NEXT-GENERATION SEQUENCING AND PHYSICAL MAPPING OF WHEAT CHROMOSOME 5D AND COMPARISON WITH ITS WILD PROGENITOR

by

BÂLÂ ANI AKPINAR

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Sabancı University
February 2015

NEXT-GENERATION SEQUENCING AND PHYSICAL MAPPING OF WHEAT CHROMOSOME 5D AND COMPARISON WITH ITS WILD PROGENITOR

APPROVED BY:

Prof. Dr. Hikmet Budak ....
(Dissertation Supervisor)

Assoc. Dr. Ali Koşar ....

Assoc. Dr. Levent Öztürk ....

Prof. Dr. Müge Türet ....

Prof. Dr. Zehra Sayers ....

DATE OF APPROVAL: 13/02/2015

ABSTRACT

NEXT-GENERATION SEQUENCING AND PHYSICAL MAPPING OF WHEAT CHROMOSOME 5D AND COMPARISON WITH ITS WILD PROGENITOR

Bâlâ Anı Akpınar

PhD Thesis 2015

Prof. Dr. Hikmet Budak (Thesis supervisor)

Keywords: Wheat, next-generation sequencing, physical mapping, chromosome 5D, comparative genomics

Wheat is a staple grain crop, essential to human nutrition and animal feed. Despite its agronomic importance, wheat genomics research has long lagged behind its counterparts, due to its genome attributes. Bread wheat genome is almost six times as large as the human genome at a size of ~17 Gigabases, and is composed of >80% repetitive elements. The hexaploid genome is organized into three related sub-genomes, giving rise to numerous paralogous and homeologous loci. In this study, we characterized the flow-sorted 5D chromosome of bread wheat, *Triticum aestivum*, through survey sequencing and physical mapping, including its repeat landscape, gene content and conservation and putative tRNA repertoire. The virtual gene order of 5D chromosome suggested several perturbations in synteny, in addition to a number of putatively wheat-specific genome rearrangements. The 5DS physical map revealed that its gene space is largely organized into gene islands with an increasing gradient towards the telomere. Physical size estimates on the physical map indicated that cytogenetic estimates may considerably underestimate the 0.63-0.67 deletion bin interval. Comparative analyses of its wild progenitor, *Aegilops tauschii* 5D chromosome shed light into wheat genome evolution. The high density 5DS physical map at ~10.5 markers/Mb and 1.34x-1.61x survey sequences of the entire chromosome provides the foundation of the reference sequencing of this chromosome and presents a valuable genomics resource that the breeders and the researchers should benefit from.

ÖZET

BUĞDAY 5D KROMOZOMUNUN YENİ-NESİL DİZİLEMESİ, FİZİKSEL HARİTALAMASI ve YABANİ ATASIYLA KARŞILAŞTIRILMASI

Bâlâ Anı Akpınar
Doktora Tezi 2015
Prof. Dr. Hikmet Budak (Tez danışmanı)

Anahtar kelimeler: Buğday, yeni-nesil dizileme, fiziksel haritalama, 5D kromozomu, karşılaştırmalı genomiks

Başlıca ekinlerimizden olan buğday, temel bir gıda maddesi ve hayvan yemi kaynağıdır. Tarımsal önemine rağmen, buğdayda genomik çalışmalar, genom özellikleri nedeniyle, uzun zamandır, diğer ekinlerin gerisinde kalmıştır. Ekmeklik buğday genomu, ~17 Gigabaz büyüklüğü ile insan genomunun neredeyse 6 katı büyüklükte olup, %80'den fazla oranda tekrarlı dizi içermektedir. Hekzaploid genomu, birbirine benzer üç alt genomdan oluştuğu için pek çok paralog ve homeolog lokusu kapsar. Bu çalışmada, tekrarlı dizi düzeni, gen içeriği ve korunması ve muhtemel tRNA içeriği dahilinde, akış sitometrisi ile saflaştırılmış, ekmeklik buğday, *Triticum aestivum*, 5D kromozomunu karakterize ettik. 5D kromozomunun sanal gen sırası, buğdaya özgü genom düzenlemelerinin yanısıra, korunmuş gen bloklarında pek çok karışıklık olabileceğini ortaya çıkardı. 5DS fiziksel haritası ise, gen düzleminin telomere doğru artan yoğunlukta gen adacıklarından oluştuğunu gösterdi. Fiziksel boyut tahminleri, 0.63-0.67 delesyon bölge aralığının sitogenetik tahminlerde ciddi ölçüde küçültülmüş olabileceğine işaret etti. Yabani atası *Aegilops tauschii* 5D kromozomu ile karşılaştırmalı analizler ise buğday genom evrimine ışık tuttu. Megabaz başına ~10.5 markör ile yüksek yoğunluklu 5DS fiziksel haritası ve 1.34x-1.61x kapsamalı tüm kromozom dizileri bu kromozomun referans dizilemesine temel teşkil etmekte olup, hem ıslahçıların hem de araştırmacıların yararlanabileceği değerli bir genomik kaynak da sunmaktadır.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my thesis advisor, Prof. Hikmet Budak, for his guidance and continuous support throughout this research. He has provided me with invaluable opportunities that, together with his expertise, greatly assisted me accomplishing my academic goals.

I would like to extend my gratitude to my thesis committee members, Prof. Müge Türet, Prof. Zehra Sayers, Assoc. Prof. Ali Koşar, and Assoc. Prof. Levent Öztürk for their precious time and constructive feedback that substantially improved this work.

I have most certainly learnt a lot from Dr. Stuart J. Lucas; his help is greatly appreciated. I am also grateful to Dr. Meral Yüce for our extensive hours in the lab and for her assistance. I am thankful to our IT System Specialist, Barış Tümer; his assistance, particularly on urgent occasions, has helped this research progress smoothly.

I would like to thank my lab members, and, in particular, Dr. Melda Kantar, for walking along this academic journey with me. I greatly appreciate and acknowledge the friendship and guidance of my former lab members, Duygu Kuzuoğlu-Öztürk and Ayşegül Altıntaş.

I cannot possibly thank enough Ahmet Sinan Yavuz for his tremendous patience in answering my each and every question and for his sincere friendship. I am grateful to have my dearest friends, Ayşe Bengisu Seferoğlu and Özlem Şenkaya; my gratitude for their love and support is beyond words. Additionally, being from similar research areas, our academic discussions with A. Bengisu Seferoğlu undeniably helped me improve myself.

I would like to thank TÜBİTAK-BİDEB for the scholarship they provided during this work. Part of this research was funded by TÜBİTAK (Grant no. 111O665).

