus regions from the A-, B- and D-genomes of *T. aestivum* cv Glenlea:

The *Ha* locus region in TaBAC502E9 (A-genome) spans only 3925 bp compared to 5330 bp representing the homoeologous region from TaBAC1551N13 (B-genome). Though the genes representing the 5' and 3' boundaries and *Gsp-1* are conserved (Fig. 5.2 and Fig. 5.3A), there is only 49% sequence identity between the regions. This is mainly due to the divergence of the intergenic regions. In TaBAC502E9, a solo LTR Angela element was found whereas in TaBAC1551N13 an internal fragment of Sabrina retrolement and two copies of *ATPase* were identified.

Similarly, there is only 40% sequence similarity observed between the homoeologous *Ha* regions in TaBAC502E9 and TaBAC1067B3 (31,607 bp; D-genome), though the genes representing the 3' boundary are conserved (Fig. 5.2 and Fig. 5.3B). In TaBAC502E9, a solo LTR Angela element was found whereas in TaBAC1067B03 the corresponding region consists of promoters of *Pina, Pinb,* two copies of *ATPase* and seven insertions of LTR retroelements belonging to Romani, Vagabond, WIS and Sabrina families. A non-LTR retroelement (Nala) and a MITE (Damocles) were also present. The immediate upstream region of the gene marking the 3' boundary harbored a 364 bp conserved sequence free of genes and repeats.

Only segmental conservation could be observed between the *Ha* loci of the B- and D-genomes of Glenlea (Fig. 5.2 and Fig. 5.3C). Two copies of *ATPase* (pseudogenes) were conserved in the B- (1378-1739-bp and 2641-3078-bp) and D-genomes (13018-13455-bp and 17098-17459-bp). Also, the B-genome has a fragment of Sabrina element

(3361-3937-bp) whereas the D-genome has a solo LTR of the same element, downstream of the *ATPase* (18717-19622-bp).

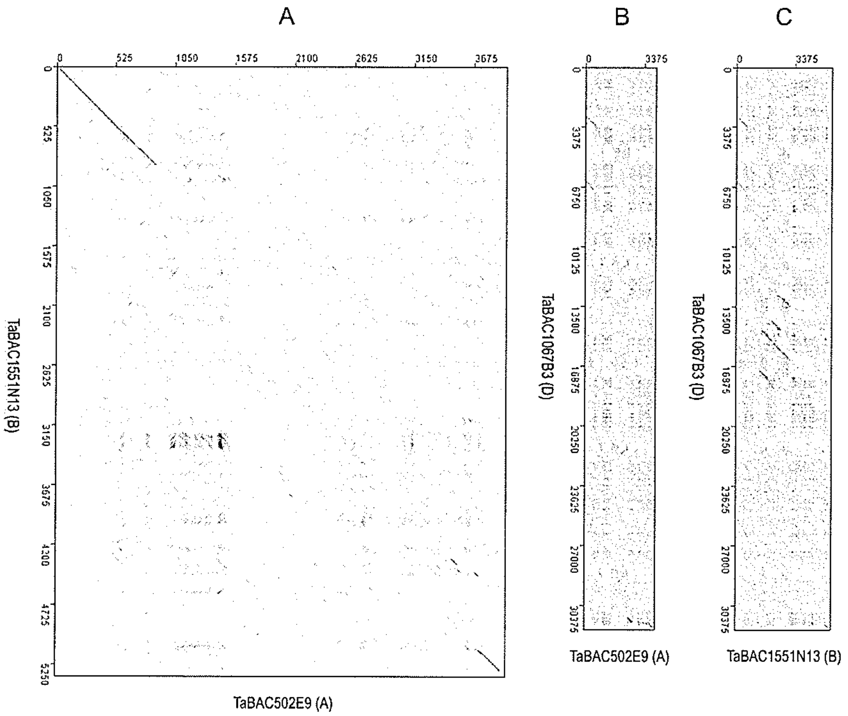


Figure 5.3 Pairwise comparison of the homoeologous *Ha* loci from *T. aestivum* cv Glenlea among themselves by dot plot analyses. **A.** A-genome (x-axis) versus B-genome (y-axis); **B.** A-genome (x-axis) versus D-genome (y-axis) and **C.** B-genome (x-axis) versus D-genome (y-axis).

5.4.6 Comparative analyses of the *Ha* locus regions

The *Ha* locus region from the A-, B- and D- genomes of Glenlea were compared to orthologous regions from *T. monococcum, Ae. tauschii, T. turgidum, T. aestivum* and homologous regions from *T. turgidum* and *T. aestivum* available in the public domain. The demarcated regions of the *Ha* loci considered for the comparison are outlined in Table 5.2.

5.4.7 A-genome of cv Glenlea vs. *T. monococcum, T. turgidum* and *T. aestivum* cv Renan

Comparison of the *Ha* locus region from the A-genome of Glenlea (spanning from 53,528 to 57452-bp; size-3925 bp) with the A^m-genome of *T. monococcum (*spanning from 25,946 to 101,101-bp; size-75,156 bp, Chantret et al. 2004) indicated the presence of sequence conservation (97% identity) of the 5' region with conservation breakpoints at 1970-bp and 1968-bp positions in the A- and A^m-genomes, respectively (Fig. 5.4 and Fig. 5.5A). This corresponds to the *Gsp*-1 gene (495 bp) present in the same orientation and 183 bp 3'UTR region. A tandem repeat of 185 bp (with repeat unit of 20 bp) is present 522 bp downstream of the 3'UTR region. The remaining conserved sequence consisted of unassigned DNA. Though a solo LTR of Angela element is present ~1.9 kb downstream to the *Gsp-1* gene in the A-genome, it is absent at the orthologous position in the A^m-genome. However, the *Ha* locus of the A^m-genome contained four insertions of Angela element, including a complete element in the genomic segment harbouring the *Pina* and *Pinb* genes which is deleted in the A genome.

Table 5.2 Demarcation of the *Ha* loci of *Triticum* and *Aegilops sp.* used for comparative analyses

BAC clone name/ Accession number	**Species/ Variety**	**Genome**	**Length of the BAC sequence (bp)**	**Start position of the *Ha* locus**[a]	**End position of the *Ha* locus**[b]	**Size (bp)**	**Reference**
TaBAC502E9	*T. aestivum* cv. Glenlea	A	172, 643	53,528	57,452	3,925	This study
TaBAC1551N13	*T. aestivum* cv. Glenlea	B	168,467	130,003	135,332	5,330	This study
TaBAC1067B3	*T. aestivum* cv. Glenlea	D	69,878	6,701[c]	38,307	31,607	This study
AY 491681	*T. monococcum*	A^m	101,101	25,946	101,101[d]	75,156	Chantret et al. 2004
CR 626926	*A. tauschi*	D	94,421	4,616	94,421[e]	89,806	Chantret et al. 2005
CR626933	*T. turgidum ssp durum* cv Langdon65	A	25,216	5,240	9,163	3,924	Chantret et al. 2005
CR626932	*T. turgidum ssp durum* cv Langdon65	B	19,229	3,946	9,269	5,324	Chantret et al. 2005
CR626929	*T. aestivum* cv Renan	A	20,745	5,247	11,877	6,631	Chantret et al. 2005
CR626930	*T. aestivum* cv Renan	B	19,274	3,946	9,267	5,322	Chantret et al. 2005
CR626934	*T. aestivum* cv Renan	D	94,398	4,617	71,215	66,599	Chantret et al. 2005

[a] Start position of the *GSP-1* gene

[b] Start position of the gene encoding the *Arabidopsis thaliana* hypothetical protein yhjx (referred as *gene 8* in the literature) marking the 3' boundary of the *Ha* locus

[c] Start position of the truncated *Pina* gene because of the lack of *GSP-D1* at the 5'end

[d] & [e] End positions of the BAC sequences in the absence of the gene marking the 3' boundary of the *Ha* locus.

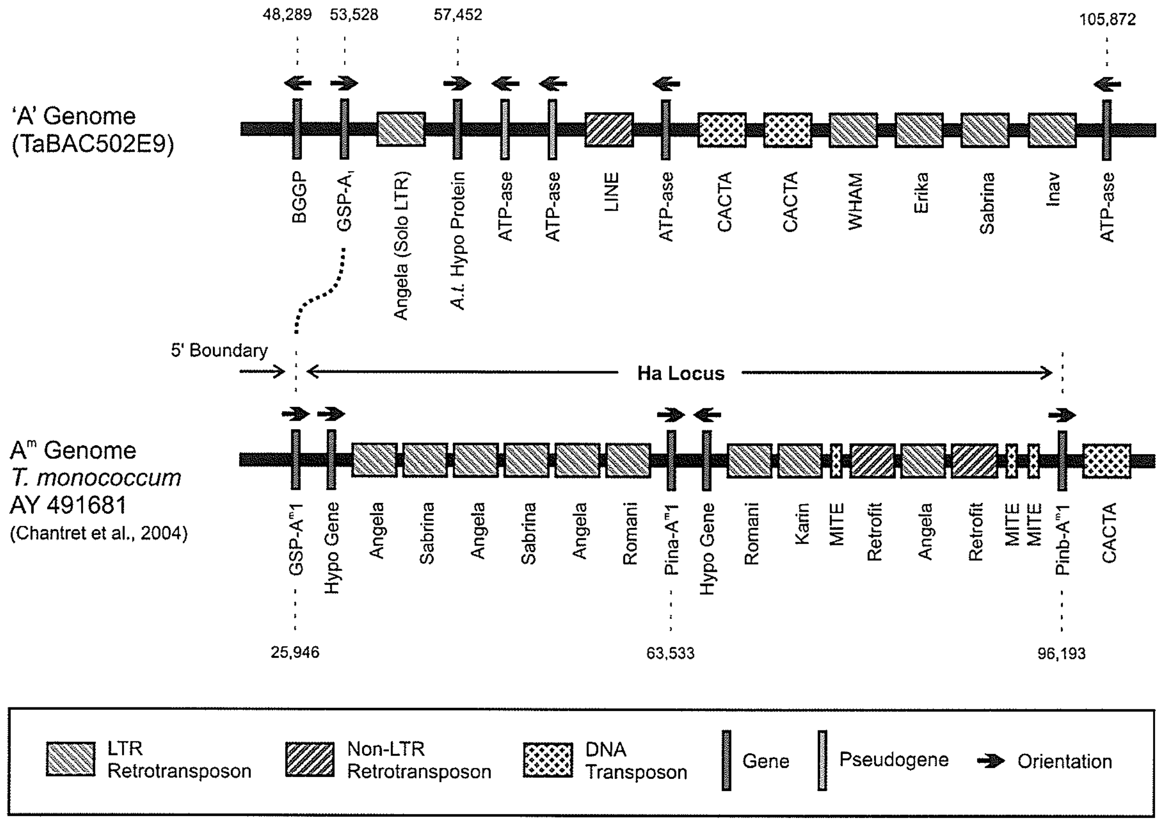


Figure 5.4 Genome organization of the *Ha* loci from the A-genome of *T. aestivum* cv Glenlea, anchored to the orthologous region from the A^{m}-genome of *T. monococcum* (AY491681). Positions of the important genes of the loci in whole BAC sequences are indicated. Transcriptional orientations of the genes are indicated by arrows. Elements are not drawn to scale

Approximately 99% sequence identity was observed between the *Ha* locus region from the A-genome of Glenlea (3925 bp) and the orthologous region from the A-genome of *T. turgidum* (3924 bp, Chantret et al. 2005) over their entire length, except for a single nucleotide (A) addition at the 1453-bp position and a SNP (G→ A) at the 1642-bp position in the *Ha* locus region of cv Glenlea (Fig. 5.5B).

Similarly, comparison of the *Ha* locus from the cv Glenlea A-genome with the corresponding region of *T. aestivum* cv Renan (Chantret et al. 2005) indicated sequence disruption (2708 bp) in Renan, with conservation (99%) at the 5' and 3' ends (Fig. 5.5C).

The same Angela retroelement spanned the conservation breakpoint in the Renan sequence and implied its truncation in Glenlea.

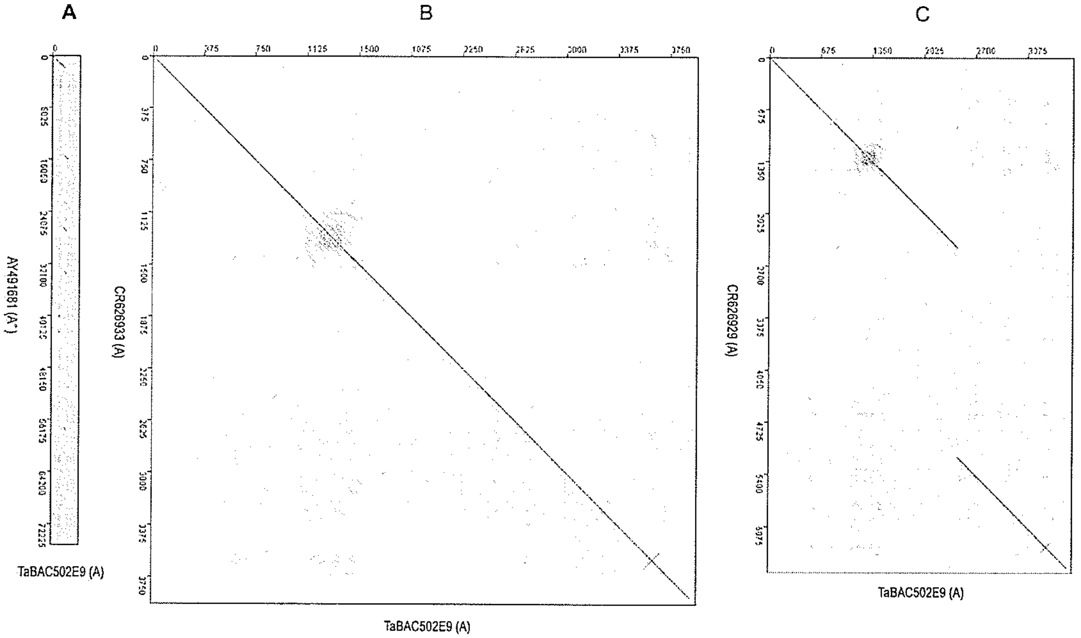


Figure. 5.5 Dot plot comparison of the *Ha* locus from the A-genome of *T. aestivum* cv Glenlea with orthologous/homologous regions. **A.** A^m-genome of *T. monococcum* (AY491681); **B.** A-genome of *T. turgidum* (CR626933) and **C.** A-genome of *T. aestivum* cv Renan (CR626929). In all pairwise comparisons, the *Ha* locus from the A-genome of *T. aestivum* cv Glenlea is represented on the x-axis and the orthologous/homologous regions on the y-axis.

5.4.8 B-genome of cv Glenlea vs. B-genomes of *T. turgidum* and *T. aestivum* cv Renan

The *Ha* locus regions from the B-genomes of cv Glenlea (5330 bp) and *T. turgidum* (5324 bp, Chantret et al. 2005) have 100% sequence identity over their entire length (Fig. 5.6A). Similarly, the sequence of the *Ha* locus from the B-genome of cv Glenlea is 100% identical with the corresponding sequence from *T. aestivum* cv Renan (5322 bp, Chantret et al. 2005) (Fig. 5.6B).

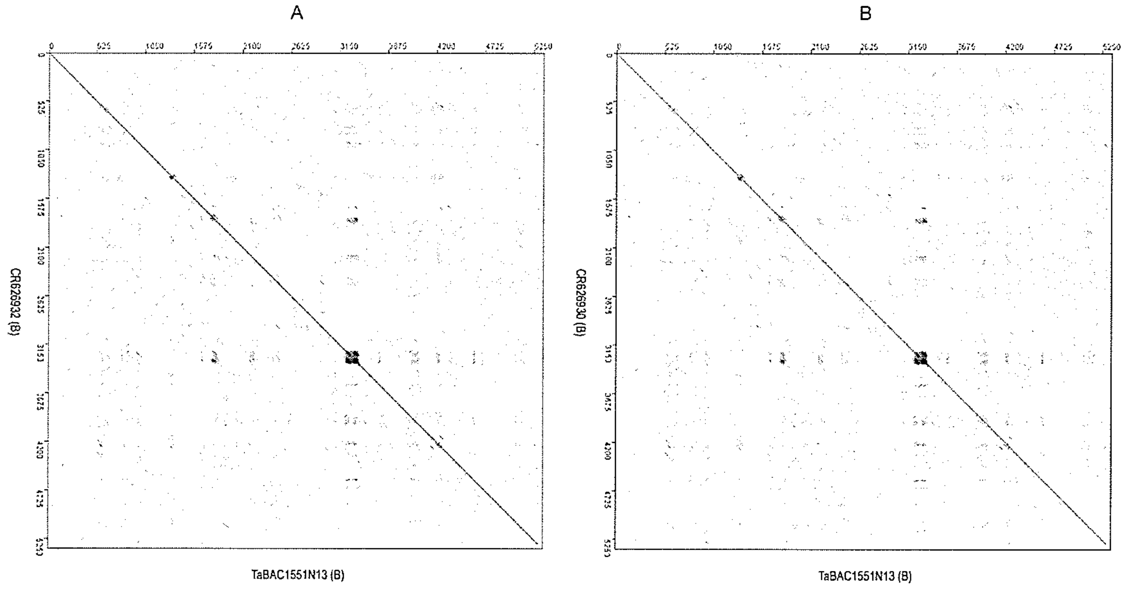


Figure. 5.6 Dot plot comparison of the *Ha* locus from the B-genome of *T. aestivum* cv Glenlea with orthologous/homologous regions. **A.** B-genome of *T. turgidum* (CR626932); **B.** B-genome of *T. aestivum* cv Renan (CR626930). In both pairwise comparisons, the *Ha* locus from the B-genome of *T. aestivum* cv Glenlea is represented on the x-axis and the orthologous/homologous regions on the y-axis.

5.4.9 D-genome of cv Glenlea vs. D-genomes of *Ae. tauschii* and *T. aestivum* cv Renan

Comparison of the partial *Ha* locus from the D-genome of Glenlea (31,607 bp) with the partial D-genome of *Ae. tauschii* (89,806 bp, Chantret et al. 2005) revealed six colinear regions of varying sizes from 254 bp to 2394 bp (Fig. 5.7 and Fig. 5.9A). The percent identity of these conserved regions varied from 87% to 99%. However, *Pinb* was the only conserved gene found in the characterized region, with Romani and Vagabond elements also accounting for the observed colinearity. The lack of *Gsp-D1* containing sequence at the 5' end of the *Ha* locus in the D-genome of Glenlea and the gene delimiting the 3' boundary of the locus in *Ae. tauschii* limited the comparative analysis of the whole loci.

In the region between the *Pina* gene and the gene marking the 3' boundary of the locus, the Glenlea sequence (31,607 bp) has ~90% similarity with the corresponding region from the D-genome of *T. aestivum* cv Renan (35473bp, Chantret et al. 2005). There is conservation of the order and orientation of *Pinb* and *ATPase* genes and retroelements such as Romani, Latidu, Vagabond, WIS and Damocles in both regions (Fig. 5.8 and Fig. 5.9B). The region between *Pina* and *Pinb* in Glenlea is only 2,745 bp compared to 17,701 bp of the corresponding region in cv Renan. The difference is due to the deletion of an ~ 14.9 kb fragment spanning the interval between *Pina* and *Pinb* in cv Glenlea as compared to cv Renan. This deleted fragment includes the entire coding region of *Pina* except for the first 21 bp, leading to the hard phenotype of cv Glenlea. The presence of solo LTRs of Romani and Vagabond retroelements in this intergenic interval suggest their involvement in segmental deletion. In the entire *Ha* locus regions, only two copies of *ATPase* were found in Glenlea as compared to three copies in Renan. As well, one additional copy of *PinB* (pseudogene) found in Renan was not found in Glenlea.