I am grateful to our family friend Mustafa Özay and his family, for their support and belief in me, without a hesitation. I owe any single piece of my success to my parents, Nesibe Akpınar and Fikri Akpınar; I cannot imagine a life, without their love, support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

APS	Adenosine 5’ Phosphosulfate
BAC	Bacterial Artificial Chromosome
BES	BAC-End Sequencing
CB	Consensus Band
ChIP	Chromatin Immunoprecipitation
CNV	Copy Number Variation
COS	Conserved Orthologous Set
CTG	Contig
DAPI	4',6-diamidino-2-phenylindole
emPCR	Emulsion PCR
EST	Expressed Sequence Tag
FISH	Fluorescence *In Situ* Hybridization
FPB	FingerPrinted Background removal
FPC	FingerPrinted Contig
Gb	Gigabases
GO	Gene Ontology
HICF	High-Information Content Fingerprinting
InDel	Insertion-Deletion
ISBP	Insertion Site-Based Polymorphism
ITMI	International Triticeae Mapping Initiative
IWGSC	International Wheat Genome Sequencing Consortium
LD	Linkage Disequilibrium
LTC	Linear Topology Contig
MAS	Marker-Assisted Selection
Mb	Megabases
MBC	Map-Based Cloning

MDA	Multiple Displacement Amplification
MTP	Minimum Tiling Path
Mya	Million years ago
NGS	Next-Generation Sequencing
PAV	Presence-Absence Variation
PCR	Polymerase Chain Reaction
PTP	PicoTiterPlate
RNA-Seq	RNA-Sequencing
SBS	Sequencing-By-Synthesis
SC	Supercontig
SNP	Single Nucleotide Polymorphism
SSR	Simple Sequence Repeat
TE	Transposable Element
TF	Transcription Factor
WGD	Whole Genome Duplication
WGP	Whole Genome Profiling
WGS	Whole Genome Shotgun

1. INTRODUCTION

Agricultural production faces major challenges as the world population continues to grow and climate changes progressively affect crop yields, while the acreage of arable lands remains essentially the same. Through the agricultural history, domestication and systemic breeding have achieved steady yield gains at the expense of genetic diversity. Consequently, the gene pools of today's elite cultivars are considerably narrow, and further improvements through breeding appear to necessitate effective exploration and utilization of the germsplasms, including wild populations and landraces.

The advances in molecular biology and reducing costs of sequencing technologies have opened up new avenues for crop improvement through genome sequencing and genomics research, which enable extensive characterization of genetic stocks and mutant collections. For the past few years, fierce efforts have unraveled genome sequences of many model and crop plants, and ongoing efforts are now directed to tackle the crop genomes that were once considered intractable.

One such crop plant, bread wheat, has a 17 Gigabase long hexaploid genome that is composed of >80% repetitive elements. Despite being an essential component of nutrition and a leading crop, the genome attributes of wheat have long hindered genomics studies. The flow sorting and physical mapping of its largest chromosome, 3B, have set the pace in wheat genomics and subsequently led to the very recent report of its reference sequencing (Paux *et al.*, 2008; Choulet *et al.*, 2014).

Here we describe the next-generation sequencing and physical mapping of bread wheat, *Triticum aestivum*, chromosome 5D, and its comparison with its counterpart from the D-genome progenitor *Aegilops tauschii*. Genomic resources generated in this study can readily be applied to map-based cloning of important genes and alleles from this chromosome. On the long run, these resources will provide a framework for the future reference sequencing of this chromosome, which represents a significant piece of the wheat genome puzzle.

2. OVERVIEW

2.1. Wheat as a leading crop

Food security is a growing concern across the globe. Roughly one in seven individuals is estimated to be under- or malnourished worldwide (Foley *et al.*, 2011). Although the prevalence of undernourishment has decreased during the past two decades, food security is likely to remain a major issue, as the world population is projected to exceed 9 billion by 2050, necessitating an estimated increase of 60% in global agricultural production to meet the food demand (FAO, 2013).

Cereal crops are the main sources of human nutrition and animal feed. Among the cereals, wheat, a cereal grain crop, currently ranks the third, following rice and maize, with an annual production of over 713 million tonnes in 2013 (http://faostat3.fao.org/). Since maize is generally used as animal feed, wheat is actually the second major constituent of human nourishment and provides nearly 1/5 of the total caloric input (Reynolds *et al.*, 2009; FAO, 2013). Wheat is the most extensively grown food crop, harvested across over 218 million hectares worldwide; due to its hardy nature, wheat is capable of growing across a wide range of environments. However, climate changes and the increasing use of crops for biofuel production hinder crop production (Foley *et al.*, 2011). In particular, climate trends are estimated to cause a 5.5% loss in wheat production between 1980-2008 (Lobell *et al.*, 2011). Further improvements on crop production to feed the growing world population will be tightly linked to increased yields.

2.2. DNA Sequencing

2.2.1. First generation sequencing technologies

The sequence of the DNA had intrigued scientists since the discovery of the “double helix” in 1953. By that time, the notion of proteins made up of amino acid residues arranged in an arbitrary but defined order was already known, and the order of the amino acid residues was attributed to the sequence of the DNA fragment encoding the corresponding protein. However, the exact mechanism was unknown. However, the experimental determination of the DNA sequence could not be achieved for 15 years, largely because DNA molecules are usually much longer than proteins and the incorporation of only 4 bases in any DNA molecule complicates the chemical separation of different DNA fragments (Hutchison, 2007).

The discovery of type II restriction nucleases cleaving DNA at specific recognition sites and the use of polyacrylamide gels for the separation of DNA fragments with different sizes had been crucial in the development of first generation sequencing methodologies. Type II restriction nucleases enabled the long DNA molecules to be cut into smaller fragments with specific ends that can be used in priming the sequencing reaction (Hutchison, 2007). Consequently, the first complete genome sequence was published in 1977, which belonged to the ~5,375 nucleotide-long genome of the ϕX174 bacteriophage (Sanger *et al.*, 1977a).

Near the end of 1977, Sanger and his colleagues described a new DNA sequencing methodology utilizing chain-terminating inhibitors (Sanger *et al.*, 1977b). Although DNA sequencing had been carried out for a couple of years prior, the introduction of this new method, commonly known as the “Sanger sequencing” or “dideoxy sequencing” today, had been pivotal. Sanger sequencing relies on the termination of the growing DNA chain by 2',3'-dideoxynucleotides (ddNTPs), modified analogues of natural 2’-deoxynucleotides (dNTPs). Since ddNTPs lack the 3’-hydroxyl group, DNA polymerase cannot elongate the complementary DNA strand, once a ddNTP is incorporated into the growing chain. In the presence of a mixture of dNTPs and ddNTPs at a certain ratio, DNA polymerase produces a mixture of nested fragments, which can be separated by gel electrophoresis to deduce the sequence of the original DNA fragment (Fig. 1) (Sanger *et al.*, 1977b). Sanger sequencing quickly became the method of choice as the first generation DNA sequencing techniques and dominated the DNA sequencing era for three decades. Over the years, Sanger sequencing had been significantly improved through technological advances, including the use of fluorescent dyes, improved detection methods and capillary electrophoresis and microfluidic platforms, and was automated (Metzker, 2005), eventually, forming the basis of the Human Genome Project (Lander *et al.*, 2001; Venter *et al.*, 2001).