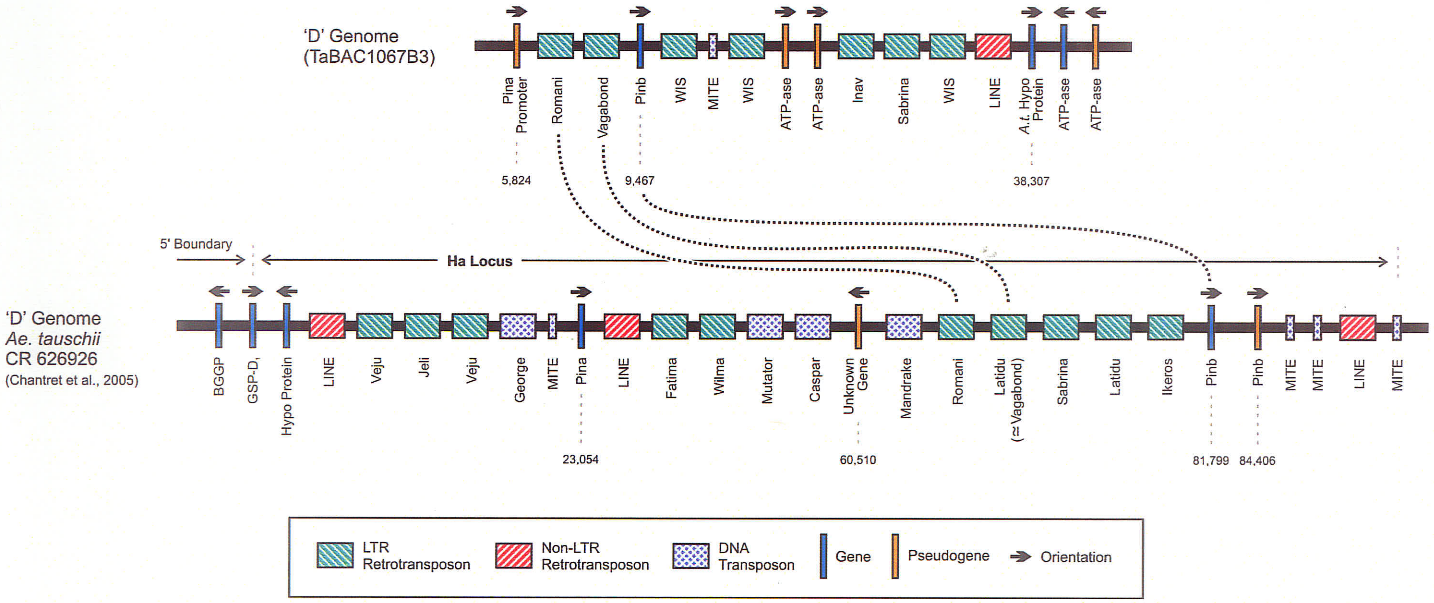


Figure 5.7 Genome organization of the *Ha* loci from the D-genome of *T. aestivum* cv Glenlea, anchored to the orthologous region from the D-genome of *Ae. tauschii* (CR626926). Positions of the important genes of the loci in whole BAC sequences are indicated. Transcriptional orientations of the genes are indicated by arrows. Elements are not drawn to scale.

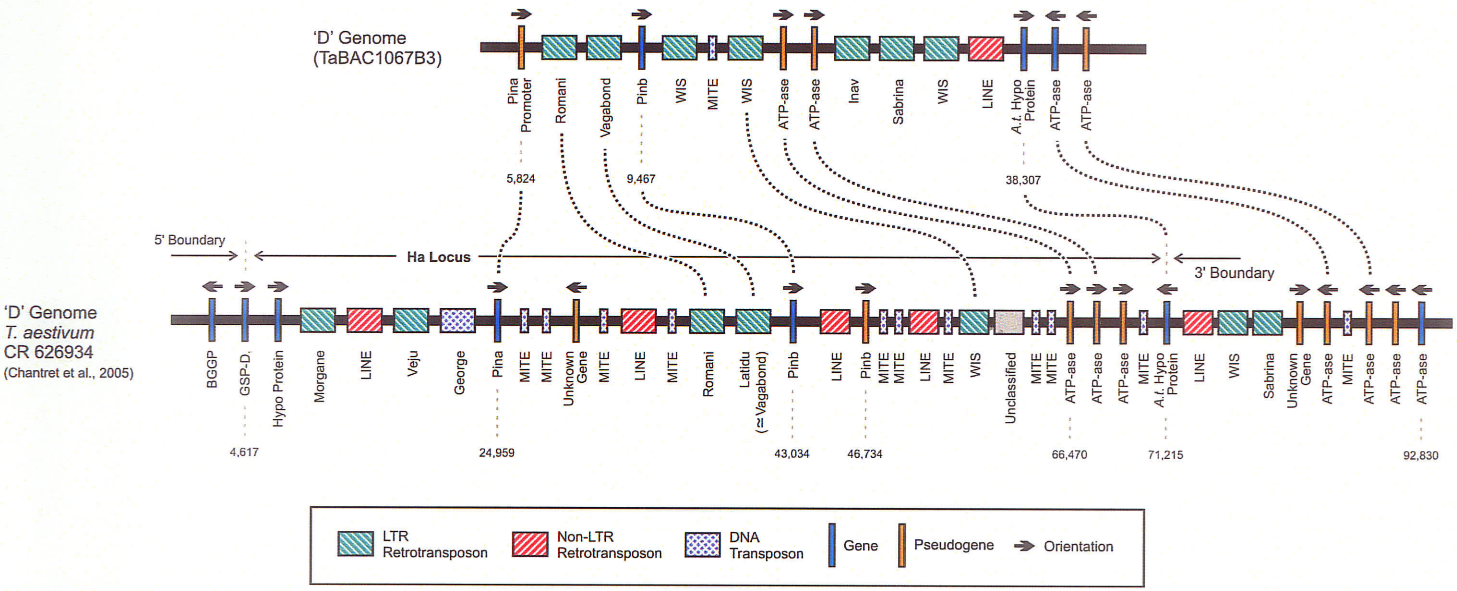


Figure 5.8 Genome organization of the *Ha* loci from the D-genomes of *T. aestivum* cv Glenlea, anchored to the orthologous region from the D-genome of *T. aestivum* cv Renan (CR626934). Positions of the important genes of the loci in whole BAC sequences are indicated. Transcriptional orientations of the genes are indicated by arrows. Elements are not drawn to scale.

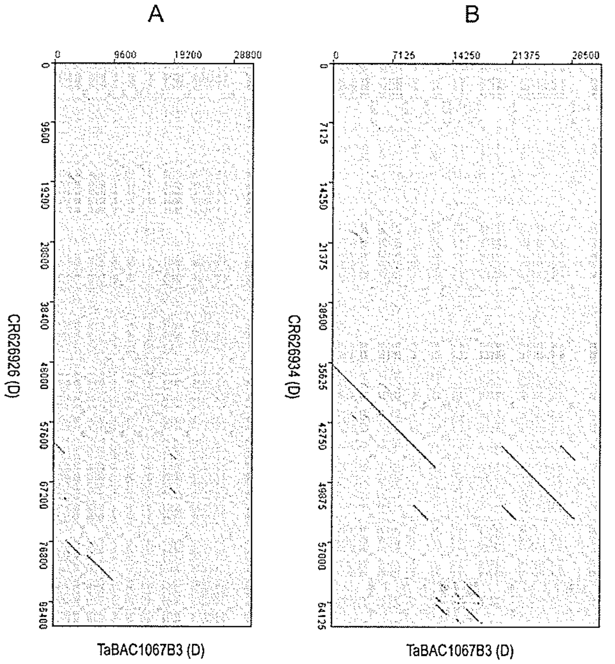


Figure 5.9 Dot plot comparison of the *Ha* locus from the D-genome of *T. aestivum* cv Glenlea with orthologous/homologous regions. **A.** D-genome of *Ae. tauschii* (CR626926) and **B.** D-genome of *T. aestivum* cv Renan (CR626934). In both pairwise comparisons, the *Ha* locus from the D-genome of *T. aestivum* cv Glenlea is represented on the x-axis and the orthologous/homologous regions on the y-axis.

5.5 Discussion

The hardness (*Ha*) locus governing wheat endosperm texture is an excellent example of how mutations from segmental deletions and nucleotide substitutions/frameshifts can cause a spectrum of variation in physical traits, in this case endosperm hardness from soft to very hard (Bhave and Morris 2008b). In this study, the homoeologous *Ha* loci regions from the A-, B- and D-genomes of *T. aestivum* cv Glenlea were sequenced, analysed and compared with the orthologous regions from diploid (*T. monococcum*, *Ae. tauschii*) and tetraploid (*T. turgidum*) ancestral genomes and related *T. aestivum* cv Renan available in the public domain, to gain insight into the molecular evolution of the locus.

5.5.1 Colinearity of the genes and the divergence of the intergenic regions at the *Ha* locus

Studies of evolution of grass genomes and comparative genomics of the members of the *Poaceae* family revealed conservation of gene order (colinearity) across genomes indicating their common origin 60 million years ago (Moore et al. 1995; Gale and Devos 1998; Keller and Feuillet 2000; Tang et al. 2008). The 5' and 3' boundary regions delimited by the presence of the orthologous genes *BGGP* and an *A. thaliana* hypothetical protein (*gene8*) found in rice and all wheat genomes studied to date (Chantret et al. 2004; Chantret et al. 2005; Li et al. 2008) were also identified in the *Ha* locus regions from the A-, B- and D-genomes of *T. aestivum* cv Glenlea with the exception of the 5'-end sequence of the D-genome which was missing. In the *Ha* loci and 3' flanking regions, the order and orientation of the genes (*Gsp-1* and/or *Pina*-truncated, *Pinb* and *ATPase*) were conserved among homoeologues, except for the variation in the copy number of *ATPase* genes. However, in the intergenic regions, the microcolinearity is disrupted because of divergence resulting from insertion and/or amplification of transposable elements and *ATPase* genes (Fig. 5.2). Presence of non-shared retroelements in the colinear positions of the intergenic intervals of the *Ha* locus of different genomes indicate that the divergence of the intergenic regions occurred after the radiation of the A, B, and D genomes from the common ancestor. In wheat, ~70% divergence was observed in the intergenic regions of the *Lr10* loci of three wheat ploidy levels (Isidore et al. 2005). In maize, the intergenic intervals among colinear genes diverged even among inbreds (Brunner et al. 2005). However, analysis of the homoeologous *Glu-1* region indicated sequence conservation both at the genic and large portions of the intergenic regions dominated by retrotransposons (Gu et al. 2006).

5.5.2 Deletions of *Pina* and *Pinb* genes in the A- and B-genomes

Transposable elements constitute 80% of the repetitive fraction which in turn accounts for ~90% of the wheat genome (Li et al. 2004). It was suggested that genomic stress caused by events such as allopolyploidization activates retroelements leading to their integration at new sites by waves of retrotransposition (Grandbastien 1998; Ammiraju et al. 2007). These elements play a crucial role in the evolution of the genome structural organization because of their abundance, diversity and ubiquitous distribution (Havecker et al. 2004; Sabot et al. 2004). Seeding of many copies of diverse groups of homologous sequences (e.g. LTRs) across the genome predisposes the integration sites towards the occurrence of unequal non homologous recombination causing the formation of solo LTRs and illegitimate recombinations which lead to deletions of varying lengths (Devos et al. 2002; Ma et al. 2004).

The *Ha* locus comprises the *Gsp-1*, *Pina* and *Pinb* genes. The latter two genes are present in diploid wheats but, with a few exceptions, are absent in the A- and B-genomes of polyploid wheats. AABB tetraploid wheats are devoid of puroindoline genes altogether while they are only present on the D-genome of AABBDD hexaploid wheats. Indeed, *Gsp-1* sequences were present on all three *Ha* homoeologues of the A-, B- and D-genomes of Glenlea but no puroindoline sequences were found downstream on the A- and B-genomes. Comparative analyses of three homoeologues revealed the presence of a 441 bp Angela solo LTR approximately 1.5 kb downstream of the *Gsp-A1*. This suggested its involvement in the genomic rearrangements leading to the deletion of *Pina* and *Pinb* genes in the A-genome. Similarly, ~2.9 kb downstream of the *Gsp-B1*, a 627 bp fragment of Sabrina retroelement belonging to the *athila* superfamily was identified. This element could have driven the segmental deletion encompassing *Pina* and *Pinb* in the B-genome. Based on the observed size of the locus in the D-genome of *T. aestivum* cv Renan (66,103 bp, CR626934), the estimated deletions are ~62.5 kb and ~61 kb in the A- and B-genomes, respectively. The presence of two different retroelements downstream of *Gsp-1,* indicating independent deletions of the fragments encompassing *Pina* and *Pinb* genes at the *Ha* loci of the A- and B-genomes of *T. aestivum* cv Renan was reported recently (Chantret et al. 2005). Inter-element and/or intra-element recombination driving segmental deletions and duplications were discovered as a major mechanism in the studies of evolution of the angiosperm genomes such as rice, maize, barley and wheat (Vitte and Panaud 2005; Vitte and Bennetzen 2006).

5.5.3 Deletion of *Pina* gene from the D genome

Glenlea is a Canadian Western extra strong wheat (CWES) characterized by a hard endosperm texture when compared to the majority of Canadian bread wheat classes. This characteristic is believed to be the result of its null PINA phenotype even though it is wild type for PINB. A deletion of thousands of nucleotides of sequence including all but the first 21 bp of the *Pina* coding sequence is responsible for this PINA null phenotype. The presence of fragments of Romani and Vagabond retroelements between the truncated *Pina* and *Pinb* genes lead us to believe that this phenotype is another case of retroelement driven deletion. In fact, this region is hypervariable in length. In Glenlea, the intergenic region is only 2745 bp but the same interval is 32 kb in *T. monococcum*, 58 kb in *Ae. tauschi*, and 18 kb in the D-genome of *T. aestivum* cv Renan (Chantret et al. 2004; Chantret et al. 2005). This finding is the first documentation of the sequence organization of the *Pina-D1b* allele resulting in a null-*Pina* associated hard endosperm phenotype. Li et al. (2008) also reported independent deletion of blocks encompassing *Pina* and *Pinb* genes in the *Ha* loci of polyploid *Aegilops* and polyploid *Triticum* species. The only exceptions reported to date are retention of *Pina* and *Pinb* in the A-genome of tetraploid *T. timopheevi* (AAGG) and A^m-genome of *T. zhukovskyi* (A^mA^mAAGG) (Li et al. 2008).

5.5.4 Possible physiological basis of maintenance of dosage of *Pin* genes in polyploid wheat

PINA and PINB belong to the α-amylase inhibitor family of proteins and their possible role in the defense against the microbial degradation of the starch molecules is suggested (Krishnamurthy et al. 2001; Jing et al. 2003; Swan et al. 2006; Li et al. 2008). Also, Li et al. (2008) hypothesised that *Pina* and *Pinb* genes are retained in all diploid (>200) accessions studied so far (Gautier et al. 2000; Lillemo et al. 2002; Massa et al. 2004; Simeone et al. 2006) because of selection pressure associated with decreased plant fitness upon deletion of single copies. In contrast, in polyploids, increase in the number of *Pin* genes could result in higher quantities of PINA and PINB proteins (See et al. 2004) which may lead to deleterious effects such as slow degradation of starch during germination. Germination is crucial in early stand establishment in natural populations to outcompete related diploids (Li et al. 2008). Moreover, studies on synthetic polyploids indicated rapid sequence elimination as an immediate response of the genome in the first few generations (Eckardt 2001). The presence of retroelements in the vicinity of the locus would promote the chances of such deletions and confer a fitness advantage (Gaut and Ross-Ibarra 2008).

5.5.5 Additional evidence in support of the hypothesis of more than one tetraploid ancestor in the origin of hexaploid wheat

Comparison of the *Ha* loci from the A genomes of *T. aestivum* cv Glenlea (this study) and T. *turgidum* (Chantret et al. 2005) revealed that there is 99% identity over their entire length. However, the corresponding region from the A-genome *T. aestivum* cv Renan (Chantret et al. 2005) revealed sequence disruption. The presence of the same Angela retroelement in truncated form (sharing exact coordinates) at the conservation breakpoint in Glenlea that is also found at the tetraploid level, implies that cv Glenlea and cv Renan may have different tetraploid ancestors with different A-genomes. This, in turn, suggests that the second haplotype found in *T. aestivum* cv Renan possibly originated from a different tetraploid progenitor which has not yet been sequenced. Gu et al. (2006) suggested that most of the intergenic retrotransposons of the *Glu-1* loci are inherited from the ancestral genomes because of the presence of shared elements at colinear positions across genomes from different ploidy levels. Isidore et al. (2005) reported that the sequences spanning the *Lr10* loci from the A genomes of tetraploid (*T. durum)* and hexaploid (*T. aestivum* cv Renan) wheat were also different. Similarly, at the *Glu-1* loci, the sequence from *T. durum* is more closely related to *T. aestivum* cv Chinese Spring than to the corresponding region from *T. aestivum* cv Renan, suggesting the involvement of more than one tetraploid ancestor in the origin of hexaploid wheat and therefore independent hexaplodization events (Gu et al. 2004;

Gu et al. 2006). Studies of the *Glu-B1* locus also supported this hypothesis (Ragupathy et al. 2008).

5.5.6 The *Ha* locus is a gene dense region

The *Ha* locus regions represents a highly dense region of the wheat genome as reported previously (Feuillet and Keller 1999; Brooks et al. 2002; Chantret et al. 2004), compared to a gene density of one gene per 75 kb observed in some randomly chosen wheat BACs (Devos et al. 2005). In maize, a gene poor region with one gene per 200 kb has been reported (Haberer et al. 2005). The presence of multiple copies of *ATPase* within the *Ha* loci and the 3' flanking regions is one of the prime causes of this high gene density (Fig. 5.2). In the current model of the wheat genome organization, the presence of gene islands with higher gene density amid vast stretches of repetitive DNA relatively free of genes was suggested (Sandhu and Gill 2002; Anderson et al. 2003). In our study the retrotransposon content was found to be 47% (TaBAC502E9), 44% (TaBAC1551N13) and 31% (TaBAC1067B03) compared to the whole genome estimate of 60-80 % (Li et al. 2004). This low content also suggests that the *Ha* locus and its flanking regions are gene rich in accordance with the emerging evidence that retroelements are predominantly localized in the heterochromatin regions with poor gene content (Bennetzen 2000c; Ma and Bennetzen 2006).

5.5.7 Mechanisms of genome evolution and its impact on agriculturally important loci

Plant genomes are dynamic entities in terms of their structural organization and size. They are driven by evolutionary forces such as transposable elements (Bennetzen 2005; Wessler 2006a and 2006b). Though there are reports of conservation of overall gene content across plant genomes (Schulman and Kalendar 2005) and gene order across grass genomes (Moore et al. 1995; Keller and Feuillet 2000), there is enormous expansion and divergence of intergenic regions due to the dynamics of retroelements, accounting for the large genome sizes of maize, barley and wheat (SanMiguel et al 1996; Gu et al. 2004; Scherrer et al. 2005). The presence of different kinds of truncated retroelements in the intergenic interval of the *Ha* loci, multiple copies of *ATPase* genes and multiple insertions of the same element (e.g. WIS element in the D-genome) indicate multiple mechanisms of genome evolution such as retroelement insertion, amplification, gene duplication and segmental deletion. While retroelement amplification and gene duplication led to expansion of the *Ha* loci implying genome evolution, the inter-element and/or intra-element recombination driven deletions, not only led to the shrinkage of the *Ha* loci but also significantly impacted phenotypes such as endosperm texture by deletion of *Pina* and *Pinb* genes. Earlier studies also documented retroelements as one of the major evolutionary forces in the dynamics of genomes affecting its size and structural organization leading to changes in gene content (genotype) and phenotype (Wicker et al. 2001; Feuillet et al. 2001; Morgante 2006).

CHAPTER 6

GENERAL DISCUSSION

Nothing in biology makes sense except in the light of evolution.

Dobzhansky (1973)

In the 1970s, there were unanswered questions about the 'C-value paradox' i.e. the lack of correlation between genome complexity with reference to the organismal complexity (Gregory 2005). Later, it was proposed that 95% of the genome was noncoding which was referred as 'junk DNA' because of the lack of contribution to the phenotype of an organism (Ohno 1972). Orgel and Crick (1980) and Doolittle and Sapienza, (1980) introduced the landmark theory of the 'selfish DNA' for transposable elements dominating the junk portion of the genome because of their unique ability to replicate and perpetuate themselves without any contribution to either housekeeping functions of the organism or obvious phenotypic effects. However, emergence of studies at the genome-scale and results of sequencing projects have changed those perspectives drastically by indicating the possible roles of transposable elements in the evolution of genomes and genes (Fedoroff 2000; Bennetzen 2005). Now it is accepted as fact that transposable elements are genomic scrap yards that they co-exist and co-evolve with the host genes, mainly because of their ability to contribute to the evolutionary success of the host genome as a whole, by the following mechanisms (Wessler 2006b):

1. Engineering chromosomal rearrangements such as inversions, deletions, duplications and translocations which help to rapidly restructure the genome favouring its adaptive responses, upon exposure to biotic and abiotic stresses.

2. Creating novel structural genetic variation such as new alleles by inserting into or in the immediate vicinity of genes, or by generating new expression patterns by donating promoter and/or other regulatory sequences to neighboring genes, alternative RNA splicing and exon shuffling.

3. Altering the expression of the genes by epigenetic mechanisms such as DNA and histone methylation upon integration in regions adjacent to the host genes.

In our study, we observed the phenotypic impacts resulting from one of the mechanisms described above namely, retroelement mediated chromosomal rearrangements in the adaptive evolution of the genome. Specifically, the molecular evolution of the *Glu-B1* locus involved a retroelement mediated duplication while evolution of the *Ha* locus involved retroelement driven deletion events. Interestingly, both genome reshaping events have implications in wheat quality, though quality *per se* which has significant relevance to agriculture has no meaning in biology.

Gene duplication was hypothesized to be the cause of overexpression of the Bx7 HMW-GS in the landrace TAA36 (Lukow et al. 1992) and in the cultivar Red River 68 (D'Ovidio et al. 1997). By characterizing the genome organization of a BAC clone harbouring the *Glu-B1* locus of Glenlea, another cultivar with the Bx7^{OE} phenotype, Cloutier et al. (2005) identified a retrotransposon mediated segmental duplication responsible for the *Bx7* gene duplication. In the present study, the gene duplication was confirmed in 43 Bx7 overexpressing cultivars in a survey of 412 accessions including 96 diploids/tetraploids, from 41 countries. Discovery of three *T. turgidum* Bx7^{OE} lines indicated that the *Bx7* gene duplication was a pre-hexaplodization event as hypothesised by Butow et al. (2004). The presence of tandem arrays of genes is a common feature of plant genomes. Among characterized genes, nearly 12-16% (in *Arabidopsis*) and 14% (in rice) are present in tandem organization (Arabidopsis Genome Initiative 2000; International Rice Genome Sequencing Project 2005). Unequal recombination between dispersed copia-like elements, resulting in tandem duplication at *white* (*w*) locus in *Drosophila* has been reported (Goldberg, 1983). Similarly, unequal recombination mediated by the tandem repeats was identified at the *a1* locus in maize (Yandequ-Nelson et al. 2006). Very recently, in tomato, retrotransposon driven duplication of a 24.7 kb region encompassing *SUN*, a major gene controlling the fruit shape resulting in elongated phenotype from the wild round type, was reported (Xiao et al. 2008).

Phylogenetic analysis of RT domains and the estimates of LTR divergence of the family of the retroelement Sasanda_EU157184-1 which mediated the duplication of the *Bx7* gene indicated that the element is at least 1.2 to 1.8 million years old, with at least five sub-families among the characterized members. The insertion time estimates based on the divergence between the two LTRs revealed the elements' recent transposition activity, inferred from the identical LTRs found in 49 of the 89 elements studied. This estimate is in line with the findings that most retroelements were found to be amplified within the last three million years in the genomes of maize, rice and wheat (SanMiguel et al. 1998; Piegu et al. 2006; Wicker and Keller 2007). However, this is the first report in wheat where large numbers of complete elements (49) have not accumulated mutations in their LTRs. Transposable elements tend to be kept silent by the host genome mainly by epigenetic mechanisms operating at the nucleotide level as well at a higher order (chromatin) level (Comai 2000; Lippman et al. 2004; Chen and Ni 2006, Cheng et al.

2006; Slotkin and Martienssen 2007). The former includes DNA methylation while the latter comprises histone modification such as acetylation and methylation, which form the basis for altering the chromatin from the active (euchromatin) to inactive (heterochromatin) state. Also, siRNA mediated specific degradation of the transposon transcripts (RNA interference-RNAi) was discovered recently (Slotkin and Martienssen 2007). Mutants such as *decrease in DNA methylation1* (*DDM1*, Vongs et al. 1993) would change the methylation landscape of the genome and therefore, in the genomic background where *DDM1* is present, transcriptionally silent transposable elements will be activated (Hirochika et al. 2000). Also, genomic-shock (McClintock 1984) caused by abiotic and biotic stresses may favour histone modification associated chromatin remodelling, resulting in the amplification of retroelements (Comai 2000; Lippman et al. 2004; Chen and Ni 2006). Stresses such as drought, tissue culture, allopolyploidization, introgressive hybridization, UV light and pathogen infection were correlated with activation of retroelements (Kalender et al. 2000; Kukuchi et al. 2003; Levy and Feldman 2004; Liu and Wendel 2000; Ramallo et al. 2008; Ansari et al. 2007). Though the genome has only a tiny fraction of the autonomous elements among the large numbers of families of mobile elements, many with thousands of copies, the retention of recognition sequences for *trans* acting factors (transposon proteins) will result in the mobility of non-autonomous elements as well.

Mutations resulting from segmental deletions and nucleotide substitutions/indels at the *Ha* locus resulted in a spectrum of variation in endosperm texture (Bhave and Morris 2008b). In hexaploid cultivar Glenlea, the presence of two different truncated retroelements downstream to *Gsp-A1* and *Gsp-B1* indicated independent deletions of the fragments encompassing the *Pina* and *Pinb* genes of the *Ha* loci from the A- and B-genomes. Similarly, the presence of fragments of Romani and Vagabond retroelements between the *Pina* (truncated) and *Pinb* genes suggest their involvement in the deletion of the coding region of *Pina,* hence leading to the hard endosperm phenotype of hexaploid cv Glenlea. Deletion of genomic DNA mediated by tandem repeats of transposable elements such as LTRs, terminal inverted repeats (TIRs) were reported in wheat (Dubcovsky and Dvorak 2007; Wicker and Keller 2007), rice (Ma et al. 2004; Vitte et al. 2007) and in other plant genomes (Vitte and Panaud 2005; Vitte and Bennetzen 2006). Specifically, Chantret et al. (2005) and Li et al. (2008) also reported independent deletion of blocks encompassing *Pina* and *Pinb* genes in the *Ha* loci of polyploid *Aegilops* and polyploid *Triticum* species. A physiological bottleneck associated with increased α-amylase inhibitor activity upon an increase in *Pina* and *Pinb* gene dosage in polyploid wheat was suggested as the reason for *puroindoline* gene deletions in polyploid wheat because it resulted in poor germination and consequently poor stand establishment unfavourable to the adaptive evolution of the species (Li et al. 2008).

The current model of genome organization suggests the presence of gene rich and gene poor regions in large plant genomes. Gene poor regions with a gene density of one gene per 75 kb and one gene per 200 kb were reported for wheat and maize, respectively (Devos et al. 2005; Haberer et al. 2005). In this context, the *Ha* locus of Glenlea represented a gene rich region with gene density estimates of one gene per ~12 kb (A genome) , one gene per ~9 kb (B genome) and one gene per ~7 kb (D genome), as reported previously (Brooks et al. 2002; Chantret et al. 2004). Tandemly repeated *ATPase* within/near the *Ha* loci region contributed significantly to the high gene density and the structural organization of this locus is in accordance with the current view of the wheat genome organization which suggests presence of gene islands (Sandhu and Gill 2002).

The understanding of genome organizations at the *Glu-B1* and the *Ha* loci have practical applications in wheat breeding through marker aided selection which will improve the precision and efficiency of selection for desirable genotypes (Sorrells 2007). The dominant left and right junction SCAR markers developed for the *Glu-B1* locus can be used to select Bx7^{OE} phenotypes. Similarly, the genome organization of the *Ha* loci can be used to generate markers for tagging the hardness phenotype. As well, the finding that the Sasanda element family have moderate numbers of copies (347 per haploid genome) and the possibilities of insertion site polymorphism resulting from its very recent activity, can be used to generate retroelement based markers such as SSAP, REMAP and IRAP which have applications in saturating linkage maps and in genotyping studies (Schulman 2007).

The study also generated additional evidence in support of the multiple polyploidization events that led to the formation of bread wheat (hexaploid T. aestivum). This was accomplished by documenting the presence of two different haplotypes both at the *Glu-B1* locus (B-genome) as well as the *Ha* locus (A-genome). More specifically, we provided evidence towards the involvement of multiple tetraploids (AABB) in the hybridization events with the D-genome donors.The presence of two different genome organizations shared between *T. turgidum* and *T. aestivum* at the *Glu-B1* locus (with and without segmental duplication) represented evidence for the multiple tetraploid hypothesis based on the B genome. Similarly, the presence of shared *Ha* locus organization in the A genomes of *T. aestivum* cv Glenlea and *T. turgidum* (Chantret et al.

2005) but distinct from the corresponding region of *T. aestivum* cv Renan (Chantret et al 2005) constituted further evidence for the contribution of more than one tetraploid (differing at the *Ha* loci of A-genome) in the evolutionary origin of the bread wheat. Gu et al. (2006) also suggested this hypothesis based on distinct *Glu-1A*-genome lineages.

Hence, it can be concluded that the adaptive evolution mediated by the repetitive fraction of the genome can have significant impact on agriculturally important loci such as *Glu-B1* and *Ha*, governing quality parameters associated with dough rheology and endosperm texture.

CHAPTER 7

LITERATURE CITED

Abad JP, De Pablos B, Osoegawa D, De Jong PJ, Martin-Gallardo A, Villasante A (2004) Genomic analysis of *Drosophila melanogaster* telomeres: Full-length copies of HeT-A and TART elements at telomeres. Mol Biol Evol 21: 1613-1619

Alix K, Ryder CD, Moore J, King GJ, Heslop-Harrison JSP (2005) The genome organization of retrotransposons in *Brassica oleracea.* Plant Mol Biol 59: 839-851

Allaby RG, Banerjee M , Brown TA (1999) Evolution of the high molecular weight glutenin loci of the A, B, D and G genomes of wheat. Genome 42: 296-307

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410

Ammiraju JSS, Zuccolo A, Yu Y, Song X, Piegu B, Chevalier F, Walling JG, Ma J, Talag J, Brar S, SanMiguel PJ, Jiang N, Jackson SA, Panaud O, Wing RA (2007) Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus *Oryza.* Plant J 52: 342-351

Anderson OD and Greene FC (1989) The characterization and comparative analysis of HMW glutenin genes from genomes A and B of hexaploid wheat. Theor Appl Genet 77: 689-700

Anderson OD, Greene FC, Yip RE, Halford NG, Shewry PR, and Malpica-Romero JM (1989) Nucleotide sequences of the two high molecular weight glutenin genes from the D genome of hexaploid bread wheat, *Triticum aestivum* L. cv Cheyenne. Nucleic Acids Res 17: 461-462

Anderson OD, Rausch C, Moullet O, Lagudah E.S. (2003) The wheat D genome HMW glutenin loci: BAC sequencing, gene distribution and retrotransposon clusters. Funct Integr Genomics 3: 56-68

Anjum FM, Khan MR, Din A, Saeed M, Pasha I, Arshad MU (2007) Wheat Gluten: High molecular weight glutenin subunits-structure, genetics and relation to dough elasticity. J Food Sci 72: R56-R63

Ansari KI, Walter S, Brennan JM, Lemmens M, Kessans S, McGahern A, Egan D, Doohan FM (2007) Retrotransposon and gene activation in wheat in response to mycotoxigenic and non-mycotoxigenic-associated Fusarium stress. Theor Appl Genet 114: 927- 937

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

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9: 211-215

Barro F, Rooke L, Bekes F, Gras P, Tatham AS, Fido R, Lazzeri PA, Shewry PR, Barcelo P (1997) Transformation of wheat with high molecular weight subunit genes results in improved functional properties. Nat Biotechnol 15: 1295-1299

Beecher B, Smidansky E, See D (2001) Mapping and analysis of hordoindolines. Theor Appl Genet 102: 833-840

Beecher B, Bettge A, Smidansky E, and Giroux MJ (2002) Expression of wild-type *Pinb* sequence in transgenic wheat complements a hard phenotype. Theor Appl Genet 105: 870-877

Bennet MD and Leitch IJ (1995) Nuclear DNA amounts in angiosperms. Ann Bot (Lon) 76: 113-176

Bennetzen JL (2000a) Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. Plant Cell 12: 1021-1029

Bennetzen JL (2000b) Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42: 251-269

Bennetzen JL (2000c) The many hues of plant heterochromatin. Genome Biol 1: reviews107.1-107.4

Bennetzen JL (2002) Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica 115: 29-36

Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev 15: 621-627

Bennetzen JL (2007) Patterns in grass genome evolution. Curr Opin Plant Biol 10: 176-181

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

Bhave M, Morris CF (2008a) Molecular genetics of puroindolines and related genes: regulation of expression, membrane binding properties and applications. Plant Mol Biol 66: 221-231

Bhave M, Morris CF (2008b) Molecular genetics of puroindolines and related genes: allelic diversity in wheat and other grasses. Plant Mol Biol 66: 205-219

Biemont C and Vieira C (2006) Junk DNA as an evolutionary force. Nature 443: 521-524

Brodie R, Roper RL, and Upton C (2004) JDOTTER: a java interface to multiple dotplots generated by dotter. Bioinformatics 20: 279-281

Brooks SA, Huang L, Gill BS, Fellers JP (2002) Analysis of 106 kb of contiguous DNA sequence from the D genome of wheat reveals high gene density and a complex arrangement of genes related to disease resistance. Genome 45: 963-972

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A (2005) Evolution of DNA sequence non homologies among maize inbreds. Plant Cell 17: 343-360

Bureau TE, White SE, Wessler SR (1994) Transduction of a cellular gene by a plant retroelement. Cell 20: 479-480

Bushuk W (1998) Wheat breeding for end-product use. Euphytica 100: 137-145

Butow BJ, Gras PW, Haraszi R, Bekes F (2002) The effects of different salts on mixing and extension parameters on a diverse group of wheat cultivars using 2g mixographs and extensigraph methods. Cereal Chem 79: 823-826

Butow BJ, Ma W, Gale KR, Cornish GB, Rampling L, Larroque O, Morell MK, Bekes F (2003) Molecular discrimination of Bx7 alleles demonstrates that a highly expressed high molecular weight glutenin allele has a major impact on wheat flour dough strength. Theor Appl Genet 107: 1524-1532

Butow BJ, Gale KR, Ikea J, Juhasz A, Bedo Z, Tamas L, Gianibelli MC (2004) Dissemination of the highly expressed Bx7 glutenin subunit (*Glu-B1al* allele) in wheat as revealed by novel PCR markers and RP-HPLC. Theor Appl Genet 109: 1525–1535

Caceres M, Ranz JM, Barbadilla A, Long M, Ruiz A (1999) Generation of a widespread *Drosophila* inversion by a transposable element. Science 285: 415-418

Caldwell KS, Dvorak J, Lagudah ES, Akhunov E, Luo MC, Wolters P, Powell W (2004) Sequence polymorphism in polyploid wheat and their D-genome diploid ancestor. Genetics 167: 941-947

Chantret N, Cenci A, Sabot F, Anderson O, Dubcovsky J (2004) Sequencing of the *Triticum monococcum hardness* locus reveals good microcolinearity with rice. Mol Genet Genomics 271: 377-386

Chantret N, Salse J, Sabot F, Rahman S, Bellec A, Laubin B, Dubois I, Dossat C, Sourdille P, Joudrier P, Gautier M, Cattolico L, Beckert M, Aubourg S, Weissenbach J, Caboche M, Bernard M, Leroy P, Chalhoub B (2005) Molecular basis of evolutionary events that shaped the *Hardness* locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). Plant Cell 17: 1033-1045

Chen ZJ, Ni Z (2006) Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. BioEssays 28: 240-252

Cheng C, Daigen M, Hirochika H (2006) Epigenetic regulation of the rice retrotransposon Tos17. Mol Genet Genomics 276: 378-390

Cloutier S, Banks T, Nilmalgoda S (2005) Molecular understanding of wheat evolution at the *Glu-B1* locus. In: Proceedings of the international conference on plant genomics and biotechnology: Challenges and opportunities. Raipur, India, p 40

Cloutier S, McCallum BD, Loutre C, Banks TW, Wicker T, Feuillet C, Keller B, Jordan MC (2007) Leaf rust resistance gene *Lr1*, isolated from bread wheat (*Triticum aestivum* L.) is a member of the large *psr567* gene family. Plant Mol Biol 65: 93-106

Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L (2005) Chromosome evolution in eukaryotes: a multi-kingdom perspective. Trends Genet 21: 673-682

Comai L (2000) Genetic and epigenetic interactions in allopolyploid plants. Plant Mol Biol 43: 387-399. Curr Opin Biotechnol 17: 168–173

D'Ovidio R, Masci S, Porceddu E, Kasarda D. (1997) Duplication of the high molecular weight glutenin subunit gene in bread wheat (*Triticum aestivum* L.) cultivar 'Red River 68'. Plant Breed 116: 525–531

Darlington HF, Rouster J, Hoffmann L, Halford NG, Shewry PR, Simpson D (2001) Identification and characterization of hordoindolines from barley grain. Plant Mol Biol 47: 785-794

De Bustos A, and Jouve N (2003) Characterization and analysis of new HMW glutenin alleles encoded by the *Glu-R1* locus of *Secale cereale*. Theor Appl Genet 107: 74-83

Devos KM, Brown JKM, Bennetzen JL (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res 12: 1075-1079

Devos KM, Ma J, Pontaroli AC, Pratt LH, Bennetzen JL (2005) Analysis and mapping of randomly chosen bacterial artificial chromosomes clones from hexaploid bread wheat. Proc Natl Acad Sci USA 102: 19243-19248

Dexter JE, Marchylo BA, MacGregor AW, Tkachuk R (1989) The structure and protein composition of vitreous and starchy durum wheat kernels. J Cereal Science 10: 19-32

Dilbirligi M, Erayman M, Sandhu D, Sidhu D, Gill KS (2004) Identification of wheat chromosomal regions containing expressed resistance genes. Genetics 166: 461-481

Dixit A, Ma KH, Yu JW, Cho EG, Park YJ (2006) Reverse transcriptase domain sequences from mungbean (*Vigna radiata*) LTR retrotransposons: Sequence characterization and phylogenetic analysis. Plant Cell Rep 25: 100-111

Dobzhansky T (1973) Nothing in Biology makes sense except in the light of evolution. Amer Biol Teach 35:125-129

Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284: 601-603

Dubcovsky J, Dvorak J (2007) Genome plasticity, a key factor in the success of polyploidy wheat under domestication. Science 316: 1862-1866

Dubreil L, Meliande S, Chiron H, Compoint JP, Quillien L, Branlard G, Marion D (1998) Effect of puroindolines on the bread making properties of wheat flour. Cereal Chem 75: 222-229

Dvorak J, Luo MC, Yang ZL, Zhang H (1998) The structure of the *Aegilops tauschii* genepool and the evolution of hexaploid wheat. Theor Appl Genet 97: 657-670.