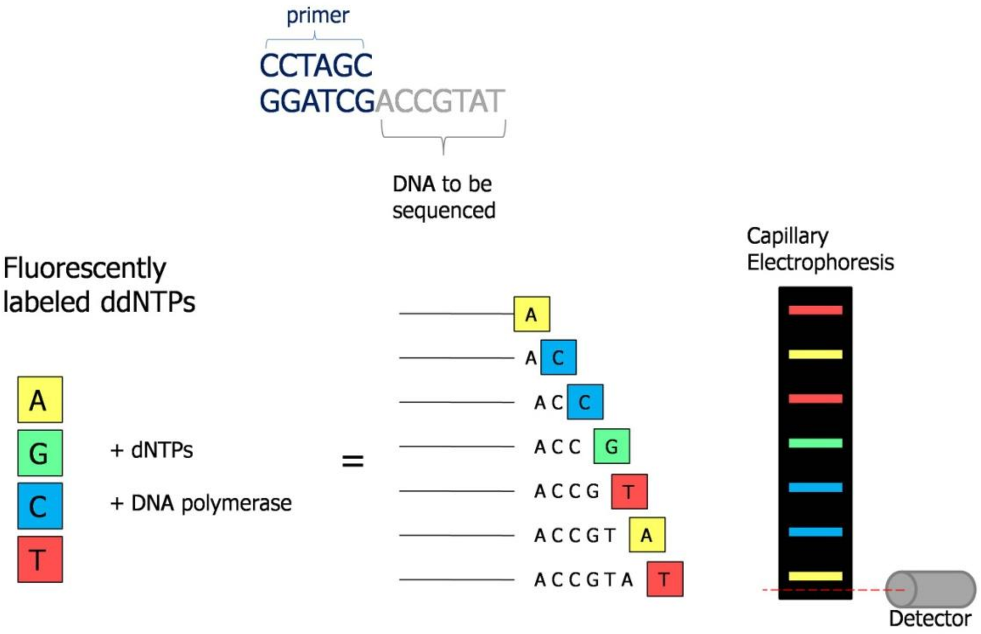


Figure 1. Schematic overview of Sanger sequencing with the current techological advances.

2.2.2. Next-Generation Sequencing technologies

The completion of the Human Genome Project marked the beginning of a new era in DNA sequencing. As the benefits of sequencing and re-sequencing of human genomes were realized, in particular, for disease research, a tremendous need for sequencing quickly built up. However, despite the remarkable success of Sanger sequencing, the inherent limitations of this sequencing methodology necessitated the development of novel sequencing approaches (Metzker, 2010). Sanger sequencing is low-throughput, tedius and costly; in fact, the Human Genome Project was completed at a cost of $2.7 billion using automated Sanger sequencing (http://www.genome.gov/).

Initially, Next-Generation Sequencing (NGS) technologies were developed for the re-sequencing purposes. Soon after, the high-throughput capacities of these NGS platforms able to carry out reasonably accurate sequencing at considerably reduced costs have led to the adoption of these technologies as the primary sequencing approach. Currently, NGS technologies are applied to a broad range of research areas, including but not limited to genomics, transcriptomics, metagenomics, forensic science, epidemiology, diagnostics and therapeutics (Metzker, 2005; Hutchison, 2007). Remarkably, the utility of NGS technologies has moved beyond the sequencing purposes; for instance, NGS platforms are being increasingly used in gene expression studies, where prior information about the sequence of a transcript is not required to detect its expression, in contrast to hybridization-based microarray platforms (Metzker, 2010).

Most NGS technologies, as well as the first-generation Sanger sequencing, are DNA polymerase-dependent, however, they differ in their template preparation, sequencing, imaging and data analysis steps (Metzker, 2010). A number of NGS plaforms are commercially available, among which Illumina/Solexa (www.illumina.com) and Roche/454 (www.454.com) platforms are currently the leading ones.

NGS technologies generally require clonal amplification of the template DNA to be sequenced, as most imaging systems are not capable of detecting single fluorescence or luminescence events. This amplification step introduces an amplification bias, in which certain sequences are replicated more than others, and may induce mutations during the amplification. Genome assemblies and sequence alignments, indeed, demonstrated an underrepresentation of AT-rich or GC-rich target sequences that are sequenced through Illumina/Solexa and Roche/454 technologies (Metzker, 2010). In order to overcome this issue, “third generation”, or sometimes referred to as “next-next-generation sequencing”, technologies are being developed, which act on single-molecule templates. However, these third generation technologies have not been widely applied in research programmes yet (Bolger *et al.*, 2014).

2.2.2.1. Illumina/Solexa platform

Solexa technology is commercialized by Illumina, hence, is generally referred to as the Illumina/Solexa platform. The Illumina/Solexa platform works on either single end or paired-end libraries, which are generated from the randomly sheared fragments of the template DNA. These sequencing libraries are then clonally amplified through solid-phase amplification, also known as the bridge amplification. Similar to the Sanger sequencing, Illumina/Solexa platform utilizes the chain terminator chemistry for DNA sequencing. Fluorescently labeled chain-terminating nucleotides are added to the growing DNA chain in a reversible manner (Metzker, 2010). The incorporation of reversible chain terminators, and thus, DNA synthesis is not highly efficient. Consequently, read lengths obtained by the Illumina/Solexa platform is generally shorter than the Roche/454 platform (Hutchison, 2007).

The most common error type in Illumina/Solexa generated sequences are base substitutions, particularly following the incorporation of a guanine base in the previous cycle (Metzker, 2010). Despite the shorter read lengths, Illumina/Solexa technology can provide better depth of coverage at reduced costs, compared to the Roche/454 platform (Metzker, 2010; You *et al.*, 2011).

2.2.2.2. Roche/454 Platform

Pyrosequencing, which forms the core of the Roche/454 platform, was first described in 1988 (Hutchison, 2007). This non-fluorescence technique relies on a number of sequential enzymatic reactions, which begins with the incorporation of a dNTP into the growing DNA chain and ends with the generation and detection of visible light (Fig. 2). The incorporation of a dNTP molecule by the action of **DNA polymerase** during a sequencing reaction releases an inorganic pyrophosphate molecule, which is converted to an ATP molecule by the **ATP sulfurylase** in the presence of Adenosine 5’ Phosphosulfate (APS). This ATP molecule is then used by **Luciferase** to convert lucferin into oxyluciferin, generating visible light in the process. The amount of generated light is proportional to the amount of ATP, which, in turn, is proportional to the initial amount of pyrophosphate molecules released, and thus, the amount of dNTPs incorporated. Finally, **Apyrase** removes all the unused dNTPs and ATPs to prevent cross-signals (Agah *et al.*, 2004).