Eagles HA, Bariana HS, Ogbonnaya FC, Rebetzke GJ, Hollamby GJ, Hendry RJ, Henschke PH, Carter M (2001) Implementation of markers in Australian wheat breeding. Aust J Agric Res 52: 1349-1356

Eckardt NA (2001) A sense of self: the role of DNA sequence elimination in allopolyploidization. Plant Cell 13: 1699-1704

Endo TR, Gill BS (1996). The deletion stocks of common wheat. J Hered 87: 295-307

Erayman M, Sandhu D, Sidhu D, Dilbirligi M, Baenziger PS, Gill KS (2004) Demarcating the gene-rich regions of the wheat genome. Nucleic Acids Res 32: 3546-3565

Ewing B, Hillier L, Wendl MC, Green G (1998) Base-calling of automated sequencer traces using *phred.* I. accuracy assessment. Genome Res 8: 175-185

Fedoroff N (2000) Transposons and genome evolution in plants. Proc Nat Acad Sci USA 97: 7002-7007

Feldman M (2001) Origin of cultivated wheat. In: Bonjean AP, Angus WJ (Eds) The world wheat book: A history of wheat breeding. First edn, Intercept, France, pp 3-56

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39: 783-791

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3: 329-341

Feuillet C, Keller B (1999) High gene density is conserved at syntenic loci of small and large grass genomes. Proc Natl Acad Sci USA 96: 8665-8670

Feuillet C, Penger A, Gellner K, Nast A, Keller B (2001) Molecular evolution of *receptor-like kinase* genes in hexaploid wheat : Independent evolution of orthologs after polyploidization and mechanisms of local rearrangements at paralogous loci. Plant Physiol 125: 1304-1313

Flavell AJ, Knox MR, Pearce SR, Ellis THN (1998) Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. Plant J 16: 643-650

Forde J, Malpica J, Halford NG, Shrewry PR, Anderson OD, Greene FC, Miflin BJ (1985) The nucleotide sequence of a HMW glutenin subunit gene located on chromosome1A of wheat *Triticum aestivum* L. Nucleic Acids Res 13: 6817-6832

Gale KR (2005) Diagnostic DNA markers for quality traits in wheat. J Cereal Sci 41: 181-192

Gale MD, Devos KM (1998) Comparative genetics in the grasses. Proc Natl Acad Sci USA 95: 1971-1974

Gao S, Gu YQ, Wu J, Coleman-Derr D, Huo N, Crossman C, Jia J, Zuo Q, Ren Z, Anderson OD, Kong X (2007) Rapid evolution and complex structural organization in genomic regions harbouring multiple prolamine genes in the polyploidy wheat genome. Plant Mol Biol 65: 189-203

Gaut BS, Ross-Ibarra J (2008) Selection on major components of angiosperm Genomes. Science 320: 484-486

Gaut BS, Morton BR, McCaigBC, Clegg MT (1996) Substitution rate comparisons between grasses and palms: Synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. Proc Nat Acad Sci USA 93: 10274-10279

Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK (2007) Recombination: an under appreciated factor in the evolution of plant genomes. Nat Rev Genet 8: 77-84

Gautier MF, Aleman ME, Guirao A, Marion D, and Joudrier P (1994) *Triticum aestivum* puroindolines, two basic cysteine-rich seed proteins: cDNA sequence analysis and developmental gene expression. Plant Mol Biol 25: 43-57

Gautier MF, Cosson P, Guirao A, Alary R, and Joudrier P (2000) Puroindoline genes are highly conserved in diploid ancestor wheats and related species but absent in tetraploid *Triticum species*. Plant Sci 153: 81–91

Gianibelli MC, Echaide M, Larroque OR, Carrillo JM, Dubcovsky J (2002) Biochemical and molecular characterization of *Glu-1* loci in Argentinean wheat cultivars. Euphytica 128: 61-73

Giles RJ, Brown TA (2006) *GluDy* allele variations in *Aegilops tauschii* and *Triticum aestivum*: implications for the origins of hexaploid wheats. Theor Appl Genet 112: 1563-1572

Gill BS, Appels R, Botha-Oberholster AM, Buell CR, Bennetzen JL, Chalhoub B, Chumley F, Dvorak J, Iwanaga M, Keller B, Li W, McCombie WR, Ogihara Y, Quetier F, Sasaki T (2004) International Genome Research on Wheat Consortium: A workshop report on wheat genome sequencing. Genetics 168: 1087-1096

Gill KS, Gill BS, Endo TR, Boyko EV (1996a) Identification and high-density mapping of gene-rich regions in chromosome group 5 of wheat. Genetics 143: 1001-1012

Gill KS, Gill BS, Endo TR, Taylor T (1996b) Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. Genetics 144: 1883-1891

Giovanni MD Cenci A, Janni M, D'Ovidio RA (2008) LTR *copia* retrotransposon and *Mutator* transposons interrupt *Pgip* genes in cultivated and wild wheats. Theor Appl Genet 116: 859-867

Giroux MJ, Morris CF (1997) A glycine to serine change in puroindoline b is associated with wheat grain hardness and low levels of starch-surface friabilin. Theor Appl Genet 95: 857-864

Goldberg ML, Sheen JY, Gehring WJ, Green MM (1983) Unequal crossing over associated asymmetrical synapsis between nomadic elements in the *Drosophila melanogaster* genome. Proc Natl Acad Sci USA 80: 5017-5021

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

Grandbastien M (1992) Retroelements in higher plants. Trends Genet 8: 103-108

Grandbastien M (1998) Activation of plant retrotransposons under stress conditions. Trends Plant Sci 3: 181-187

Greenwell P, Schofeld JD (1986) A starch granule protein associated with endosperm softness in wheat. Cereal Chem 63: 379-380

Gregory TR (2005) Genome size evolution in animals. In: Gregory TR (Ed) The evolution of the genome. Elsevier Academic press, MA, USA, pp 4-87

Gu YQ, Coleman-Derr D, Kong X, Anderson OD (2004) Rapid genome evolution revealed by comparative sequence analysis of orthologous regions from four *Triticeae* genomes. Plant Physiol 135: 459-470

Gu YQ, Salse J, Coleman-Derr D, Dupin A, Crossman C, Lazo GR, Huo N, Belcram H, Ravel C, Charmet G, Charles M, Anderson OD, Chalhoub B (2006) Types and rates of sequence evolution at the high-molecular-weight glutenin locus in hexaploid wheat and its ancestral genomes. Genetics 174: 1493-1504

Haberer G, Young S, Bharti AK, Gundlach H, Raymond C, Fuks G, Butler E, Wing RA, Rounsley S, Barren B, Nusbaum C, Mayer KFX, Messing J (2005) Structure and architecture of the maize genome. Plant Physiol 139: 1612-1624

Halford NG, Forde J, Anderson OD, Greene FC, Shewry PR (1987) The nucleotide and deduced amino acid sequence of an HMW glutenin subunit gene from chromosome 1B of bread wheat (*Triticum aestivum* L.) and comparison with those of genes from chromosome 1A and 1D. Theor Appl Genet 75: 117-126

Halford NG, Field JM, Blair H, Urwin P, Moore K, Robert L, Thompson R, Flavell RB, Tatham AS, and Shewry PR (1992) Analysis of HMW glutenin subunits encoded by chromosome 1A of bread wheat (*Triticum aestivum* L.) indicates quantitative effects on grain quality. Theor Appl Genet 83: 373-378

Han JS, Szak ST, Boeke JD (2004) Transcriptional disruption by the L1 retrotransposon and the implications for mammalian transcriptomes. Nature 429: 268-274

Harberd NF, Flavell RB Thompson RD (1987) Identification of a transposon-like insertion in a *Glu-1* allele of wheat. Mol Gen Genet 209: 326-332

Havecker ER, Gao X, Voytas DF (2004) The diversity of LTR retrotransposons. Genome Bio 5: 225.1-225.6

Hawkins JS, Hu G, Rapp RA, Grafenberg JL, Wendel JF (2008) Phylogenetic determination of the pace of transposable element proliferation in plants: *copia* and LINE-like elements in *Gossypium.* Genome 51: 11-18

Higgins D, Thompson J, Gibson T, Thompson JD, Higgins DG, Gibson TJ (1994) ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680

Hirochika H, Okamoto H, Kakutani T (2000) Silencing of retrotransposons in Arabidopsis and reactivation by the *ddm1* mutation. Plant Cell 12: 357-368

Hoen DR, Park KC, Elrouby N, Yu Z, Mohabir N, Cowan RK, Bureau TE (2006) Transposon-mediated expansion and diversification of a family of *ULP*-like genes. Mol Biol Evol 23: 1254-1268

Hogg AC, Sripo T, Beecher B, Martin JM, Giroux MJ (2004) Wheat puroindolines interact to form friabilin and control wheat grain hardness. Theor Appl Genet 108: 1089-1097

Holligan D, Zhang X, Jiang N, Pritham EJ, Wessler SR (2006) The transposable element landscape of the model legume *Lotus japonicus.* Genetics 174: 2215-2228

Hong BH, Rubenthaler GL and Allan RE (1989) Wheat pentosans: cultivar variation and relationship to kernel hardness. Cereal Chemistry 66: 369-373

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9: 868-877

Huang XQ, Röder MS (2005) Development of SNP assays for genotyping the puroindoline b gene for grain hardness in wheat using pyrosequencing. J Agric Food Chem 53: 2070-2075

Huang XQ, Cloutier S (2007) Hemi-nested touchdown PCR combined with primer-template mismatch PCR for rapid isolation and sequencing of low molecular weight glutenin subunit gene family from a hexaploid wheat BAC library. BMC Genetics 8: 18. DOI: 10.1186/1471-2156-8-18

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

Huang L, Brooks SA, Li W, Fellers JP, Trick HN, Gill BS (2003) Map-based cloning of leaf rust resistance gene *Lr21* from the large and polyploid genome of bread Wheat. Genetics 164: 655-664

Ikeda TM, Ohnishi, Nagamine T, Oda S, Hisatomi T, Yano H (2005) Identification of new puroindoline genotypes and their relationship to flour texture among wheat cultivars. J Cereal Sci 41: 1-6

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

Isidore E, Scherrer B, Chalhoub B, Feuillet C, Keller B (2005) Ancient haplotypes resulting from extensive molecular rearrangements in the wheat A genome have been maintained in species of three different ploidy levels. Genome Res 15: 526-536

Jiang N, Bao Z, Temnykh S, Cheng Z, Jiang J, Wing RA, McCouch SR, Wessler SR (2002) *Dasheng:* A recently amplified non-autonomous long terminal repeat element that is a major component of pericentric regions in rice. Genetics 161: 1293-1305

Jiang N, Bao S, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature 431: 569-573

Jing W, Demeoe AR, Vogel HJ (2003) Conformation of a bactericidal domain of puroindoline a: structure and mechanism of action of a 13-residue antimicrobial peptide. J Bacteriol 185: 4938-4947

Jolly CJ, Rahman S, Kortt A, Higgins TJV (1993) Characterisation of the wheat Mr 15000 grain-softness protein and analysis of the relationship between its accumulation in the whole seed and grain softness. Theor Appl Genet 86: 589-597

Jolly CJ, Glenn G, Rahman S (1996) *Gsp-1* genes are linked to the grain hardness locus (*Ha*) on wheat chromosome 5D. Proc Natl Acad Sci USA 93: 2408-2413

Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19: 68-72

Juhasz A, Larroque OR, Tamas L, Hsam SLK, Zeller FJ, Bekes F, Bedo Z (2003) Bankuti 1201-an old Hungarian wheat variety with special storage protein composition. Theor Appl Genet 107: 697-704

Juretic N, Hoen DR, Huynh ML, Harrison PM, Bureau TE (2005) The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. Genome Res 15: 1292-1297

Jurka J (2000) Repbase update: A database and an electronic journal of repetitive elements. Trends Genet 16: 418-420

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update a database of eukaryotic repetitive elements. Cytogenet Genome Res 110 (1-4): 462-467

Kalendar R, Grob T, Regina M, Suoniemi A, Schulman AH (1999) IRAP and REMAP: two new retrotransposon-based DNA fingerprinting techniques. Theor Appl Genet 98: 704-711

Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH (2000) Genome evolution of wild barley (*Hordeum spontaneum*) by BARE1 retrotransposon dynamics in response to sharp microclimate divergence. Proc Natl Acd Sci USA 97: 6603-6607

Kashkush K, Feldman M, Levy AV (2003) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat Genet 33: 102-106

Kazazian HH Jr (2004) Mobile elements: drivers of genome evolution. Science 303: 1626-1632

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

Kellogg EA (2001) Evolutionary history of the grasses. Plant Physiol 125: 1198-1205

Kidwell MG (2002) Transposable elements and the evolution of genome size in eukaryotes. Genetica 115: 49-63

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16: 111-120

Kong XY, Gu YQ, You FM, Dubcovsky J, Anderson OD (2004) Dynamics of the evolution of orthologous and paralogous portions of a complex locus region in two genomes of allopolyploid wheat. Plant Mol Biol 54: 55-69

Krishnamurthy K, Giroux MJ (2001) Expression of wheat puroindoline genes in transgenic rice enhances grain softness. Nature Biotech 19: 162-166

Krishnamoorthy K, Balconi C, Sherwood JE, Giroux MJ (2001) Wheat puroindoline enhance fungal disease resistance in transgenic rice. Mol Plant Microbe Interact 14: 1255-1260

Kronmiller BA, Wise RP (2008) TEnest: automated chronological annotation and visualization of nested plant transposable elements. Plant Physiol 146: 45-59

Kubis S, Schmidt T, Heslop-Harrison JS (1998) Repetitive DNA elements as a major component of plant genomes. Ann Bot 82 (supplement A): 45-55

Kukuchi K, Terauchi K, Wada M, Hirano HY (2003) The plant MITE *mPing* is mobilized in anther culture. Nature 421: 167-170

Kumar A, Bennetzen JL (1999) Plant retrotransposons. Ann Rev Genet 33: 479-532
Lai J, Li Y, Messing J, Dooner HK (2005) Gene movement by helitron transposons contributes to the haplotype variability of maize. Proc Natl Acad Sci USA 102: 9068-9073

Lall IPS, Maneesha, Upadhyaya KC (2002) Panzee, a *copia-* like retrotransposon from the grain legume, pigeonpea (*Cajanus cajan* L.). Mol Genet Genomics 267: 271-280

Law CN, Young CF, Brown JWS, Snape JW, Worland AJ (1978) The study of grain protein control in wheat using whole chromosome substitution lines. In: Seed protein improvement by nuclear techniques. International Atomic Energy Agency, Vienna, Austria, pp 483-502

Levy AA, Feldman M (2004) Genetic and epigenetic reprogramming of the wheat genome upon allopolyploidization. Biol J Linnaean Soc 82: 607-613

Li W, Zhang P, Fellers JP, Friebe B, Gill BS (2004) Sequence composition, organization, and evolution of the core *Triticeae* genome. Plant J 40: 500–511

Li QY, Yan YM, Wang AL, Zhang YZ, Hsam SLK, Zeller FJ (2006) Detection of HMW glutenin subunit variation among 205 cultivated emmer accessions (*T. turgidum* ssp. *dicoccum*). Pl Breeding 125: 120-124

Li W, Huang L, Gill BS (2008) Recurrent deletions of puroindoline genes at the grain hardness locus in four independent lineages of polyploid wheat. Plant Physiol 146: 200-212

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

Lippman ZG, Black AV, Vaughn MW, Dedhia N, McCombie RW (2004) Role of transposable elements in heterochromatin and epigenetic control. Nature 430: 471-476

Liu B, Wendel JF (2000) Retrotransposon activation followed by rapid repression in introgressed rice plants. Genome 43: 874-880

Lönnig WE, Saedler H (2002) Chromosome rearrangements and transposable elements. Annu Rev Genet 36: 389-410

Lukow OM, Payne PI, Tkachuk R (1989) The HMW Glutenin subunit composition of Canadian wheat cultivars and their association with bread-making quality. J Sci Food Agric 46: 451-260

Lukow OM, Forsyth SA, Payne PI (1992) Over-production of HMW glutenin subunits coded on chromosome 1B in common wheat, *Triticum aestivum*. J Genet Breed 46: 187-192

Lukow OM, Preston KR, Watts BM, Malcolmson LJ, Cloutier S (2002) Measuring the influence of wheat protein in bread making: From damage control to genetic manipulation of protein composition in wheat. In: Ng PKW, Wrigley CW (Eds) Wheat quality elucidation-The Bushuk legacy. First edn, American Association of Cereal Chemists, Inc., St Paul,USA, pp 50-64

Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labelling kit and sizing of restriction fragments by capillary electrophoresis. Genomics 82: 378-389

Lynch M (2007) Mobile genetic elements. In: The origins of genome architecture. Sinauer Associates, Inc., Sunderland, USA, pp 151-191

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290: 1151-1155

Ma J, Bennetzen JL (2006) Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. Proc Natl Acad Sci USA 103: 383-388

Ma W, Zhang W, Gale KR (2003) Multiplex-PCR typing of high molecular weight glutenin subunit alleles in wheat. Euphytica 134: 51-60

Ma J, Devos KM, Bennetzen JL (2004) Analysis of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res 14: 860-869

Mackie AM, Sharp PJ, Lagudah ES (1996) The nucleotide and derived amino acid sequence of a HMW Glutenin gene from *Triticum tauschii* and comparison with those from the D genome of bread wheat. J Cereal Sci 24: 73-78

Madlung A, Comai L (2004) The effect of stress on genome regulation and structure. Ann Bot 94: 481-495

Marchylo BA, Lukow OM, Kruger JE (1992) Quantitative variation in high molecular weight glutenin subunit 7 in some Canadian wheats. J Cereal Sci 15: 29-37

Marillonnet S, Wessler SR (1997) Retrotransposon insertion into the maize waxy gene results in tissue specific RNA processing. Plant Cell 9: 967-78

Marillonnet S, Wessler SR (1998) Extreme structural heterogeneity among the members of a maize retrotransposon family. Genetics 150: 1245-1256

Marino-Ramirez L, Lewis KC, Landsman D, Jordan IK (2005) Transposable elements donate lineage specific regulatory sequences to host genomes. Cytogenet Genome Res 110: 333-341

Martin JM, Meyer FD, Smidansky ED, Wanjugi H, Blechl AE, Giroux MJ (2006) Complementation of the pina (null) allele with the wild type *Pina* sequence restores a soft phenotype in transgenic wheat. Theor Appl Genet 113: 1563-1570.