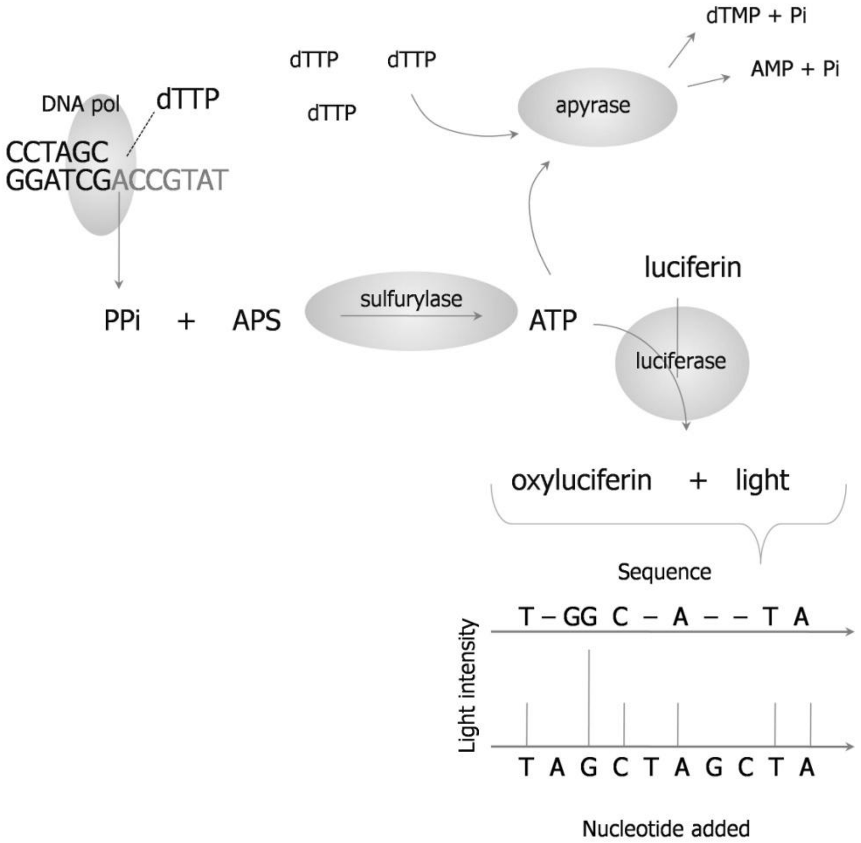


Figure 2. Pyrosequencing chemistry. DNA pol: DNA polymerase, sulfurylase: ATP sulfurylase, dTTP: Thymidine triphosphate, dTMP: Thymidine monophosphate, ATP: Adenosine triphosphate, AMP: Adenosine monophosphate, PPi: Pyrophosphate.

Margulies and his colleagues were the first ones to describe the use of pyrosequencing in an NGS system, which is commercialized by 454 Life Sciences (Roche Applied Sciences, Basel, Switzerland). In order to perform high-throughput DNA sequencing, the Roche/454 platform combines an emulsion-based method with pyrosequencing carried out inside picoliter-sized wells of a solid support. Prior to sequencing, the DNA template is randomly sheared to generate a sequencing library of small DNA fragments. Each DNA fragment is captured by a bead through the base-pairing of adapter sequences and clonally amplified in an oil-water emulsion, which is called emulsion PCR (emPCR). A Sequencing-By-Synthesis (SBS) reaction following pyrosequencing chemistry takes place within the picoliter-sized wells of a PicoTiterPlate (PTP) device, where each nucleotide flows through the pico-wells one at a time (Margulies, 2005).

In contrast to other sequencing approaches, pyrosequencing does not involve any chain termination. Instead, nucleotides are supplied at a defined and sequential order at limiting amounts (Hutchison, 2007). The limiting amounts of the dNTP supply, however, complicate the sequencing of long homopolymer repeats as it might lead to incomplete extension by the DNA polymerase (Metzker, 2005). Homopolymer repeats are the major sources of errors in Roche/454 platforms (Hutchison, 2007).

2.3. Crop genome sequencing

The first plant genome sequence was published in 2000, which was the 125 Mb-long genome of the model plant, *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000). Since the promises of a whole genome sequence offer for crop improvement have long been recognized, the first crop genome sequence, that of rice, was published soon after (Goff *et al.*, 2002; Yu *et al.*, 2002). Both of these studies relied on traditional Sanger sequencing of a minimal set of overlapping Bacterial Artificial Chromosomes (BACs).

Genome sequencing efforts basically proceed through two approaches. Initial genome sequencing projects adopted a clone-by-clone approach, as described above. This approach includes laborious cloning steps and requires physical mapping of the BAC clones to guide the sequence assembly, which may not cover the entire genome. A more recent approach is the Whole Genome Shotgun (WGS), which involves the direct sequencing of different sized fragments of the genome to be sequenced. WGS eliminates the need for the cloning and physical mapping steps at the cost of accuracy, particularly in repetitive regions (Jackson *et al.*, 2011).

Automated Sanger sequencing had been the method of choice for early plant sequencing projects (Metzker, 2010). Despite the long reads obtained by Sanger sequencing (up to 1 kb), this methodology is low-throughput and both time- and resource-intensive. Therefore, NGS technologies initially developed for re-sequencing purposes are being increasingly employed in *de novo* sequencing projects. NGS platforms circumvent tedious steps of Sanger sequencing, such as bacterial cloning, and also provide multiplexing options (Varshney *et al.*, 2009). NGS methodologies, however, typically suffer from short read lengths, which complicate the subsequent sequence assembly, usually resulting in fragmented assemblies (Bolger *et al.*, 2014).

Despite the advances in sequencing technologies and the decreasing costs, plant genome sequencing has been mostly limited to small genomes with low repetitive content (Fig. 3). Genome sizes in crop plants vary greatly in size. The ploidy level and repeat/transposable element content of the genome account for most of the variation in crop genome sizes (Feuillet *et al.*, 2011). These two factors constitute the major challenges in crop genome sequencing.

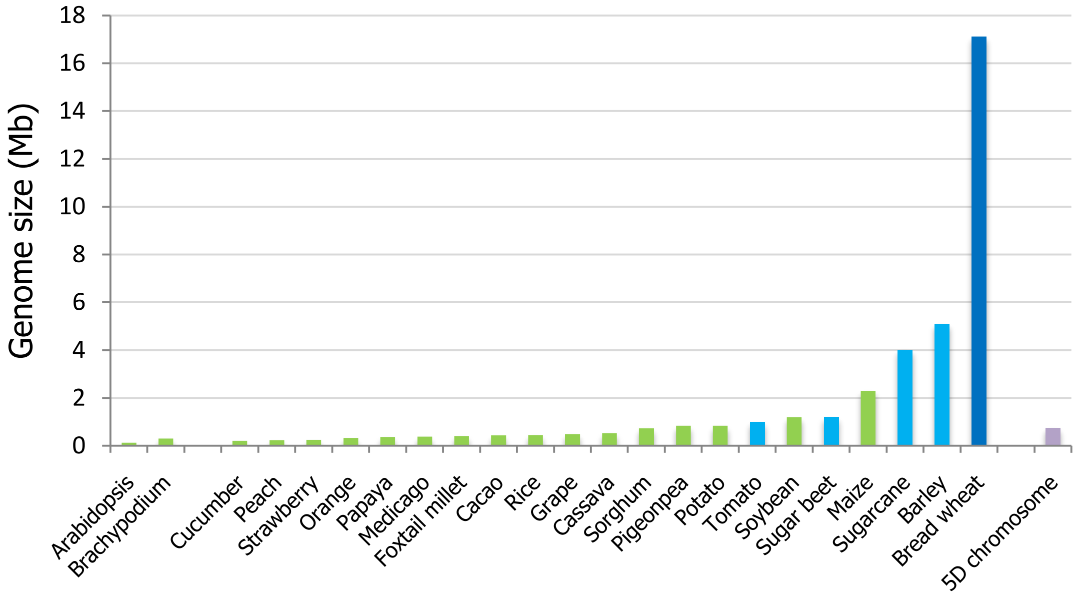


Figure 3. Status of crop genome sequencing. Green bars indicate a 'finished' genome sequence, while light blue bars indicate high-quality draft genome. Bread wheat genome indicated by a dark blue bar is at the draft status currently. The 748 Mb 5D chromosome is indicated by the purple bar at the right end. Adapted from Metzker *et al.*, 2010.