Massa AN, Morris CF (2006) Molecular evolution of the puroindoline-a, puroindoline-b and grain softness protein-1 genes in the tribe Triticeae. J Mol Evol 63: 526-536

Massa AN, Morris CF, Gill BS (2004) Sequence diversity of puroindoline-a, puroindoline-b, and the grain softness protein genes in *Aegilops tauschii* Coss. Crop Sci 44: 1808-1816

Matsuoka Y, Tsunewaki K (1996) Wheat retrotransposon families identified by reverse transcriptase domain analysis. Mol Biol Evol 13: 1384-1392

Matsuoka Y, Tsunewaki K (1999) Evolutionary dynamics of Ty1-*copia* group retrotransposons in grass shown by reverse transcriptase domain analysis. Mol Biol Evol 16: 208-21

Mattern JM, Morris R, Schmidt JW, Johnson VA (1973) Location of genes for kernel properties in wheat cultivar Cheyenne using chromosome substitution lines. In: Sears, E. R., Sears, L. M. S. (Eds) Proceedings of the 4th International Wheat Genetics Symposium, University of Missouri, Columbia, MO,USA, pp 703-707

McCarthy EM, Liu J, Lizhi G, McDonald JF (2002) Long terminal repeat retrotransposons of *Oryza sativa.* Genome Bio 3: research0053.1-0053.11

McClintock B (1950) The origin and behaviour of mutable loci in maize. Proc Natl Acad Sci USA 36: 344-355

McClintock B (1956) Controlling elements and the gene. Cold Spring Harb Symp Quant Biol 21: 197-216

McClintock B (1984) The significance of responses of the genome to challenge. Science 225: 792-801

McClintock B (1987) The discovery and characterization of transposable elements: The collected papers of Barbara McClintock. Garland Publishing Inc., New York, 636p

Moolhuijzen P, Dunn DS, Bellgard M, Carter M, Jia J, Kong X, Gill BS, Feuillet C, Breen J, Appels R (2007) Wheat genome structure and function: genome sequence data and the International wheat genome sequencing consortium. Aust J Agri Res 58: 470-475

Moore G, Devos KM, Wang Z, Gale MD (1995) Grasses line up and form a circle. Curr Biol 5: 737-739

Morgante M (2006) Plant Genome organization and diversity: the year of the junk. Current Opin Biotech 17: 168-173

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

Morris CF (2002) Puroindolines: the molecular genetic basis of wheat grain hardness. Plant Mol Biol 48: 633-647

Morris CF, Lukow OM, Perron CE (1998) Grain hardness, dough mixing and pan bread performance among wheats differing in puroindoline hardness mutation. Cereal Foods World 43: 533

Morris CF, Lillemo M, Simeone MC, Giroux MJ, Babb SL, Kidwell KK (2001) Prevalence of puroindoline grain hardness genotypes among historically significant North American spring and winter wheats. Crop Sci 41: 218-228

Morrison WR, Law CN, Wylie LJ, Coventry AM, Seekings J (1989) The effect of group 5 chromosomes on the free polar lipids and breadmaking quality of wheat. J Cereal Sci 9: 41-51

Morrison WR, Greenwell P, Law CN, Sulaiman BD (1992) Occurrence of friabilin, a low molecular weight protein associated with grain softness, on starch granules isolated from some wheats and related species. J Cereal Sci 15: 143-149

Muehlbauer GJ, Bhau BS, Syed NH, Heinen S, Cho S, Marshall D, Pateyron S, Buisine N, Chalhoub B, Flavell AJ (2006) A *hAT* superfamily transposase recruited by the cereal genome. Mol Genet Genomics 275: 553-563

Naeem HA, Sapirstein HD (2007) Ultra-fast separation of wheat glutenin subunits by reversed-phase HPLC using a superficially porous silica-based column. J Cereal Sci 46: 157-168

Nilmalgoda S D, Cloutier S, Walichnowski AZ (2003) Construction and characterization of a bacterial artificial chromosome (BAC) library of hexaploid wheat (*Triticum aestivum* L.) and validation of genome coverage using locus-specific primers. Genome 46: 870-878

Ohno S (1972) So much 'junk' DNA in our genome. In: Smith H H. (Ed) Evolution of Genetic Systems. Brookhaven Symposium on Biology, Gordon and Breach, NY, USA, Vol 26, pp 366-370

Orgel LE and Crick FHC (1980) Selfish DNA: The ultimate parasite. Nature 284: 604-607

Pardue ML, Rashkova S, Casacuberta E, DeBaryshe PG, George JA, Traverse KL (2005) Two retrotransposons maintain telomeres in *Drosophila*. Chromosome Res 13: 443-453

Paterson AH, Bowers JE, Paterson DG, Estill JC, Chapman BA (2003) Structure and evolution of cereal genomes. Curr Opin Genet Dev 13: 644-650

Payne PI (1987) Genetics of wheat storage proteins and the effect of allelic variation on bread making quality. Ann Rev Plant Physiol 38: 141-153

Payne PI, Lawrence GJ (1983) Catalogue of alleles for the complex gene loci,*Glu-A1,Glu-B1* and *Glu-D1* which code for high-molecular-weight subunits of glutenin in hexaploid wheat. Cereal Res Commun 11: 29-35

Payne PI, Law CN, Mudd E (1980) Control by homoeologous group 1 chromosomes of the high molecular weight subunits of glutenin, a major protein of wheat endosperm. Theor Appl Genet 58: 113-120

Payne PI, Holt LM, Law CN (1981) Structural and genetical studies on the high-molecular-weight subunits of wheat glutenin. Theor Appl Genet 60: 229-236

Payne PI, Holt LM, Jackson EA, Law CN (1984) Wheat storage proteins: their genetics and their potential for manipulation by plant breeding. Philos Trans R Soc London 304: 359-371

Payne PI, Nightingale MA, Krattiger AF, Holt LM (1987) The relationship between HMW glutenin subunit composition and the bread-making quality of British-grown wheat varieties. J Sci Food Agric 40: 51-65

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, Panaud O (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis,* a wild relative of rice. Genome Res 16: 1262-1269

Pomeranz Y, Williams PC (1990) Wheat hardness: its genetic, structural and biochemical background, measurements and signifcance. In: Pomeranz, Y. (Ed) Advances in cereal science and technology. American Association of Cereal Chemists, St Paul, Minnesota, USA, Vol 10, pp 471-544

Puig M, Caceres M, Ruiz A (2004) Silencing of a gene adjacent to the breakpoint of a widespread Drosophila inversion by a transposon induced antisense RNA. Proc Natl Acad Sci USA 101: 9013-9018

Qi L, Echalier B, Friebe B, Gill BS (2003) Molecular characterization of a set of wheat deletion stocks for use in chromosome bin mapping of ESTs. Funct Integr Genomics 3: 39-55

Qi LL, Echalier B, Chao S, Lazo GR, Butler GE, Anderson OD, Akhunov ED, Dvorak J, Linkiewicz AM, Ratnasri A, Dubcovsky J, Bermudez-Kandianis CE, Greene RA, Kantety R, La Rota CM, Munkvold JD, Sorrels SF, Sorells ME, Dilbirligi M, Sidhu D, Erayman M, Randhawa HS, Sandhu D, Bondareva SN, Gill KS, Mahmoud AA, Ma X, Miftahudin GJP, Conley EJ, Nduati V, Gonzalez-Hernandez JL, Anderson JA, Peng JH, Lapitan NLV, Hossain KG, Kalavacharla V, Kianian SF, Pathan MS, Zhang DS, Nguyen HT, Choi D, Fenton RD, Close TJ, McGuire PE, Qualset CO, Gill BS (2004) A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. Genetics 168: 701-712

Queen RA, Gribbon BM, James C, Jack P, Flavell AJ (2004) Retrotransposon-based molecular markers for linkage and genetics diversity analysis in wheat. Mol Genet Genomics 271: 91-97

Radovanovic N, Cloutier S (2003) Gene-assisted selection for high molecular weight glutenin subunits in wheat doubled haploid breeding programs. Mol Breeding 12: 51-59

Radovanovic N, Cloutier S, Brown D, Humphreys DG, Lukow OM (2002) Genetic variance for gluten strength contributed by high molecular weight glutenin proteins. Cereal Chem 79: 843-849

Ragupathy R, Naeem HA, Reimer E, Lukow OM, Sapirstein HD, Cloutier S (2008) Evolutionary origin of the segmental duplication encompassing the wheat *GLU-B1* locus encoding the overexpressed Bx7 (Bx7^{OE}) high molecular weight glutenin subunit. Theor Appl Genet 116: 283-296

Rahman S, Jolly JC, Skerritt JH, Wallosheck A (1994) Cloning of a wheat 15-kD grain softness protein: GSP is a mixture of puroindoline-like polypeptides. Eur J Biochem 223: 917-925

Rakszegi M, Bekes F, Lang L, Tamas L, Shewry PR, Bedo Z (2005) Technological quality of transgenic wheat expressing an increased amount of HMW glutenin subunit. J Cereal Sci 42: 15-23

Ramallo E, Kalendar R, Schulman AH, Martínez-Izquierdo JA (2008) Reme1, a *copia* retrotransposon in melon, is transcriptionally induced by UV light. Plant Mol Biol 66: 137-150

Röder MS, Korzun V Wendehake K, Plaschke J, Tixier MH, Leroy P, Ganal MW (1998) A microsatellite map of wheat. Genetics 149: 2007-2023

Sabot F, Simon D, Bernard M (2004) Plant transposable elements, with an emphasis on grass species. Euphytica 139: 227-247

Sabot F, Sourdille P, Chantret N, Bernard M (2006) Morgane, a new LTR retrotransposon group, and its subfamilies in wheat. Genetica 128: 439-447

Saitou N, Nei M (1987) The neighbour-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425

Sakata K, Nagamura Y, Numa H, Antonio BA, Nagasaki H, Idonuma A, Watanabe W, Shimizu Y, Horiuchi I, Matsumoto T, Sasaki T, Higo K (2002) RiceGAAS: an automated annotation system and database for rice genome sequence. Nucleic Acids Res 30: 98-102

Salamini FH, Ozkan A, Brandolini R, Schafer-Pregl Martin W (2002) Genetics and geography of wild cereal domestication in the near east. Nat Rev Genet 30: 429-441

Sandhu D, Gill KS (2002) Gene-containing regions of the wheat and the other genomes. Plant Physiol 128: 803-811

SanMiguel P, Bennetzen JL (1998) Evidence that a recent increase in Maize genome size was caused by the massive amplification of intergene retrotransposons. Ann Bot 82: 37-44

SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274: 765-768.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20: 43-45

SanMiguel P, Ramakrishna W, Bennetzen JL, Busso CS, Dubcovsky J (2002) Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome $5A^{m}$. Funct Integr Genomics 2: 70-80

Scherrer B, Isidore E, Klein P, Kim JS, Belleb A, Chalhoub B, Keller B, Feuillet C (2005) Large intraspecific haplotype variability at the *Rph7* locus results from rapid and recent divergence in the barley genome. Plant Cell 17: 361-374

Schuelke M (2000) An economic method for the fluorescent labelling of PCR fragments. Nat Biotechnol 18: 233–234

Schulman AH, (2007) Molecular markers to assess genetic diversity. Euphytica 158: 313-321

Schulman A H, Kalendar RA (2005) A movable feast: diverse retrotransposons and their contribution to barley genome dynamics. Cytogenet Genome Res 110: 598-605

See DR, Giroux M, Gill BS (2004) Effect of multiple copies of puroindoline gene on grain softness. Crop Sci 44: 1248-1253

Shewry PR, Halford NG, Tatham AS (1992) High molecular weight subunits of wheat glutenin. J Cereal Sci 15: 105-120

Shewry PR, Gilbert SM, Savage AWJ, Tatham AS, Wan YF, Belton PS, Wellner N, D'Ovidio R, Bekes F, Halford NG (2003) Sequence and properties of HMW subunit 1Bx20 from pasta wheat (*Triticum durum*) which is associated with poor end use properties. Theor Appl Genet 106: 744-750

Shewry PR, Halford NG, Lafiandra D (2006) The high molecular weight subunits of glutenin. In: Wrigley C, Bekes F, Bushuk W (Eds) Gliadin and Glutenin-The unique balance of wheat quality. AACC international, St Paul, Minnesota, USA, pp 143-169

Sidhu D, Gill KS (2004) Distribution of genes and recombination in wheat and other eukaryotes. Plant Cell Tissue Organ Cult 79: 257-270

Simeone M, Lafiandra D (2005) Isolation and characterization of friabilins genes in rye. J Cereal Sci 41: 115-122

Simeone M, Gedye KR, Mason-Gamer R, Gill BS, Morris CF (2006) Conserved regulator elements identified from a comparative puroindoline gene sequence survey of *Triticum* and *Aegilops* diploid taxa. J Cereal Sci 44: 21-33

Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 8: 272-285

Smith DB, Flavell RB (1975) Characterization of wheat genome by reassociation kinetics. Chromosoma 50: 223-242

Soderlund C, Longden I, Mott R (1997) FPC: a system for building contigs from restriction fingerprinted clones. Bioinformatics 13: 523-535

Soltis DE, Soltis PL (1999) Polyploidy: recurrent formation and genome evolution. Trends Ecol Evol 14: 348-351

Somers DJ, Isaac P, Edwards K (2004) A high density microsatellite map for bread wheat (*Triticum aestivum* L.). Theor Appl Genet 109: 1105-1114

Song R, Llaca V, Messing J (2002) Mosaic organization of orthologous sequences in grass genomes. Genome Res 12: 1549-1555

Sonnhammer ELL, Durbin R (1996) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167: 1-10

Sorrells ME (2007) Application of new knowledge, technologies and strategies to wheat improvement. Euphytica 157: 299-306

Sorrells ME, Rota ML, Bermudez-Kandianis CE, Greene RA, Kantety R, Munkvold JD, Miftahudin MA, Ma X, Gustafson PJ, Qi LL, Echalier B, Gill BS, Matthews DE, Lazo GR, Chao S, Anderson OD, Edwards H, Linkiewicz AM, Dubcovsky J, Akhunov ED, Dvorak J, Zhang D, Nguyen HT, Peng J, Lapitan NLV, Gonzelez-Hernandez JL, Anderson JA, Hossain K, Kalavacharla V, Kianian SF, Choi D, Close TJ, Dilbirligi M, Gill KS, Steber C, Walker-Simmons MK, Mcguire PE, Qualset CO (2003) Comparative DNA sequence analysis of Wheat and Rice genomes. Genome Res 13: 1818-1827

Sourdille P, Perretant MR, Charmet G, Leroy P, Gautier MF, Joudrier P, Nelson JC, Sorrells ME, Bernard M (1996) Linkage between RFLP markers and genes affecting kernel hardness in wheat. Theor Appl Genet 93: 580-586

Stein N (2007) Triticeae genomics: advances in sequence analysis of large genome cereal crops. Chromosome Res 15: 21-31

Sugiyama T, Rafalski A, Peterson D, Soll D (1985) A wheat HMW glutenin subunit gene reveals a highly repeated structure. Nucleic Acids Res 13: 8729-8737

Suoniemi A, Anamthawat-Jonsson K, Arna T, Schulman AH (1996) Retrotransposon *BARE*-1 is a major dispersed component of the barley (*Hordeum vulgare* L.) genome. Plant Mol Biol 30: 1321-1329

Sutton KH (1991) Quantitative and qualitative variation among high molecular weight subunits of glutenin detected by reversed phase high performance liquid chromatography. J Cereal Sci 14: 25-34

Swan CG, Bowman JGP, Martin JM, Giroux MJ (2006) Increased puroindoline levels slow ruminal digestion of wheat (*Triticum aestivum* L.) starch by cattle. J Anim Sci 84: 641-650

Syed NH, Flavell AJ (2006) Sequence-specific amplification polymorphisms (SSAPs): a multi-locus approach for analyzing transposon insertions. Nature Protocols 1: 2746-2752

Takeda S, Sugimoto K, Otsuki H, Hirochika H (1998) Transcriptional activation of the tobacco retrotransposon *Tto1* by wounding and methyl jasmonate. Plant Mol Biol 36: 365-376

Talbert LE, Smith LY, Blake NK (1998) More than one origin of hexaploid wheat is indicated by sequence comparison of low copy DNA. Genome 41: 402-407

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596-1599

Tanchak MA, Schernthaner JP, Giband M, Altosaar I (1998) Tryptophanins: isolation and molecular characterization of oat cDNA clones encoding proteins structurally related to puroindoline and wheat grain softness proteins. Plant Sci 137-184

Tang YM, You-Zhi MA, Li LC, Xing-Guo YE (2005) Identification and characterization of reverse transcriptase domain of transcriptionally active retrotransposons in wheat genomes. J Integrative Plant Biol 47: 604-612

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH (2008) Synteny and collinearity in plant genomes. Science 320: 486-488

Tatusova TA, Madden TL (1999) BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 174: 247-250

Thompson RD, Bartels D, Harberd NP (1985) Nucleotide sequence of a gene from chromosome ID of wheat encoding a HMW glutenin subunit. Nucleic Acids Res 13: 6833-6846

Tipples KH, Kilborn RH, Preston KR (1994) Bread wheat quality defined. In: Wheat: production, properties and quality. Bushuk W, Rasper VF (Eds) Chapman and Hall, Glasgow, UK, pp 25-35

Tranquilli G, Lijavetzky D, Muzzi G, and Dubcovsky J (1999) Genetic and physical characterization of grain texture related loci in diploid wheat. Mol Gen Genet 262: 846-850

Turnbull KM, Rahman S (2002) Endosperm texture in wheat. J. Cereal Sci 36: 327-337

Turnbull KM, Turner M, Mukai Y, Yamamoto M, Morell MK, Appels R, Rahman S (2003) The organization of genes tightly linked to the *Ha* locus in *Aegilops tauschii*, the D genome donor to wheat. Genome 46: 330-338

Varagona MJ, Purugganan M, Wessler SR (1992) Alternative splicing induced by insertion of retrotransposons into maize waxy gene. Plant Cell 4: 811-820

Vawser MJ, Cornish GB (2004) Overexpression of HMW glutenin subunit *Glu-B1* 7x in hexaploid wheat varieties (*Triticum aestivum* L.). Aust J Agri Res 55: 577-588

Vicient CM, Jääskeläinen MJ, Kalendar R, Schulman AH. (2001) Active retrotransposons are a common feature of grass genomes. Plant Physiol 125: 1283-1292

Vicient CM, Kalendar R, Schulman AH (2005) Variability, recombination, and mosaic evolution of the barley bare-1 retrotransposon. J Mol Evol 61: 275-291.

Vision TJ, Brown DG, Tanksley SD (2000) The origin of genomic duplications in Arabidopsis. Science 290: 2114-2117

Vitte C, Bennetzen JL (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci USA 103: 17638-17643

Vitte C, Panaud O (2003) Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice *Oryza sativa* L. Mol Biol Evol 20(4): 528-540

Vitte C, Panaud O (2005) LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenet Genome Res 110: 91-107

Vitte C, Ishii T, Lamy F, Brar DS, and Panaud O (2004). Genomic paleontology provides evidence for two distinct origins of Asian rice (*Oryza sativa L.*). Mol Genet Genomics 272: 504-511

Volff JN (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays 28: 913-922

Vongs A, Kakutani T, Martienssen RA, and Richards EJ (1993) *Arabidopsis thaliana* DNA methylation mutants. Science 260: 1926-1928

Voytas DF, Cummings MP, Konieczny AK, Ausubel FM, Rodermel SR (1992). *Copia-* like retrotransposons are ubiquitous among plants. Proc Natl Acad Sci USA 89: 7124-7128

Wang Q, Dooner HK (2006) Remarkable variation in maize genome structure inferred from haplotype diversity at the *bz* locus. Proc Natl Acad Sci USA 103: 17644-17649

Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S, Lu Z, Wong GK, Long M, Wang J (2006) High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell 18: 1791-1802

Weegels PL, Van de Pijpekamp AM, Graveland A, Hamer RJ, Schofield JD (1996) Depolymerisation and re-polymerisation of wheat glutenin during dough processing. I. Relationships between glutenin macropolymer content and quality parameters. J Cereal Sci 23: 103-111

Wessler SR (1996) Plant retrotransposons: Turned on by stress. Curr Biol 6: 959-961

Wessler SR (2006a) Transposable elements and the evolution of eukaryotic genomes. Proc Natl Acad Sci USA 103: 17600-17601

Wessler SR (2006b) Eukaryotic transposable elements: Teaching old genomes new tricks. In: Lynn Helena Caporale (Ed) The implicit Genome. Oxford University press, NY, USA, pp138-162

White SE, Habera LF, Wessler SR (1994) Retrotransposons in the flanking regions of normal plant genes: A role for *copia*-like elements in the evolution of gene structure and expression. Proc Natl Acad Sci USA 91: 11792-11796

Wicker T, Keller B (2007) Genome-wide comparative analysis of *copia* retrotransposons in *Triticeae*, rice and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. Genome Res 17: 1072-1081

Wicker T, Stein J, Albar L, Feuillet C, Schlagenhauf E, Keller B (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. Plant J 26: 307-316

Wicker T, Matthews DE, Keller B (2002) TREP: A database for Triticeae repetitive elements. Trends Plant Sci 7: 561-562

Wicker T, Yahiaoui N, Guyot R, Schlagenhauf E, Liu ZD, Dobcovsky J (2003) Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A^m genomes of wheat. Plant Cell 15: 1186-1197

Wicker T, Zimmermann W, Perovic D, Paterson AH, Ganal M, Graner Stein N (2005) A detailed look at 7 million years of genome evolution in a 439 kb contiguous sequence at the barley *Hv-elF4E* locus: recombination, rearrangements and repeats. Plant J 41: 184-194

Wicker T, Yahiaoui N, Keller B (2007a) Contrasting rates of evolution in *Pm3* loci from three wheat species and rice. Genetics 177: 1207-1216

Wicker T, Sabot F, Huna-Van A, Bennetzen JL, Copy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007b) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8: 973-982

Wolfe KH (2001) Yesterday's polyploids and the mystery of diploidization. Nature Rev Genet 2: 333-341

Wrigley CW (1996) Giant proteins with flour power. Nature 381: 738-739

Xiao H, Jiang N, Schaffner E, Stockinger EJ, Van der Knaap E (2008) A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Science 319: 1527-1530

Xu Z, Wang H (2007) LTR_ FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35: W265-W268

Yan L, Loukoianov A, Tranquilli G, Helguera M, Fahima T, Dubcovsky J (2003) Positional cloning of the wheat vernalization gene *VRN1*. Proc Natl Acad Sci USA 100: 6263-6268.

Yan L, Loukoianov A, Blechl A, Tranquilli G, Ramakrishna W, SanMiguel P, Bennetzen JL, Echenique V, Dubcovsky J (2004) The wheat *VRN2* gene is a flowering repressor down-regulated by vernalization. Science 303: 1640-1644

Yan Y, Zheng J, Xiao Y, Yu J, Hu Y, Cai M, Li Y, Hsam SLK, Zeller FJ (2004) Identification and molecular characterization of a novel y-type *Glu-D^{t}1* glutenin gene of *Aegilops tauschii.* Theo Appl Genet 108: 1349–1358

Yandeau-Nelson MD, Xia Y, Li J, Neuffer MG, Schnable PS (2006) Unequal sister chromatid and homolog recombination at a tandem duplication of the a1 locus in maize. Genetics 173: 2211-2226

Zhang J (2003) Evolution by gene duplication: an update. Trends Eco Evol 18: 292-298

APPENDICES

Appendix I. High molecular weight glutenin subunits (HMW-GS) protein extraction and SDS-PAGE analysis (Radovanovic and Cloutier 2003)

Seeds were first crushed in a mortar and pestle and then an appropriate amount of 1x extraction buffer calculated by multiplying the weight of the sample in mg by 0.02 to give the volume in ml was added (3x extraction buffer contained 20% Glycerol, 0.02% Pyronin Y, 0.125 M Tris-Cl, pH 6.8 and 4% SDS). β-mercaptoethanol to a final concentration of 8% , was added and samples were vortexed several times over an hour at room temperature. The samples were then incubated at 95°C for 2.5 minutes and placed in -20°C storage until use. A vertical electrophoresis unit (Hoefer Scientific, Inc, Holliston, MA, USA) was used to conduct gel electrophoresis on 8 μl sub samples in 12% SDS-PAGE gel for approximately 3 hours at 40 mA in a buffer containing 0.1% SDS, 0.192M Glycine and 0.025 M Tris. Gels were stained with 1% Coomassie Blue in ethanol and Blakley's stain for 1.5 days. After staining, gels were soaked in 5% glycerol for about 4 hours and dried at room temperature for several days.

Appendix II. Sample preparation and RP-HPLC analysis of HMW-GS (Hamid and Sapirstein 2007)

A. Sample preparation

Around 50 mg of crushed endosperm sample was washed twice with 500 μl of 50% 1-propanol for 15 minutes at room temperature, followed by centrifugation for 3 minutes at 15,000g. These wash solutions containing mainly gliadins, were discarded. The residue was again washed with 500 μl of 50% 1-propanol to remove any remaining soluble proteins and glutenin subunits were extracted with 150 μl extraction buffer (0.08 M Tris–HCl containing 50% 1-propanol, pH 7.5) in the presence of DTT (1% w/v) for 30 minutes at 60°C. The extract was then alkylated with 150 μl of buffer containing 4-vinylpyridine (4%) at 60°C for 30 minutes, vortexing every 6 minutes. After centrifugation at 15,000*g* for three minutes 1.0 μl aliquots of supernatant were transferred to auto-sampler vials.

B. RP-HPLC analysis of glutenin subunits

An Agilent HPLC 1100 system with a Poroshell 300SB-C_8 column (5 μm particle size, 300 A° pore size, 2.1×75 mm), a binary solvent delivery system, online vacuum degasser and diode array detector with a 6 mm path length, and 1.7 μl flow cell (Agilent Technologies Inc., Wilmington, DE, USA) was used. The Poroshell column was preceded by a guard column-Poroshell 300SB-C_8 2.1×12.5 mm, 5 μm. Deionized water (solvent A) and Acetonitrile (solvent B) each containing 0.1% (v/v) TFA were used for the elution of glutenin subunits. Detector wavelength was set to 206 nm. Agilent

ChemStation software (version 10.01) was used for integration and quantitative analysis of chromatograms using a peak width response time of 0.05 minutes.

Appendix III. List of *Triticum* accessions surveyed for the duplication at the *Glu-B1* locus

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
1	AS11527	*Ae. speltoides*	PGRC	Czech	-	-	n/a	-
2	AS11585	*Ae. speltoides*	PGRC	Canada	-	-	n/a	-
3	CN11649	*T. monococcum* subsp *monococcum*	PGRC	unknown	-	-	n/a	-
4	CN37617	*T. monococcum* subsp *monococcum*	PGRC	unknown	-	-	n/a	-
5	CN12440	*T. monococcum* subsp *monococcum*	PGRC	unknown	-	-	n/a	-
6	CN37655	*T. monococcum* subsp *monococcum*	PGRC	unknown	-	-	n/a	-
7	CN11756	*T. monococcum* subsp *monococcum*	PGRC	unknown	-	-	n/a	-
8	CWI19490	*T. monococcum* RL5444	CIMMYT	unknown	-	-	n/a	-
9	BW31215	*Ae. squarrosa*	CIMMYT	unknown	-	-	n/a	-
10	BW31229	*Ae. squarrosa*	CIMMYT	unknown	-	-	n/a	-
11	BW31242	*Ae. squarrosa*	CIMMYT	unknown	-	-	n/a	-
12	BW31236	*Ae. squarrosa*	CIMMYT	unknown	-	-	n/a	-
13	BW31260	*Ae. squarrosa*	CIMMYT	unknown	-	-	n/a	-
14	BW31246	*Ae. squarrosa*	CIMMYT	unknown	-	-	n/a	-
15	CN30817	*Ae. tauschii*	PGRC	unknown	-	-	n/a	-
16	CN30818	*Ae. tauschii*	PGRC	unknown	-	-	n/a	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
17	CN30939	*Ae. tauschii*	PGRC	unknown	-	-	n/a	-
18	CN30948	*Ae. tauschii*	PGRC	unknown	-	-	n/a	-
19	CN40033	*Ae. tauschii*	PGRC	unknown	-	-	n/a	-
20	CN12223	*T. turgidum* var 0-10-758	PGRC	Czech	7	-	7	-
21	CN32158	*T. turgidum* var Alaska	PGRC	USA	-	-	n/a	-
22	CN2644	*T. turgidum* var Branco	PGRC	Portugal	7OE	+	7	+
23	CN51254	*T. turgidum* var Melanopus	PGRC	Russia	7	-	7	-
24	CN1748	*T. turgidum* var Tetra Canthatch	PGRC	Canada	7	-	7*	-
25	DW7195	*T. turgidum* var IC 3042	CIMMYT	unknown	-	-		-
26	CN2677	*T. turgidum*	PGRC	Russia	-	-	n/a	-
27	CN12232	*T. turgidum*	PGRC	Czech	-	-	n/a	-
28	CN10545	*T. turgidum*	PGRC	Canada	-	-	n/a	-
29	CN10547	*T. turgidum*	PGRC	Canada	-	-	n/a	-
30	CN11002	*T. turgidum*	PGRC	Canada	-	-	n/a	-
31	CN11003	*T. turgidum*	PGRC	Canada	7	-	7	-
32	CN11573	*T. turgidum*	PGRC	Canada	-	-	n/a	-
33	CN11579	*T. turgidum*	PGRC	Canada	-	-	n/a	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
34	CN12222	*T. turgidum*	PGRC	Czech	7OE	+	7	+
35	CN12224	*T. turgidum*	PGRC	Canada	-	-	n/a	-
36	CN12225	*T. turgidum*	PGRC	Czech	7OE	+	7	+
37	CN12226	*T. turgidum*	PGRC	Czech	-	-	n/a	-
38	CN12227	*T. turgidum*	PGRC	Czech	7	+	7	-
39	CN12230	*T. turgidum*	PGRC	Czech	-	-	n/a	-
40	CN12233	*T. turgidum*	PGRC	Czech	-	-	n/a	-
41	CN51246	*T. turgidum*	PGRC	Canada	-	-	n/a	-
42	CN51253	*T. turgidum*	PGRC	Russia	7	-	7	-
43	CN51255	*T. turgidum*	PGRC	Russia	7	-	7	-
44	CN51256	*T. turgidum*	PGRC	Russia	7	-	7	-
45	CN51257	*T. turgidum*	PGRC	Russia	7	-	7	-
46	CN51258	*T. turgidum*	PGRC	Russia	7	-	7	-
47	CN51259	*T. turgidum*	PGRC	Russia	7	-	7	-
48	CN51263	*T. turgidum*	PGRC	Russia	7	-	7	-
49	CN51265	*T. turgidum*	PGRC	Russia	7	-	7	-
50	CN51269	*T. turgidum*	PGRC	USA	-	-	na	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
51	CN33858	*T. turgidum* subsp *turgidum* var Alaska	PGRC	unknown	-	-	n/a	-
52	CN11761	*T. turgidum* subsp *turgidum* var Gentile	PGRC	unknown	-	-	n/a	-
53	CN40848	*T. turgidum subsp turgidum*	PGRC	unknown	7	-	7	-
54	CN32492	*T. turgidum* subsp *dicoccoides*	PGRC	unknown	-	-	n/a	-
55	CN32493	*T. turgidum* subsp *dicoccoides*	PGRC	unknown	-	-	n/a	-
56	CN32494	*T. turgidum* subsp *dicoccoides*	PGRC	unknown	-	-	n/a	-
57	CN32495	*T. turgidum* subsp *dicoccoides*	PGRC	unknown	-	-	n/a	-
58	CN37613	*T. turgidum* subsp *dicoccoides*	PGRC	unknown	-	-	n/a	-
59	CWI18234	*T. dicoccoides* PI 355459	CIMMYT	unknown	-	-	n/a	-
60	CWI15345	*T. dicoccoides*	CIMMYT	unknown	-	-	n/a	-
61	CWI16984	*T. dicoccoides* PI 190919	CIMMYT	unknown	-	-	n/a	-
62	CWI17128	*T. dicoccoides* PI 256029	CIMMYT	unknown	-	-	n/a	-
63	CWI19112	*T. dicoccoides* PI467007	CIMMYT	unknown	-	-	n/a	-
64	CWI19119	*T. dicoccoides* PI 467033	CIMMYT	unknown	-	-	n/a	-
65	CWI18902	*T. dicoccoides* PI 428024	CIMMYT	unknown	-	-	n/a	-
66	-	*T. dicoccoides* 5310	Dr.Fedak	unknown	-	-	n/a	-
67	-	*T. dicoccoides* 5315	Dr.Fedak	unknown	7	-	7	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
68	CN7773	*T. turgidum* subsp *durum* var Abu Fashit	PGRC	unknown	-	-	n/a	-
69	CN7786	*T. turgidum* subsp *durum* var Sbei	PGRC	unknown	-	-	n/a	-
70	CN10163	*T. turgidum* subsp *durum* var Golden Ball	PGRC	unknown	-	-	n/a	-
71	CN32483	*T. turgidum* subsp *dicoccum* Var Early spelt	PGRC	unknown	-	-	n/a	-
72	CN32486	*T. turgidum* subsp *dicoccum*	PGRC	unknown	-	-	n/a	-
73	CN32488	*T. turgidum* subsp *dicoccum*	PGRC	unknown	-	-	n/a	-
74	CWI19153	*T. dicoccum* CI 8641	CIMMYT	unknown	-	-	n/a	-
75	CWI36567	*T. dicoccum* 500002	CIMMYT	unknown	-	-	n/a	-
76	CWI19155	*T. dicoccum* CI8643	CIMMYT	unknown	-	-	n/a	-
77	DW7193	*T. dicoccum* 1442	CIMMYT	unknown	-	-	n/a	-
78	DW7191	*T. dicoccum* 1409	CIMMYT	unknown	-	-	n/a	-
79	CN32496	*T. turgidum* subsp *carthlicum*	PGRC	unknown	7	-	7	-
80	CN32498	*T. turgidum* subsp *carthlicum*	PGRC	unknown	7	-	7	-
81	CN32500	*T. turgidum* subsp *carthlicum*	PGRC	unknown	7	-	7	-
82	CN32502	*T. turgidum* subsp *carthlicum*	PGRC	unknown	7	-	7	-
83	CN40686	*T. turgidum* subsp *carthlicum*	PGRC	unknown	-	-	n/a	-
84	DW7196	*T. carthlicum* IC 12180	CIMMYT	unknown	7	-	7	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
85	CWI5222	*T. carthlicum* TK 97-500	CIMMYT	unknown	7	-	7	-
86	CWI44150	*T. carthlicum* 001513	CIMMYT	unknown	7	-	7	-
87	CWI44453	*T. carthlicum* 020280	CIMMYT	unknown	7	-	7	-
88	CWI44562	*T. carthlicum* 042110	CIMMYT	unknown	7	-	7	-
89	CWI44563	*T. carthlicum* 042111	CIMMYT	unknown	7	-	7	-
90	CWI44464	*T. carthlicum* 020465	CIMMYT	unknown	7	-	7*	-
91	CN1839	*T. turgidum* subsp *turanicum*	PGRC	unknown	-	-	n/a	-
92	CN10543	*T. turgidum* subsp *turanicum*	PGRC	unknown	-	-	n/a	-
93	CN11587	*T. turgidum* subsp *turanicum*	PGRC	unknown	7	-	7*	-
94	CN51900	*T. turgidum* subsp *polonicum*	PGRC	unknown	-	-	n/a	-
95	CN51903	*T. turgidum* subsp *polonicum*	PGRC	unknown	-	-	n/a	-
96	CN51933	*T. turgidum* subsp *polonicum*	PGRC	unknown	-	-	n/a	-
97	CN1849	*T. aestivum* subsp *spelta*	PGRC	unknown	7	+	7	-
98	CN12229	*T. aestivum* subsp *spelta*	PGRC	unknown	7	+	7	-
99	CN12261	*T. aestivum* subsp *spelta*	PGRC	unknown	-	-	n/a	-
100	CN37599	*T. aestivum* subsp *spelta*	PGRC	unknown	-	-	n/a	-
101	CN2674	*T. aestivum* subsp *compactum* Var Indur compactum	PGRC	unknown	-	-	n/a	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
102	CN12213	*T. aestivum* subsp *compactum*	PGRC	unknown	7	-	7*	-
103	CN33802	*T. aestivum* subsp *compactum*	PGRC	unknown	7	-	7	-
104	CNI42787	*T. aestivum* subsp *compactum* var KVL 2263	CIMMYT	unknown	-	-	n/a	-
105	CN11647	*T. aestivum* subsp *spherococcum*	PGRC	unknown	-	-	n/a	-
106	CN11739	*T. aestivum* subsp *spherococcum*	PGRC	unknown	-	-	n/a	-
107	CN33803	*T. aestivum* subsp *spherococcum*	PGRC	unknown	7	-	7*	-
108	CN33892	*T. aestivum* subsp *spherococcum*	PGRC	unknown	-	-	n/a	-
109	CN33893	*T. aestivum* subsp *spherococcum*	PGRC	unknown	-	-	n/a	-
110	G3893	*T. spherococcum*	PGRC	unknown	-	-	n/a	-
111	CWI42988	*T. spherococcum*	PGRC	unknown	7	-	7*	-
112	CN99032	*T. aestivum* subsp *macha*	PGRC	unknown	-	-	n/a	-
113	CN33898	*T. aestivum* subsp *aestivum* var Hsin ShuKuan#1	PGRC	China	-	-	n/a	-
114	CN29735	*T. aestivum* var Fengcheung No 1	PGRC	China	7	-	7*	-
115	CN30572	*T. aestivum* var Beijing No 6	PGRC	China	7	-	7*	-
116	CN33902	*T. aestivum* var Ho chun No 12	PGRC	China	7	-	7	-
117	CN42881	*T. aestivum* var Chengdu Guangtou	PGRC	China	7	-	7	-
118	CN42882	*T. aestivum* var Mazha Mai	PGRC	China	7	-	7*	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
119	PI447402	*T. aestivum* var Cai Zi Huang	NSGC	China	7	-	7*	-
120	PI447405	*T. aestivum* var Fu Mai No 3	NSGC	China	7	-	7*	-
121	PI447403	*T. aestivum* var Wan Nian No 2	NSGC	China	7	-	7*	-
122	PI462151	*T. aestivum* var Shu Chou Wheat #3	NSGC	China	7	-	7*	-
123	PI531193	*T. aestivum* var JG1	NSGC	China	-	-	n/a	-
124	PI531188	*T. aestivum* var NING 7840	NSGC	China	7	-	7*	-
125	PI462141	*T. aestivum* var Li Yang Wong Shu Bai	NSGC	China	7	-	7*	-
126	PI481542	*T. aestivum* var Su Mai No 3	NSGC	China	7	-	7*	-
127	PI462149	*T. aestivum* var Nan Tong DA Huang PI	NSGC	China	7	-	7*	-
128	PI531191	*T. aestivum* var NING 8331	NSGC	China	7	-	7*	-
129	PI531189	*T. aestivum* var NING 8026	NSGC	China	7	-	7*	-
130	PI481544	*T. aestivum* var Xiang Mai No 1	NSGC	China	7	-	7*	-
131	PI447404	*T. aestivum* var Yang Mai No 1	NSGC	China	7	-	7*	-
132	CN11191	*T. aestivum* var New Pusa	PGRC	India	7	-	7*	-
133	CN44171	*T. aestivum* var Girija	PGRC	India	7	-	7*	-
134	CN10679	*T. aestivum* var Kenphad 25	PGRC	India	7	-	7*	-
135	PI337371	*T. aestivum* var Sonalika	NSGC	India	-	-	n/a	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
136	CN12036	*T. aestivum* var Tabasi	PGRC	Iran	7	-	7	-
137	CN11401	*T. aestivum* var Rayhany	PGRC	Iran	7	-	7	-
138	CN9577	*T. aestivum* var Azar	PGRC	Iran	7	-	7*	-
139	CN5904	*T. aestivum*	PGRC	Iran	7	-	7	-
140	CN11962	*T. aestivum* var Shahpassand	PGRC	Iran	7	-	7	-
141	CN9497	*T. aestivum* var Ajelea	PGRC	Iraq	-	-	n/a	-
142	-	*T aestivum* var TAA36	CRC	Israel	7OE	+	7	+
143	CN11206	*T. aestivum* var Norin 75	PGRC	Japan	7	-	7*	-
144	CN32076	*T. aestivum* var Nobeoka Bozu	PGRC	Japan	7	-	7	-
145	CN32077	*T. aestivum* var Nyu Bay	PGRC	Japan	7	-	7	-
146	CN44024	*T. aestivum* var Gogatsukomugai	PGRC	Japan	7	-	7*	-
147	CN44025	*T. aestivum* var Ikuzai #1	PGRC	Japan	-	-	n/a	-
148	CItr12699	*T. aestivum* var Norin10	PGRC	Japan	7	-	7	-
149	PI197130	*T. aestivum* var Sanshukomugi	NSGC	Japan	7	-	7	-
150	PI157584	*T. aestivum* var Seu Seun 27	NSGC	Japan	7	-	7*	-
151	PI197128	*T. aestivum* var Shinchunaga	NSGC	Japan	7	-	7	-
152	PI197129	*T. aestivum* var Shirasaya No1	NSGC	Japan	7	-	7	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
153	CN12209	*T. aestivum* var Dorziyeh Safra	PGRC	Jordan	-	-	n/a	-
154	CN11204	*T. aestivum*	PGRC	Jordan	-	-		-
155	CN42522	*T. aestivum*	PGRC	Nepal	7	+	7	-
156	CN29871	*T. aestivum*	PGRC	Nepal	-	-	n/a	-
157	CN29858	*T. aestivum*	PGRC	Nepal	-	-	n/a	-
158	CN29856	*T. aestivum*	PGRC	Nepal	7	-	7	-
159	CN29854	*T. aestivum*	PGRC	Nepal	7	-	7	-
160	CN9934	*T. aestivum*	PGRC	Pakistan	-	-	n/a	-
161	CN9936	*T. aestivum*	PGRC	Pakistan	-	-	n/a	-
162	CN10241	*T. aestivum*	PGRC	Pakistan	-	-	n/a	-
163	CN12251	*T. aestivum*	PGRC	Pakistan	-	-	n/a	-
164	CN30213	*T. aestivum*	PGRC	Pakistan	-	-	n/a	-
165	CN10000	*T. aestivum* var Dvina	PGRC	Russia	7	-	7*	-
166	CN10068	*T. aestivum* var Ferrugineum-87	PGRC	Russia	-	-	n/a	-
167	CN10167	*T. aestivum* var Golubka	PGRC	Russia	7	-	7*	-
168	CN10169	*T. aestivum* var Gorkorskaja-15	PGRC	Russia	7	-	7*	-
169	CN10529	*T. aestivum* var Irkutskaja 49	PGRC	Russia	7	-	7*	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
170	CN10620	*T. aestivum* var Jakutjanka	PGRC	Russia	7	-	7	-
171	CN10639	*T. aestivum* var Kamalinka	PGRC	Russia	7	-	7*	-
172	CN10644	*T. aestivum* var karagandinskaja	PGRC	Russia	7	-	7*	-
173	CN10868	*T. aestivum* var Klein Trou	PGRC	Russia	-	-	n/a	-
174	CN10898	*T. aestivum* var Koktunkulskaja 332	PGRC	Russia	7	-	7*	-
175	CN10902	*T. aestivum* var Kostoff's Triple hybrid	PGRC	Russia	7	-	7*	-
176	CN10905	*T. aestivum* var Krasnaja Svenzda	PGRC	Russia	7	-	7	-
177	CN10907	*T. aestivum* var Krasnojarskaja 1103	PGRC	Russia	7	-	7*	-
178	CN10908	*T. aestivum* var Krasnokutka 3	PGRC	Russia	7	-	7*	-
179	CN11132	*T. aestivum* var Minskaja	PGRC	Russia	7	-	7*	-
180	CN11141	*T. aestivum* var Moskovka	PGRC	Russia	7	-	7*	-
181	CN11239	*T. aestivum* var odess kaja 13	PGRC	Russia	7	-	7	-
182	CN11255	*T. aestivum* var onohoiskaja 4	PGRC	Russia	7	-	7*	-
183	CN11263	*T. aestivum* var Pamjat Urala	PGRC	Russia	7	-	7*	-
184	CN11309	*T. aestivum* var Pobeda	PGRC	Russia	7	-	7*	-
185	CN11961	*T. aestivum* var Severodvinskaja 1	PGRC	Russia	7	-	7*	-
186	CN12102	*T. aestivum* var udarnica	PGRC	Russia	7	-	7*	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
187	CItr1667	*T. aestivum* var Beloglina	NSGC	Russian Federation	7	-	7	-
188	CItr3756	*T. aestivum* var Carina	NSGC	Russian Federation	7	-	7	-
189	CItr4795	*T. aestivum* var Ladoga	NSGC	Russian Federation	7	-	7*	-
190	CItr251908	*T. aestivum* var Hungarian	NSGC	Russian Federation	7	-	7*	-
191	PI280452	*T. aestivum* var Iskra	NSGC	Russian Federation	7	-	7*	-
192	PI302424	*T. aestivum* var Bezostaja 1	NSGC	Russian Federation	7	-	7*	-
193	PI361879	*T. aestivum* var Kavkaz	NSGC	Russian Federation	7	-	7*	-
194	PI592051	*T. aestivum* var Zvezda	NSGC	Russian Federation	-	-	n/a	-
195	01C0202953	*T. aestivum* var Vega	RICP,CZECH	Russia	7	-	7*	-
196	CN5835	*T. aestivum* var 71GN No 115	PGRC	Syria	7	-	7*	-
197	CN11461	*T. aestivum*	PGRC	Austria	7	-	7*	-
198	CN10099	*T. aestivum* var Fruher Tiroler Bin	PGRC	Austria	7	-	7*	-
199	CN9834	*T. aestivum* var Comeback	PGRC	Austria	7	-	7*	-
200	CN10622	*T. aestivum* var Janetzkis Fruher	PGRC	Austria	-	-	n/a	-
201	CN10623	*T. aestivum* var Janetzkis Jabo	PGRC	Austria	7	-	7	-
202	PI383388	*T. aestivum* var Extrem	NSGC	Austria	7	-	7*	-
203	CN9514	*T. aestivum* var Alfy 2	PGRC	Belgium	7	+	7	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
204	CN10104	*T. aestivum* var Fylby	PGRC	Belgium	-	-	n/a	-
205	CN10234	*T. aestivum* var Hybride Du Jubile	PGRC	Belgium	7	-	7*	-
206	CN10631	*T. aestivum* var Jufy 2	PGRC	Belgium	-	-	n/a	-
207	CN11288	*T. aestivum* var Phoebus	PGRC	Belgium	-	-	n/a	-
208	CN10981	*T. aestivum* var Lom	PGRC	Bulgaria	7	-	7*	-
209	CN40750	*T. aestivum*	PGRC	Bulgaria	-	-	n/a	-
210	CN40618	*T. aestivum* var Lada	PGRC	Bulgaria	7	-	7*	-
211	CN10904	*T. aestivum* var Kozlodui	PGRC	Bulgaria	7	+	7	-
212	CN40617	*T. aestivum* var Katja A-1	PGRC	Bulgaria	7	+	7	-
213	PI294982	*T. aestivum* var Buffum	NSGC	Bulgaria	7	-	7*	-
214	PI294962	*T. aestivum* var Poljana	NSGC	Bulgaria	7	-	7*	-
215	01C0102607	*T. aestivum* var Vega	RICP,CZECH	Bulgaria	7	-	7/7*	-
216	01C0100187	*T. aestivum* subsp *aestivum* var Samorinska	RICP,CZECH	CSK	7	-	7*	-
217	01C0200103	*T. aestivum* subsp *aestivum* var Sandra	RICP,CZECH	CSK	7	-	7*	-
218	01C0100319	*T. aestivum* subsp *aestivum* var Vega	RICP,CZECH	CSK	7	-	7*	-
219	01C0200129	*T. aestivum* subsp *aestivum* var Maja	RICP,CZECH	CSK	7	-	7	-
220	01C0100285	*T. aestivum* subsp *aestivum* var Hana	RICP,CZECH	CSK	7	-	7*	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
221	01C0100320	*T. aestivum* subsp *aestivum* var Torysa	RICP,CZECH	CSK	7	-	7*	-
222	CN371171	*T. aestivum* var Kinsman	PGRC	Denmark	7	+	7	-
223	CN32504	*T. aestivum*	PGRC	Denmark	7	-	7	-
224	PI125093	*T. aestivum* var Vilmorin 27	NSGC	France	7	+	7	-
225	Citr13723	*T. aestivum* var Druchamp	NSGC	France	-	-	n/a	-
226	Citr6017	*T. aestivum* var Touse	NSGC	France	-	-	n/a	-
227	CItr6709	*T. aestivum* var Japhet	NSGC	France	7	+	7	-
228	PI262231	*T. aestivum* var Etoile De Choisy	NSGC	France	7	-	7*	-
229	PI262228	*T. aestivum* var Poncheau	NSGC	France	7	+	7	-
230	PI167419	*T. aestivum* var Nord Desprez	NSGC	France	7	+	7	-
231	PI262223	*T. aestivum* var Cappelle Desprez	NSGC	France	7	+	7	-
232	CN12168	*T. aestivum* var Werna	PGRC	France	-	-	n/a	-
233	CN44173	*T. aestivum* var Prinqual	PGRC	France	-	-	n/a	-
234	CN42949	*T. aestivum* var Cargimarec	PGRC	France	7	+	7	-
235	CN43797	*T. aestivum* var Ventura	PGRC	France	7	-	7	-
236	CN42403	*T. aestivum* var Arcane	PGRC	France	7	-	7*	-
237	CN12442	*T. aestivum*	PGRC	France	-	-	n/a	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
238	PI201195	*T. aestivum* var Heines VII	NSGC	Germany	-	-	n/a	-
239	CN12003	*T. aestivum* var strubes Fortschritt	PGRC	Germany	7	-	7*	-
240	CN10098	*T. aestivum* var Froubou	PGRC	Germany	7	-	7*	-
241	CN10193	*T. aestivum* var Heines Peko	PGRC	Germany	-	-	n/a	-
242	CN10624	*T. aestivum* var Janetzkis Probat	PGRC	Germany	7	-	7*	-
243	CN10209	*T. aestivum* var Hohenhein	PGRC	Germany	7	-	7*	-
244	01C0101031	*T. aestivum* subsp *aestivum* var Bankuti 1201	RICP,CZECH	Hungary	7	-	7*	-
245	01C0100613	*T. aestivum* subsp *aestivum* var Bankuti	RICP,CZECH	Hungary	7	-	7*	-
246	CN9821	*T. aestivum* var Colonias	PGRC	Hungary	7	-	7*	-
247	CN52368	*T. aestivum* var Eszterhazai No 18	PGRC	Hungary	7	-	7*	-
248	CN11932	*T. aestivum* var San Marino	PGRC	Italy	7	-	7*	-
249	CN9604	*T. aestivum* var Baudi	PGRC	Italy	-	-	n/a	-
250	CN9511	*T. aestivum* var Albimonte	PGRC	Italy	7	-	7*	-
251	CN11642	*T. aestivum*	PGRC	Poland	7	-	7*	-
252	CN9950	*T. aestivum* var Da Maia	PGRC	Portugal	7	-	7	-
253	CN10060	*T. aestivum* var Farrpo	PGRC	Portugal	7	-	7*	-
254	CN12186	*T. aestivum*	PGRC	Romania	7	-	7*	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
255	CN39713	*T. aestivum* var Lovrin 32	PGRC	Romania	7	-	7	-
256	CN9491	*T. aestivum* var Acadimia 48	PGRC	Romania	7	-		
257	CN11765	*T. aestivum*	PGRC	Romania	7	-	7*	-
258	CItr9351	*T. aestivum* var Yeoman II	NSGC	UK	-	-	n/a	-
259	CItr7338	*T. aestivum* var Red Marvel	NSGC	UK	7	+	7	-
260	PI278583	*T. aestivum* var Setter	NSGC	UK	-	-	n/a	
261	PI278572	*T. aestivum* var Benefactress	NSGC	UK	-	-	n/a	-
262	PI278562	*T. aestivum* var Victor	NSGC	UK	-	-	n/a	-
263	CItr6316	*T. aestivum* var Gold Drop	NSGC	UK	-	-	n/a	-
264	PI243192	*T. aestivum* var Hybrid 46	NSGC	UK	-	-	n/a	-
265	CItr6731	*T. aestivum* var Benefactor	NSGC	UK	7	-	7*	-
266	PI193125	*T. aestivum* var Little Joss	NSGC	UK	-	-	n/a	-
267	01C0203281	*T. aestivum* subsp *lutescense* var Musket	RICP,CZECH	GBR	7	-	7	-
268	01C0203285	*T. aestivum* subsp *lutescense* var Tonic	RICP,CZECH	GBR	7	-	7*	-
269	CN10210	*T. aestivum* var Holdfast	PGRC	UK	7	+	7	-
270	CN9635	*T. aestivum* var Bersee	PGRC	UK	7	+	7	-
271	CN9975	*T. aestivum* var Dominator	PGRC	UK	-	-	n/a	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
272	CN12149	*T. aestivum* var Warden	PGRC	UK	-	-	n/a	-
273	CN11998	*T. aestivum* var Squarhead's Master	PGRC	UK	-	-	n/a	-
274	01C0105549	*T. aestivum* subsp *lutescens* var Maja	RICP,CZECH	Yugoslavia	7	-	7*	-
275	-	*T. aestivum* var AC Minto	CRC	Canada	7	-	7*	-
276	-	*T. aestivum* var AC Vista	CRC	Canada	7OE	+	7	+
277	-	*T. aestivum* var Bluesky	CRC	Canada	7OE	+	7	+
278	-	*T. aestivum* var Columbus	CRC	Canada	7	-	7*	-
279	CN38927	*T. aestivum* var Katepwa	PGRC	Canada	7	-	7*	-
280	-	*T. aestivum* var RL 4452	CRC	Canada	7OE	+	7	+
281	-	*T. aestivum* var Roblin	CRC	Canada	7OE	+	7	+
282	CN51820	*T. aestivum* var Wild Cat	PGRC	Canada	7OE	+	7	+
283	-	*T. aestivum* var CDC Teal	CRC	Canada	7OE	+	7	+
284	-	*T. aestivum* var AC Corinne	CRC	Canada	7OE	+	7	+
285	-	*T. aestivum* var Burnside	CRC	Canada	7OE	+	7	+
286	-	*T. aestivum* var Glenavon	CRC	Canada	7OE	+	7	+
287	CN43694	*T. aestivum* var BW90	PGRC	Canada	7OE	+	7	+
288	-	*T. aestivum* var Arcola	CRC	Canada	7	-	7	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
289	CN51812	*T. aestivum* var Biggar	PGRC	Canada	7OE	+	7	+
290	CN44167	*T. aestivum* var Laura	PGRC	Canada	7OE	+	7	+
291	CN42929	*T. aestivum* var HY 320	PGRC	Canada	7OE	+	7	+
292	CN11189	*T. aestivum* var Neepawa	PGRC	Canada	7	-	7*	-
293	CN44146	*T. aestivum* var HY358	PGRC	Canada	7OE	+	7	+
294	-	*T. aestivum* var Glenlea	CRC	Canada	7OE	+	7	+
295	CN44438	*T. aestivum* var Oslo	PGRC	Canada	7OE	+	7	+
296	CN39039	*T. aestivum* subsp *aestivum* var Thatcher	PGRC	Canada	-	-	n/a	-
297	CItr14193	*T. aestivum* var Red River 68	USDA-ARS	USA	7OE	+	7	+
298	-	*T. aestivum* var Nordic	CRC	USA	7OE	+	7	+
299	PI520297	*T. aestivum* var Stoa	USDA-ARS	USA	7	-	7*	-
300	CN10641	*T. aestivum* var Kanred	PGRC	USA	7	+	7	-
301	CItr14108	*T. aestivum* var Chinese Spring	USDA-ARS	USA	7	-	7	-
302	CItr11666	*T. aestivum* var Cheyenne selection	NSGC	USA	7	-	7*	-
303	CItr8178	*T. aestivum* var Hope	NSGC	USA	-	-	n/a	-
304	CItr8885	*T. aestivum* var Cheyenne	NSGC	USA	7	-	7*	-
305	CWI16281	*T. aestivum* var Prospur	CIMMYT	USA	7OE	+	7	+