A high quality genome sequence has multiple uses. Genome sequences of model plants provide clues into plant biology, which can be used to identify similar genes, structural features or networks in economically important crops or to investigate evolutionary history through comparative analyses. In particular, genome sequencing in crops allows for the exploration and exploitation of the genetic diversity found within a germplasm. Structural variations, such as Copy Number Variation (CNV) or Presence-Absence Variations (PAVs) underlying the polymorphisms observed among individuals can be detected by mapping re-sequencing data on a high quality reference genome (Feuillet *et al.*, 2011). These polymorphisms can then be used to design several molecular markers assisting in map-based cloning of agronomically important traits or Marker-Assisted Selection (MAS) (Varshney *et al.*, 2009; Morrell *et al.*, 2011). Remarkably, comparative analysis of genome sequences may indicate Linkage Disequilibrium (LD) patterns of related genomes, which can be utilized to target the most efficient genome segments for introgression, that is, regions with low LD (Varshney *et al.*, 2009).

In the absence of a complete genome sequence, high-throughput and low coverage survey sequences are also capable of revealing certain aspects of the genome, thereby, offering a wide range of application areas, summarized below.

2.4. Applications of NGS technologies in crops

2.4.1. Gene expression and regulation

Prior to the introduction of NGS technologies, sequencing in crop species was already widespread, mostly in terms of transcript sequences. Sequencing of Expressed Sequence Tags (ESTs) even provided clues into important agronomic traits, such as drought tolerance (Ergen and Budak, 2009). Currently, over 6 million ESTs from four crops, maize, soybean, wheat and rice, are deposited in the EST database of National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/genbank/dbest, last accessed 22.01.2015). With the technical improvements and decreasing costs, NGS technologies are now beginning to dominate transcriptome profiling and gene expression studies.

RNA-Sequencing (RNA-Seq) utilizes deep sequencing through NGS platforms to identify and quantify transcripts of an organism expressed under certain conditions and offers several advantages over the traditional hybridization-based microarray platforms. Microarrays require prior knowledge of genome of transcriptome sequences in order to design hybridization probes, suffer from background noise due to cross-hybridizations, have a low dynamic range, involve complicated normalization steps and provide low resolution data. In contrast, RNA-Seq is able to provide *de novo* sequences, which is particularly important for species lacking extensive genomic or transcriptomic sequence data, at single base resolution with low to no background noise. As RNA-Seq does not impose an upper limit for quantification, transcripts with very low or very high expression levels can be detected (Wang *et al.*, 2009).

Targeting the expressed portion of the genome greatly reduces the complexity of genome, particularly those with high repetitive contents, such as wheat. In an experimental design, referred as 'exome capture', probes derived from expressed sequences are used to capture coding sequences, prior to sequencing. This approach is particularly useful for *Triticeae* genomes, where repeat contents usually exceed 80% (Smith and Flavell, 1975). Consequently, exome capture enables the sequencing of the protein-coding regions to provide much higher coverages (Winfield *et al.*, 2012). Additionally, aligning these exome sequences against transcriptomes of related species reveal SNPs, CNVs, duplications and deletions, which can be efficiently used for genotyping (Saintenac *et al.*, 2011; Wendler *et al.*, 2014). Remarkably, exome capture has been used to screen and detect chemically induced mutations F2 populations in both diploid rice and hexaploid wheat (Henry *et al.*, 2014). Exome capture may be a better alternative to RNA-Seq for detecting variations, as it enables targeted sequencing and enrichment for specific transcripts.

Recently, NGS technologies are integrated with chromatin immunoprecipitation (ChIP), a technique called as ChIP-Sequencing, to explore epigenetic modifications or DNA-protein interactions (Varshney *et al.*, 2009). Interestingly, RNA-Seq and ChIP-Seq were used in combination to reveal targets of the transcription factor, VRN1, which is involved in vernalization pathway, an important trait for *Triticeae* tribe (Deng *et al.*, 2015).

2.4.2. Molecular markers

Crop breeding depends on genetic diversity for crop improvement. One aspect of the widespread applications of NGS technologies with remarkable implications on breeding has been the design of molecular markers. Through sequencing and re-sequencing, polymorphisms, such as SNPs, CNVs and PAVs, as well as Insertion-Deletions (InDels) or Simple Sequence Repeats (SSRs) can be readily identified, from which numerous molecular markers can be designed (Feuillet *et al.*, 2011). Although SNP discovery is most efficient in the presence of a reference genome sequence, You and his colleagues have come up with a pipeline for SNP discovery without a reference genome sequence (You *et al.*, 2011). This approach utilizes relatively longer reads of one NGS platform (such as Roche/454) or sequence assemblies as reference sequences to map shorter reads of another platform (such as Illumina/Solexa). Resulting nucleotide differences are filtered against SNP proximity and depth, to avoid misidentification (You *et al.*, 2011). Recently, unique sequences flanking the insertion site of a Transposable Element (TE) were used to design Insertion Site-Based Polymorphism (ISBP) markers, which are particularly useful for crops with highly repetitive genomes, such as wheat and barley (Paux *et al.*, 2010). Additionally, variations within coding sequences, ESTs or conserved orthologous sequences can also be used to design gene-associated molecular markers (Quraishi *et al.*, 2009; Varshney *et al.*, 2009). Different types of molecular markers characterize different regions of the genome (coding or repetitive, for example). Thus, saturation of genetic maps with various types of molecular markers is crucial.

In general, molecular markers are utilized to explore genetic diversity in germplasm collections, identification of phlygenetic relationships to define cultivars, characterization of genetic resources and association mapping of agronomic traits (Edwards and Batley, 2010). An essential tool for modern breeding is Marker-Assisted Selection (MAS), the use of molecular markers tightly linked to traits of interest to track the trait through crosses. MAS is particularly useful for traits that are difficult to score, under complex genetic and/or environmental control, that manifest late in development or under particular conditions such as pathogen infection, or that exhibit low heritability (Akpinar *et al.*, 2013). Introgression of traits through interspecific crosses leads to the co-transfer of linked segments, which may have unprecedented, negative effects on crop performance, a phenomenon called as 'linkage drag'. A number of back-crosses are required to eliminate or minimize this linkage drag. Tightly linked molecular markers flanking both sides of the target gene can define the desired segment precisely, and thus, enable efficient transfer of the trait. Additionally, early selection of traits through the use of molecular markers, stable across environments and conditions, can considerably accelerate back-crossing steps (Edwards and Batley, 2010; Akpinar *et al.*, 2013).

Molecular markers also aid in Map-Based Cloning (MBC) of agronomically important traits. MBC constructs a high-density genetic map covering the chromosomal segment suspected to contain causal gene of a trait. This 'mini' map, integrated with a physical map, is then used for chromosome walking to eventually isolate the gene (Varshney *et al.*, 2006).

High-throughput sequences generated by NGS platforms provide an important source for the design and development of a variety of molecular markers that can be used to saturate the genetic maps and facilitate their integration with the physical maps or other genetic resources, which, in turn, can be used for MAS or MBC purposes in breeding programmes.

2.4.3. Comparative genomics and crop evolution

Cereal genomes exhibit a remarkable level of conservation, which allows researchers to the trace back and reconstitute the ancestral grass genome (Bolot *et al.*, 2009; Pont *et al.*, 2013; Murat *et al.*, 2014). Consequently, these related grass genomes share conserved blocks of genes which are colinear and are referred as 'syntenic' blocks (The International Brachypodium Initiative, 2010). The high conservation and syntenic relationships among grasses have contributed to the identification of conserved genes or chromosomal rearrangements from low-coverage NGS data in species lacking a reference genome sequence, such as wheat and barley (Mayer *et al.*, 2011; Wicker *et al.*, 2011; Vitulo *et al.*, 2011; Akpinar *et al.*, 2014; Lucas *et al.*, 2014). Notably, syntenic relationships and comparative genomics have provided the means to fine-map several important genes in species with limited genetic and genomic resources, such as the wild wheat germplasm, to access the genetic diversity maintained within (Zhang *et al.*, 2010; Wu *et al.*, 2013; Ouyang *et al.*, 2014; Wang *et al.*, 2014).