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
306	CWI35704	*T. aestivum* var Wheaton	CIMMYT	USA	-	-	n/a	-
307	BW386	*T. aestivum* var Bajio	CIMMYT	Mexico	7OE	+	7	+
308	CWI3767	*T. aestivum* var Tobari F66	CIMMYT	Mexico	-	-	n/a	-
309	CItr12606	*T. aestivum* var Klein Universal	USDA-ARS	Argentina	7OE	+	7	+
310	CWI77253	*T. aestivum* var Klein Sendero	CIMMYT	Argentina	7OE	+	7	+
311	BW12005	*T. aestivum* var Victoria INTA	CIMMYT	Argentina	7OE	+	7	+
312	CWI33193	*T. aestivum* var Buck Nandu	CIMMYT	Argentina	-	-	n/a	-
313	CWI3764	*T. aestivum* var Klein Toledo	CIMMYT	Argentina	7	-	7*	-
314	BWI1255	*T. aestivum* var Buck Pucara	CIMMYT	Argentina	7OE	+	7	+
315	BW464	*T. aestivum* var Calidad	CIMMYT	Argentina	7OE	+	7	+
316	CWI14942	*T. aestivum* var Klein Sin Rival	CIMMYT	Argentina	7	-	7*	-
317	BW15246	*T. aestivum* var Pampa INTA	CIMMYT	Argentina	7OE	+	7	+
318	CWI33350	*T. aestivum* var Retacon INTA	CIMMYT	Argentina	7OE	+	7	+
319	BW4689	*T. aestivum* var Klein Atlas	CIMMYT	Argentina	7OE	+	7	+
320	CWI14048	*T. aestivum* var Universal II	CIMMYT	Argentina	7OE	+	7	+
321	BW779	*T. aestivum* var Tezanos Printos Precoz	CIMMYT	Argentina	7OE	+/-	7	+
322	CN9631	*T. aestivum* var Benvenuto 3085	PGRC	Argentina	7	-	7*	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
323	CN10020	*T. aestivum* var El Gaucho	PGRC	Argentina	7OE	+	7	+
324	CN9544	*T. aestivum* var Argentina	PGRC	Argentina	7	-	7*	-
325	CN10862	*T. aestivum* var Klein Otto Wulff	PGRC	Argentina	7	-	7*	-
326	CN11791	*T. aestivum* var Rio Negro	PGRC	Argentina	7	-	7*	-
327	PI382144	*T. aestivum* var Encruzilhada	NSGC	Brazil	-	-	n/a	-
328	CItr12019	*T. aestivum* var Fronteira	NSGC	Brazil	-	-	n/a	-
329	CItr12470	*T. aestivum* var Frontana	NSGC	Brazil	7	-	7*	-
330	PI351654	*T. aestivum* var Surpresa	NSGC	Brazil	-	-	n/a	-
331	CN11323	*T. aestivum* var Preludio	PGRC	Brazil	7	-	7*	-
332	CN44009	*T. aestivum* var Trintecinco	PGRC	Brazil	7	-	7*	-
333	CN44011	*T. aestivum* var Toropi	PGRC	Brazil	7OE	+	7	+
334	CN12082	*T. aestivum* var Tropeano	PGRC	Brazil	7	-	7*	-
335	CN44002	*T. aestivum* var Colonias	PGRC	Brazil	-	-	n/a	-
336	CN42519	*T. aestivum* var Cinquentenario	PGRC	Brazil	-	-	n/a	-
337	CN10084	*T. aestivum* var Fortaleza	PGRC	Brazil	7	-	7*	-
338	CN11100	*T. aestivum* var Mentana	PGRC	Brazil	7	-	7*	-
339	CN10028	*T. aestivum* var Equator	PGRC	Brazil	-	-	n/a	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
340	CN11098	*T. aestivum* var Menkemen	PGRC	Colombia	7	-	7*	-
341	CN9524	*T. aestivum* var Andes	PGRC	Colombia	7	-	7*	-
342	CN9666	*T. aestivum* var Bonza	PGRC	Colombia	7	-	7*	-
343	CN9661	*T. aestivum* var Bola Picota	PGRC	Colombia	-	-	n/a	-
344	CN12624	*T. aestivum*	PGRC	Colombia	7	-	7*	-
345	-	*T. aestivum* var Tambillo 1	PGRC	Ecuador	-	-	n/a	-
346	-	*T. aestivum* var Atacatzo 1	PGRC	Ecuador	7	-	7*	-
347	CN11057	*T.aestivum* subsp *aestivum* var Maria Escobar	PGRC	Peru	-	-		-
348	CN12358	*T. aestivum*	PGRC	Peru	-	-	n/a	-
349	CN12268	*T. aestivum*	PGRC	Peru	7	-	7	-
350	CN11698	*T. aestivum*	PGRC	Peru	-	-	n/a	-
351	CN11058	*T. aestivum* var Maribal 50	PGRC	Peru	7	-	7	-
352	CN10196	*T. aestivum* var Helvia	PGRC	Peru	7	-	7*	-
353	PI191937	*T aestivum* var Americano 44D	USDA-ARS	Uruguay	7OE	+	7	+
354	CN11969	*T aestivum* var Sinvalocho	PGRC	Uruguay	7OE	+	7	+
355	CN9591	*T. aestivum* var Bage	PGRC	Uruguay	7	-	7	-
356	CN9832	*T. aestivum* var Combate	PGRC	Uruguay	-	-	n/a	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
357	CN10856	*T. aestivum* var Klein Credito	PGRC	Uruguay	7OE	+	7	+
358	CN12090	*T. aestivum* var Trintani	PGRC	Uruguay	-	-	n/a	-
359	CN11243	*T. aestivum* var olaeta Calandria	PGRC	Uruguay	7OE	+	7	+
360	CN12361	*T. aestivum*	PGRC	Congo	-	-	n/a	-
361	CN12666	*T. aestivum*	PGRC	Congo	-	-	n/a	-
362	CN12425	*T. aestivum*	PGRC	Egypt	-	-	n/a	-
363	CN11137	*T. aestivum* var Mokhtar Improved	PGRC	Egypt	-	-	n/a	-
364	CN9794	*T. aestivum*	PGRC	Egypt	7	-	7	-
365	CN11720	*T. aestivum*	PGRC	Egypt	-	-	n/a	-
366	CN10150	*T. aestivum* var Giza 141	PGRC	Egypt	-	-	n/a	-
367	CN11307	*T. aestivum*	PGRC	Ethiopia	7	-	7*	-
368	CN6024	*T. aestivum*	PGRC	Ethiopia	7	-	7	-
369	CN6171	*T. aestivum*	PGRC	Ethiopia	7	-	7*	-
370	CN40895	*T. aestivum*	PGRC	Ethiopia	7	-	7	-
371	CN2646	*T. aestivum* var Camadi	PGRC	Ethiopia	7	-	7	-
372	CN6030	*T. aestivum*	PGRC	Ethiopia	7	-	7	-
373	CN6025	*T. aestivum*	PGRC	Ethiopia	7	-	7	-