The draft sequences of all 21 chromosomes of bread wheat has been published very recently (Mayer *et al.*, 2014). These draft sequences provided valuable insights into wheat evolution. The D-genome of the modern bread wheat was revealed to result from an ancient hybridization between A and B genomes, which explains the observation that both A and B genomes are more similar to the D-genome than to each other, despite the relatively recent incorporation of the D-genome into the bread wheat genome (Marcussen *et al.*, 2014).

2.5. Wheat genome evolution

Extensive research on genome biology and evolution suggests that the *Poaceae* family of grasses, including cereals, co-evolved from a common ancestor, with 5 ancestral chromosomes, approximately 55-75 Mya (Gill *et al.*, 2004; Murat *et al.*, 2014). Reconstruction of the ancestral karyotpe indicates that modern genomes of major cereals, rice, wheat, barley, sorghum and maize, are variations of this ancestral genome through different chromosome breakage, fusion and duplication events. As a result, rice has 12 basic sets of chromosomes, while sorghum and maize each have 10 and the *Triticeae* tribe, including wheat and barley, has 7 basic sets of chromosomes, which share extensive homology (Salse, 2012). Cereals vary in ploidy levels (the presence of one or more genome copies, or sub-genomes), while barley is a diploid organism, wheat, from the same tribe, can be diploid, tetraploid (durum wheat) or hexaploid (bread wheat) (Feuillet *et al.*, 2007). Nevertheless, all cereals, and the majority of the grass species, are considered as diploidized paleopolyploids, due to the shared ancestral genome duplications (Murat *et al.*, 2014).

Wheat genome evolution had profound effects on the genome size and structure of modern wheat species. The modern bread wheat genome has been shaped by three hybridization and two Whole Genome Duplication (WGD) events (Marcussen *et al.*, 2014). Recent research suggests that approximately 6.5 million years ago (Mya) *Triticum* and *Aegilops* species diverged from their common ancestor, forming A and B genome lineages. The first hybridization event of the bread wheat evolution involved these two genome lineages ~5.5 Mya, giving rise to the D genome lineage (Marcussen *et al.*, 2014). The second hybridization event between *Triticum urartu* (AA genome) and an unknown relative of *Aegilops speltoides* from the Sitopsis section (BB genome) was followed by a WGD event, giving rise to the tetraploid *Triticum turgidum* (AABB genome). This species was domesticated, and several *T. turgidum* subspecies had been cultivated for thousands of years. Although most of these cultivars are no longer commercially produced, durum wheat, *T. turgidum* ssp. *durum*, is still an economically important crop (Feuillet *et al.*, 2007). Finally, a third hybridization event, dating back to only ~10.000 years ago, combined the tetraploid *T. turgidum* genome (AABB genome) and diploid *Aegilops tauschii* genome (DD genome, from the D lineage) and formed the hexaploid *Triticum aestivum,* modern bread wheat, genome, through the second WGD event (AABBDD genome). Consequently, modern bread wheat contains three related but divergent sub-genomes, which are organized into an ‘allohexaploid’ genome constitution (Fig. 4).

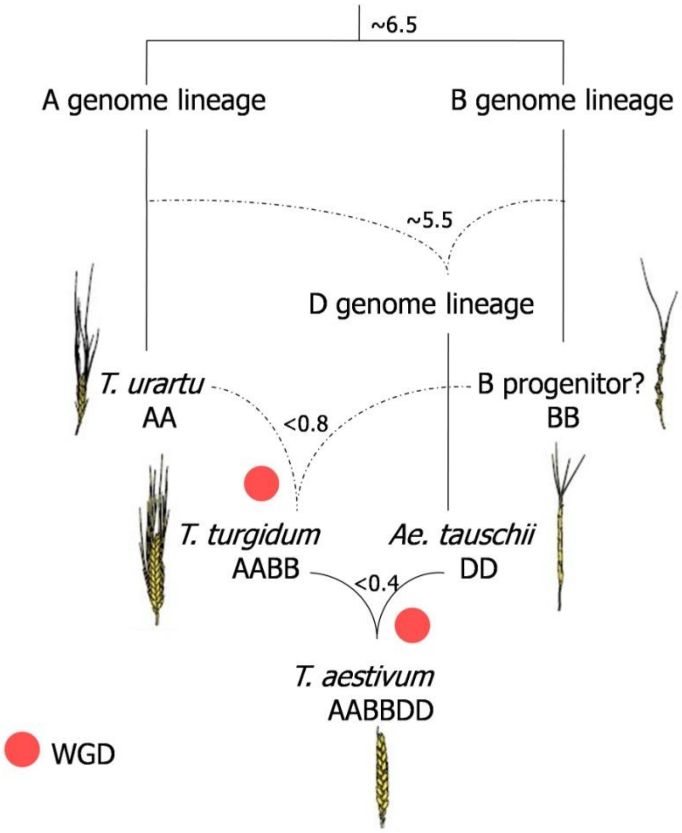


Figure 4. Recently proposed model for the genome evolution of bread wheat. Numbers denote estimated dates of the paleohistoric events in Mya. Whole Genome Duplication (WGD) events are indicated by red circles. Adapted from Marcussen *et al.*, 2014.

Paleohistory of the wheat genome suggests that bread wheat is a diploidized paleopolyploid; that is, despite being hexaploid, wheat acts as diploid. Diploidization refers to the elimination of duplicated gene redundancy in polyploid genomes, either at the structural level, through gene deletion, or, at the functional level, through neo- or sub-functionalization, pseudogenization and concerted evolution. Although diploidization efficiently turns duplicated gene copies back into singleton status, certain genes, such as Transcription Factors (TFs), are diploidization-resistant, and thus, are retained as paralogous copies (Murat *et al.*, 2014). Therefore, modern bread wheat genome is composed of diploidization-sensitive and diploidization-resistant blocks with numerous paralogous, related and pseudogenic loci across the entire genome (Pont *et al.*, 2013).

2.6. Wheat genomics

The hexaploid bread wheat genome is a grave challenge for genomics research. With a genome size of approximately 17 Gigabases (Gb), bread wheat genome is almost three times as large as the human genome (Mayer *et al.*, 2014). By nature, the allohexaploid genome contains several homeologous and paralogous loci within the related yet divergent sub-genomes. Additionally, *Triticeae* genomes have a marked abundance of repetitive elements, making up to >80% of the entire genome (Smith and Flavell, 1975), which complicates genome sequencing and subsequent assembly of the sequences. These attributes of its genome have long hindered genomics research on bread wheat, and achieving the sequencing of its huge and complex genome has been considered as practically impossible or highly unfeasible until very recently (Paux *et al.*, 2008).

While the isolation of individual chromosomes using flow cytometry has been reported four decades ago in hamsters and humans, flow cytometric sorting of plant chromosomes were complicated due to the low levels of metaphase synchronization and the presence of cell walls (Doležel *et al.*, 2012). Fortunately, advances in chromosome sorting techniques enabled isolation of individual chromosomes from plants by flow cytometry (Kubaláková *et al.*, 2002; Simková *et al.*, 2008; Safár *et al.*, 2010). Flow-cytometric sorting of chromosomes greatly reduces the genome complexity; rather than the entire genome, parts of the genome can be studied one at a time. Additionally, the use of flow-sorted chromosomes eliminates complicating homeologous and paralogous loci found elsewhere in the genome, thereby allowing the identification of chromosome-specific features. Consequently, the International Wheat Genome Sequencing Consortium (IWGSC), a collaborative platform of several research groups from public or private institutions, employed a chromosome-by-chromosome approach to tackle the daunting task of sequencing the bread wheat genome. In this approach, each chromosome is allocated to a specific research group for the ultimate goal of reference sequencing (Fig. 5) (http://www.wheatgenome.org/).

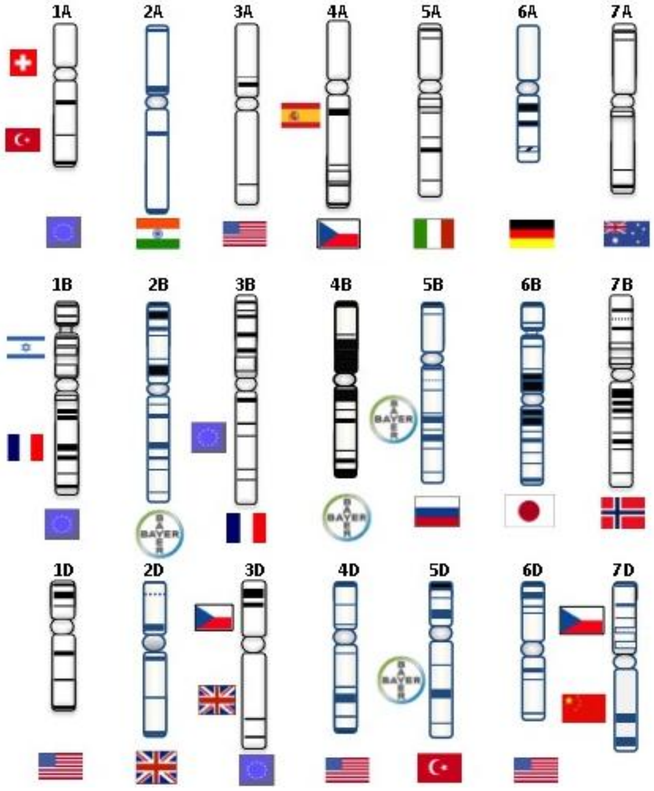


Figure 5. Bread wheat chromosomes allocated to research groups from different countries for the ultimate goal of reference sequencing (http://www.wheatgenome.org/).

The IWGSC approach to sequence the entire bread wheat genome to a reference quality involves sequencing each flow-sorted chromosome by the clone-by-clone sequencing strategy. This strategy includes the construction of the physical maps from BAC libraries of flow-sorted chromosomes (Fig. 6). As the first step, the isolated chromosome or chromosome arm is fragmented and cloned into BAC vectors to generate a chromosome-specific BAC library. These BAC clones are then fingerprinted using the high-throughput SNaPshot^{TM} High-Information Content Fingerprinting (HICF) procedure. Briefly, each BAC clone is digested with 4 rare cutters producing 3' overhangs, and a frequent cutter producing blunt ends. Different overhangs are labeled with four different fluorescent dyes, and restriction patterns (or fingerprints) are analyzed by capillary electrophoresis (Luo *et al.*, 2003). The fingerprints are compared through computational software to determine clone overlaps and build the BACs into a preliminary physical map. While traditionally FingerPrinted Contig software (Nelson *et al.*, 2005) was widely used to construct physical maps from BAC fingerprints, recently introduced Linear Topology Contig (LTC) software is increasingly employed in recent studies due to its ability to produce fewer and longer contigs and to allow evaluation of clone overlaps (Frenkel *et al.*, 2010). The next step following the physical map construction is selecting the 'Minimum Tiling Path (MTP)', that is the minimal set of overlapping BAC clones covering the entire physical map, which will be used for further refinement of the physical map and for clone-by-clone sequencing efforts. BAC-based physical maps serve as framework to guide the assembly of genomic sequences, and also present valuable sources for various applications, such as map-based gene cloning (Stein, 2007).

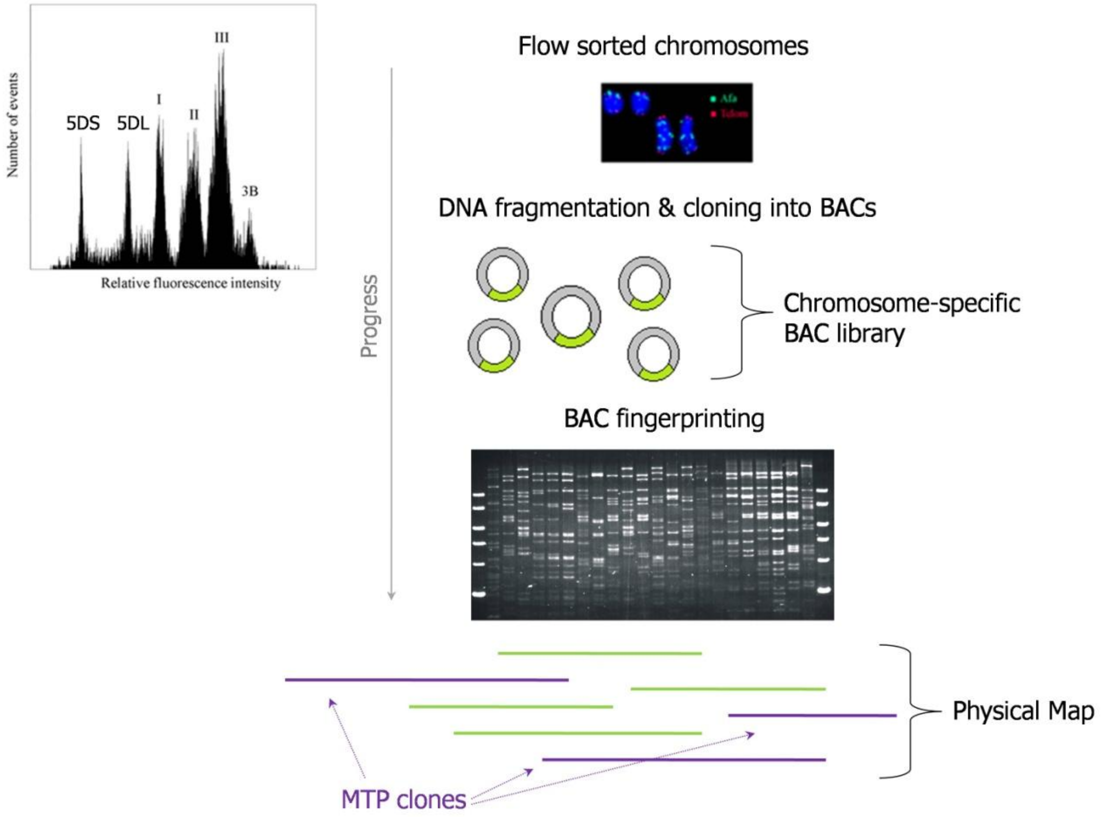


Figure 6. Schematic overview of the physical mapping of sorted chromosomes. A sample flow karyogram from which 5D chromosome arms and 3B chromosome, represented by single peaks, can be flow-sorted is given on top left. I, II, III on flow karyogram correspond to composite peaks that contains multiple chromosomes. Chromosomes are stained with DAPI (blue) and the purity of the sorted chromosomes are determined through the telomeric microsatellites (Kubaláková *et al.*, 2002).