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
374	CN12153	*T. aestivum* var Warigo	PGRC	Kenya	-	-	n/a	-
375	CN12388	*T. aestivum*	PGRC	Kenya	-	-	n/a	-
376	CN11917	*T. aestivum* var Sabanero	PGRC	Kenya	-	-	n/a	-
377	CN10892	*T. aestivum* var Koalisie	PGRC	Kenya	-	-	n/a	-
378	CN10750	*T. aestivum*	PGRC	Kenya	-	-	n/a	-
379	CN10719	*T. aestivum* var Kenya Farmer	PGRC	Kenya	7	-	7*	-
380	CN9818	*T. aestivum* var Cobbs 1066	PGRC	Kenya	-	-	n/a	-
381	CN10661	*T. aestivum*	PGRC	Kenya	-	-	n/a	-
382	CItr12471	*T. aestivum* var Kenya 58	NSGC	Kenya	7	-	7	-
383	PI320108	*T. aestivum* var Santa Elena	NSGC	Kenya	7	-	7*	-
384	CN11281	*T. aestivum* var Penkop	PGRC	South Africa	-	-	n/a	-
385	CN11059	*T. aestivum* var Marquillo	PGRC	South Africa	7	-	7*	-
386	CN9995	*T. aestivum* var Duiken	PGRC	South Africa	-	-	n/a	-
387	CN9984	*T. aestivum* var Dromedaris	PGRC	South Africa	-	-	n/a	-
388	CN9946	*T. aestivum* var Daeraad	PGRC	South Africa	-	-	n/a	-
389	01C0203832	*T. aestivum* var lutescens var Sabre	RICP,CZECH	Australia	7	-	7*	-
390	Aus29472	*T. aestivum* var Kukri	AWCC	Australia	7OE	+	7	+

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
391	Aus30031	*T. aestivum* var Chara	AWCC	Australia	7OE	+	7	+
392	CItr4996	*T. aestivum* var DART	NSGC	Australia	-	-	n/a	-
393	CItr4984	*T. aestivum* var Major	NSGC	Australia	-	-	n/a	-
394	CItr4981	*T. aestivum* var White Federation	NSGC	Australia	7	-	7*	-
395	CItr4121	*T. aestivum* var John Brown	NSGC	Australia	7	-	7*	-
396	CItr4067	*T. aestivum* var Pacific Bluestem	NSGC	Australia	-	-	n/a	-
397	CItr5125	*T. aestivum* var Bunyip	NSGC	Australia	-	-	n/a	-
398	CItr4733	*T. aestivum* var Hard Federation	NSGC	Australia	7	-	7*	-
399	CItr1697	*T. aestivum* var BAART	NSGC	Australia	-	-	n/a	-
400	CItr4608	*T. aestivum* var Jumbuck	NSGC	Australia	7	-	7*	-
401	CItr4734	*T. aestivum* var Federation	NSGC	Australia	-	-	n/a	-
402	CN10112	*T. aestivum* var Gabo	PGRC	Australia	-	-	n/a	-
403	CN10207	*T. aestivum* var Hofed	PGRC	Australia	-	-	n/a	-
404	CN9625	*T. aestivum* var Bencubbin	PGRC	Australia	7	-	7*	-
405	CN9682	*T. aestivum* var Bungulla	PGRC	Australia	7	-	7*	-
406	PI483064	*T aestivum* var Sunstar	USDA-ARS	Australia	7	-	7*	-
407	Aus30426	*T. aestivum* var Otane	AWCC	New Zealand	7OE	+	7	+

SI.No	Accesion No	Accession name	Source	Country of origin	HMW-GS (SDS-PAGE)	43 bp indel	18 bp indel	Left and right LTR junction
408	CN9851	*T. aestivum* var Cross 7	PGRC	New Zealand	-	-	n/a	-
409	CN10202		PGRC	New Zealand	-	-	n/a	-
410	CN9541		PGRC	New Zealand	7	+	7	-
411	CN6622		PGRC	New Zealand	-	-	n/a	-
412	CN12035		PGRC	New Zealand	-	-	n/a	-

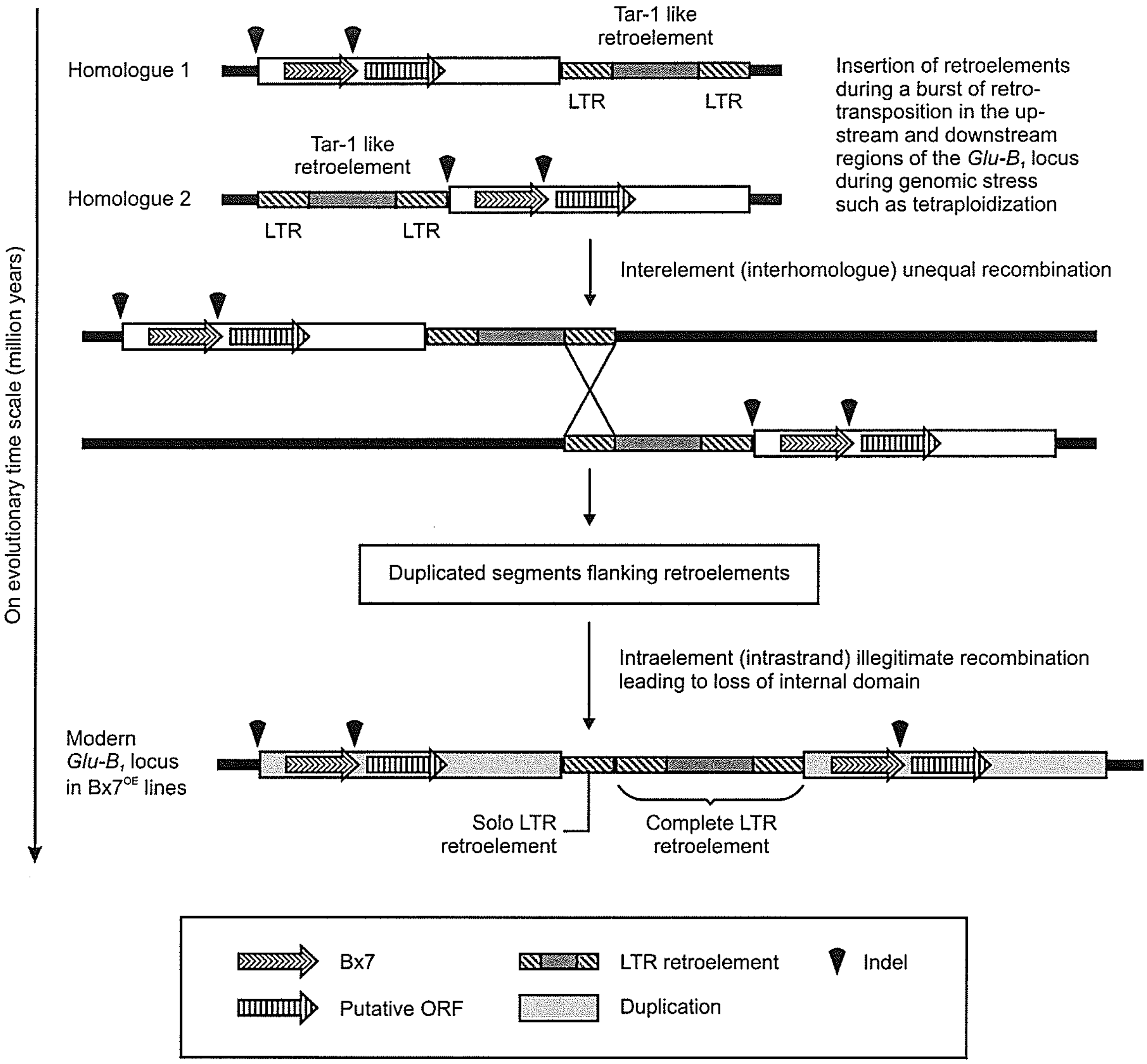
Note: n/a not applicable

Appendix IV. RP-HPLC analyses of a subset of *Triticum* accessions

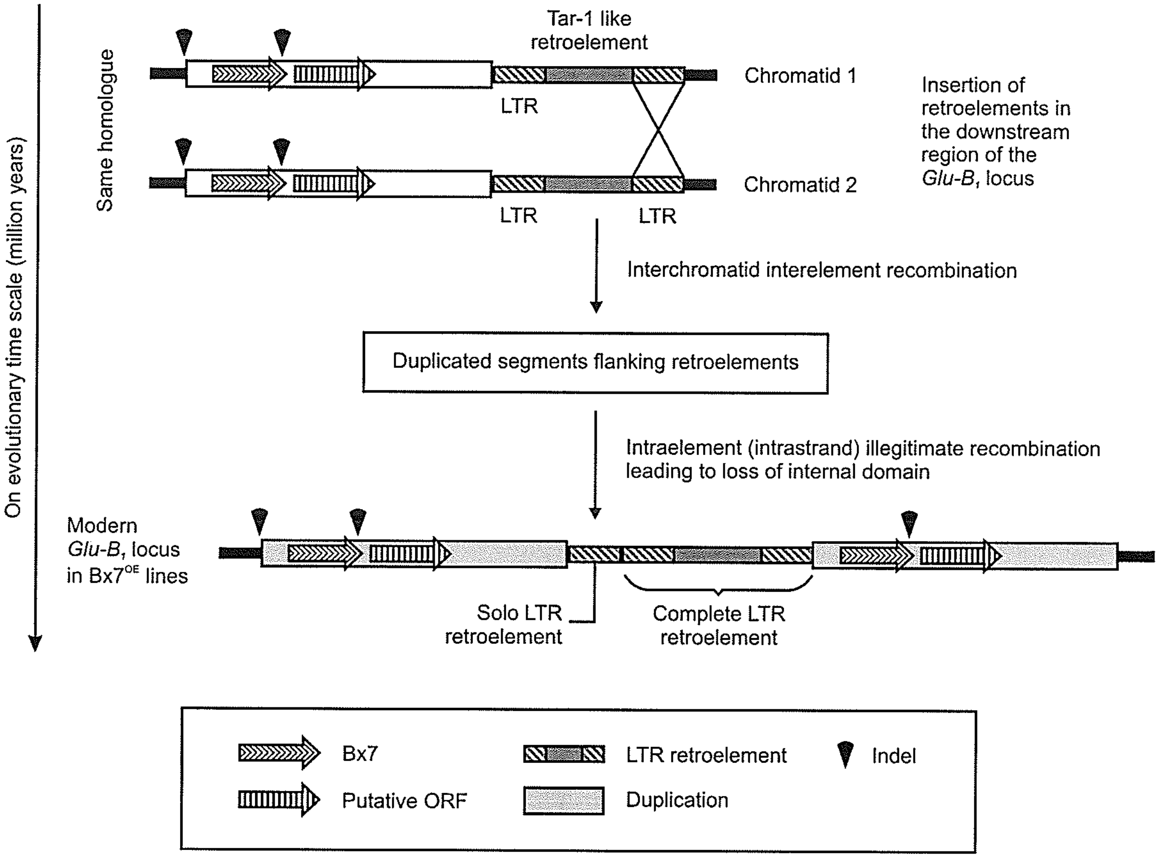
					% of Bx7 subunit to total HMW-GS			
SI No	Accession No	Accession name	Species	Country of origin	Analysis 1	Analysis 2	Average	Parameters
1	-	AC Vista	*T. aestivum*	Canada	43.38	41.59	42.49	
2	-	Bluesky	*T. aestivum*	Canada	41.97	40.96	41.47	
3	CN 44438	Oslo	*T. aestivum*	Canada	42.96	41.90	42.43	
4	CItr 14193	Red River 68	*T. aestivum*	USA	40.57	40.77	40.67	
5	-	RL 4452	*T. aestivum*	Canada	37.16	36.72	36.94	
6	-	Roblin	*T. aestivum*	Canada	43.16	43.34	43.25	
7	CN 51820	Wildcat	*T. aestivum*	Canada	37.88	38.64	38.26	
8	PI 191937	Americano 44D	*T. aestivum*	Uruguay	41.09	39.54	40.32	
9	-	CDC Teal	*T. aestivum*	Canada	37.73	38.98	38.36	
10	-	AC Corinne	*T. aestivum*	Canada	42.01	41.44	41.73	
11	-	Burnside	*T. aestivum*	Canada	40.47	41.31	40.89	
12	-	Glenavon	*T. aestivum*	Canada	40.43	41.69	41.06	
13	CN 43694	BW90	*T. aestivum*	Canada	44.16	44.32	44.24	
14	-	Nordic	*T. aestivum*	USA	44.50	43.98	44.24	
15	-	TAA36	*T. aestivum*	Israel	43.56	43.57	43.57	
16	CItr 12606	Klein Universal	*T. aestivum*	Argentina	39.82	39.69	39.76	
17	CN 51812	Biggar	*T. aestivum*	Canada	40.97	40.57	40.77	
18	CN 44167	Laura	*T. aestivum*	Canada	43.29	42.57	42.93	
19	CN 42929	HY320	*T. aestivum*	Canada	41.24	41.56	41.40	
20	CN 44146	HY358	*T. aestivum*	Canada	39.86	38.57	39.22	
21	CN 11969	Sinvalocho	*T. aestivum*	Argentina	44.72	43.42	44.07	
22	-	Glenlea	*T. aestivum*	Canada	38.12	38.67	38.40	
23	CWI 16281	Prospur	*T. aestivum*	USA	37.08	37.23	37.16	

					% of Bx7 subunit to total HMW-GS			
24	BW 386	Bajio	*T. aestivum*	Mexico	37.27	37.43	37.35	
25	CWI 77253	Klein Sendero	*T. aestivum*	Argentina	46.75	47.54	47.15	
26	BW 12005	Victoria INTA	*T. aestivum*	Argentina	41.04	40.34	40.69	
27	BWI 1255	Buck Pucara	*T. aestivum*	Argentina	41.58	42.32	41.95	
28	BW 464	Calidad	*T. aestivum*	Argentina	43.94	44.71	44.33	
29	BW 152416	Pampa INTA	*T. aestivum*	Argentina	43.54	44.44	43.99	
30	CWI 33350	Retacon INTA	*T. aestivum*	Argentina	40.32	41.79	41.06	
31	BW 4689	Klein Atlas	*T. aestivum*	Argentina	47.01	47.46	47.24	
32	CWI 14048	Universal II	*T. aestivum*	Argentina	41.81	40.9	41.36	
33	BW 779	Tezanos Printos Precoz	*T. aestivum*	Argentina	42.78	42.80	42.79	
34	CN 10020	El Gaucho	*T. aestivum*	Argentina	45.37	44.75	45.06	
35	CN 44011	Toropi	*T. aestivum*	Brazil	39.58	40.42	40.00	
36	CN 10856	Klein Credito	*T. aestivum*	Uruguay	44.49	44.30	44.40	
37	CN 11243	Olaeta Calandria	*T. aestivum*	Uruguay	40.24	40.58	40.41	
38	Aus 29472	Kukri	*T. aestivum*	Australia	44.18	44.30	44.24	
39	Aus 30426	Otane	*T. aestivum*	New Zealand	42.20	43.67	42.94	
40	Aus 30031	Chara	*T. aestivum*	Australia	40.88	41.13	41.01	
							Mean	41.74
							Max	47.24
							Min	36.94
							SD	2.51
Tetraploid accessions with Bx7^{OE} subunit								
41	CN 2644	Branco	*T. turgidum*	Portugal	64.12	62.45	63.29	
42	CN 12222	CN 12222	*T. turgidum*	Czech	64.21	63.1	63.66	

					% of Bx7 subunit to total HMW-GS				
43	CN 12225	CN 12225	*T. turgidum*	Czech	68.83	74.36	71.60		
								Mean	66.18
								Max	71.60
								Min	63.29
								SD	4.69
Checks (Hexaploid cultivars)									
44	PI 447404	Yang Mai No.1	*T. aestivum*	China	29.92	29.27	29.60		
45	CItr 6731	Benefactor	*T. aestivum*	UK	30.05	30.46	30.26		
46	01C0100613	Bankuti	*T. aestivum*	Hungary	30.80	30.50	30.65		
47	CWI 14942	Klein Sin Rival	*T. aestivum*	Argentina	30.25	30.54	30.40		
48	CN 10719	Kenya Farmer	*T. aestivum*	Kenya	29.02	30.58	29.80		
49	01C0200129	Maja	*T. aestivum*	Czech Republic	30.73	30.89	30.81		
50	CItr 8885	Cheyenne	*T. aestivum*	USA	27.61	27.55	27.58		
51	CN 38927	Katepwa	*T. aestivum*	Canada	28.55	29.44	29.00		
52	CN 11189	Neepawa	*T. aestivum*	Canada	29.57	29.55	29.56		
53	PI 520297	Stoa	*T. aestivum*	USA	28.96	28.90	28.93		
54	-	AC Minto	*T. aestivum*	Canada	30.09	30.08	30.09		
55	-	Columbus	*T. aestivum*	Canada	28.95	28.96	28.95		
								Mean	29.63
								Max	30.81
								Min	27.58
								SD	0.91
Check (Tetraploid accession)									
56	CN 51263	CN 51263	*T. turgidum*	Russia	56.65	57.42	57.04		


Tar-1 like
retroelement
Homologue 1
LTR
LTR
Insertion of retroelements during a burst of retro-transposition in the up-stream and downstream regions of the Glu-B1 locus during genomic stress such as tetraploidization
Tar-1 like
retroelement
Homologue 2
LTR
LTR
Interelement (interhomologue) unequal recombination
Duplicated segments flanking retroelements
Intraelement (intrastrand) illegitimate recombination leading to loss of internal domain
Modern Glu-B1 locus in Bx7OE lines
Solo LTR retroelement
Complete LTR retroelement
On evolutionary time scale (million years)
Bx7
Putative ORF
LTR retroelement
Duplication
Indel

Appendix V.A Model for the evolution of *Glu-B1* locus with segmental duplication involving inter-homologue unequal recombination.


On evolutionary time scale (million years)
Same homologue
Tar-1 like retroelement
Chromatid 1
LTR
Chromatid 2
LTR
LTR
Insertion of retroelements in the downstream region of the Glu-B1 locus
Interchromatid interelement recombination
Duplicated segments flanking retroelements
Intraelement (intrastrand) illegitimate recombination leading to loss of internal domain
Modern Glu-B1 locus in Bx7OE lines
Solo LTR retroelement
Complete LTR retroelement
Bx7
Putative ORF
LTR retroelement
Duplication
Indel

Appendix V.B Model for the evolution of *Glu-B1* locus involving intra-homologue (inter-chromatid) recombination.