The feasibility of this approach was first demonstrated on the longest bread wheat chromosome, 3B, which is 1 Gb long (Paux *et al.*, 2008). Chromosome 3B is the only chromosome among all 21 chromosomes of bread wheat, which can be sorted from the standard flow karyogram. Due to its remarkable size, 3B is represented as an individual peak, whereas other chromosomes are represented within three composite peaks (Fig. 6). Fortunately, the plasticity of the bread wheat genome capable of tolerating aneuploidy allowed for the construction of large cytogenetic stocks, which are used to isolate all remaining chromosomes and chromosome arms (Endo and Gill, 1996; Safár *et al.*, 2010).

Following the construction of 3B physical map (Paux *et al.*, 2008), five more physical map reports ensued, for chromosomes 1AL (Lucas *et al.*, 2013), 1AS (Breen *et al.*, 2013), 1BL (Philippe *et al.*, 2013), 1BS (Raats *et al.*, 2013) and 6A (Poursarebani *et al.*, 2014). While the physical maps of 1A, 1B and B chromosomes relied on SNaPshot^{TM} HICF (Luo *et al.*, 2003), the 6A physical map was constructed using a Whole Genome Profiling (WGP) approach.

While the physical mapping projects are in progress towards the ultimate goal of reference sequencing (http://www.wheatgenome.org/), bread wheat genome structure and organization have been under close scrutiny for the past few years through the use of NGS technologies. Initial attempts targeted NGS of selected BAC clones from chromosome 3B chromosome (Wicker *et al.*, 2011) or BAC-End Sequencing (BES) of chromosome-specific BAC libraries of chromosomes 3B, 1AL and 3AS (Paux *et al.*, 2006; Lucas *et al.*, 2012; Sehgal *et al.*, 2012). Additionally, survey sequencing of chromosomes 4A, 5A, 5D, 6B, 7BS and 7DS have also been published (Berkman *et al.*, 2011; Berkman *et al.*, 2012; Vitulo *et al.*, 2011; Hernandez *et al.*, 2012; Tanaka *et al.*, 2013; Lucas *et al.*, 2014). These sequences have been highly informative on the genome structure and organization of the bread wheat; presented a general view of the gene space, gene conservation, putative microRNA and tRNA encoding genes, repetitive landscape and comparative relationships with related grasses; and have been a rich source for the development of several molecular markers. The entire bread wheat genome has also been sequenced to a 5X coverage, the highest achieved at the time (Brenchley *et al.*, 2012). Besides bread wheat, the draft genome sequences of its two diploid progenitors, *T. urartu* and *Ae. tauschii*, were published, providing valuable insight into the A and D genomes, respectively (Ling *et al.*, 2013; Jia *et al.*, 2013). Finally, the draft genome sequences of all 21 bread wheat chromosomes and the reference sequencing of chromosome 3B have been reported very recently (Mayer *et al.*, 2014; Choulet *et al.*, 2014).

2.7. Chromosome 5D in an agronomical context

At a size of 748 Mb (258 Mb short arm, 5DS; 490 Mb long arm, 5DL), chromosome 5D is the second largest chromosome of the D-genome and constitutes approximately 4.4% of the entire bread wheat genome (Safár *et al.*, 2010). The 5D chromosome harbors a number of agronomically important genetic loci. Among these, *Pina-D1* and *Pinb-D1* alleles located within the *Ha* locus for grain **Ha**rdness are responsible for the grain texture, which determines the end-use-quality of wheat. The protein products of these alleles, puroindolines a and b, confer the grain a soft texture. The absence of either of these proteins, conversely, results in a hard texture, which is the case for durum wheat. The *Ha* locus is located on the short arm of chromosome 5D (5DS, hereafter) (Morris, 2002). Both chromosome arms also carry *Pro1* and *Pro2* genes that are related to the protein content of the grain (Mcintosh *et al.*, 2008). The long arm of the 5D chromosome, 5DL hereafter, is attributed to at least 2 vernalization loci, *Vrn-D1* and *Vrn-D4*. Vernalization, exposure to low temperatures before germination, affects the flowering time in wheat, and the length of the vernalization required varies among the genetic stocks (Mcintosh *et al.*, 2008; Yoshida *et al.*, 2010; Zhang *et al.*, 2012). Additionally, *Lr1* gene mapped to the 5DL provides resistance against the leaf rust disease, causing major yield losses globally (Cloutier *et al.*, 2007). A few additional loci are mapped to both 5DS and 5DL, conferring resistance against different *Puccinia* strains (such as *Lr57*, *Yr40*, and *Sr30*) or *Blumeria graminis* (such as *Pm2*, *Pm4*, and *Pm35*), the causal agent of powdery mildew; however, only *Lr1* gene could be cloned to date (Mcintosh *et al.*, 2008). Considering the arms race with the pathogen evolution and the disease response, cloning and characterization of further loci related to biotic stress will remain an important issue.

3. MATERIALS and METHODS

3.1. Isolation of 5D chromosome by flow cytometry

The seeds for the double ditelosomic 5D line (2n=40+2t5DS+2t5DL) of *Triticum aestivum* L. cv. Chinese Spring were kindly provided by Prof. B.S. Gill (Kansas State University, Manhattan, USA). Liquid suspensions of intact mitotic chromosomes from synchronized root tips of young seedlings were used to sort short and long arms of chromosome 5D as described by Vrána *et al*. (Vrana *et al.*, 2000). The purities of the sorted fractions were determined by Fluorescence *In Situ* Hybridization (FISH) with probes for *Afa* and telomeric repeats. Briefly, three batches of 1000 chromosomes were sorted onto microscopic slide into 10µl drop of PRINS buffer supplemented with 2.5% sucrose. The sample was air-dried and sorted chromosomes were analyzed with FISH (Kubaláková *et al.*, 2002; Hernandez *et al.*, 2012).

3.2. Next-Generation Sequencing of 5D chromosome

Prior to sequencing, chromosomal DNA was purified from chromosome arms sorted in 40µl deionized water and subsequently amplified by Multiple Displacement Amplification (MDA) using the illustra GenomiPhi DNA Amplification kit (GE Healthcare, Chalfont St. Giles, United Kingdom) as reported previously (Simková *et al.*, 2008).

Next-generation sequencing of sorted chromosome arms were carried out on GS FLX Titanium platform (454 Life Sciences, Roche Diagnostics Corporation, Indianapolis, IN, USA) as outlined in Lucas *et al.* (2014). Shotgun sequencing libraries were prepared using the GS FLX Titanium Rapid Library Kit (Product no. 05608228001, 454 Life Sciences). The library quantification was performed on Agilent 2100 Bioanalyzer using the High Sensitivity DNA Analysis Kit (Agilent Technologies, Santa Clara CA, USA). Enrichment, amplification and sequencing steps were performed using GS FLX Titanium emPCR (Product no. 05618428001, 454 Life Sciences) and Sequencing Kits (Product no. 05233526001, 454 Life Sciences).

The experimental procedures were performed following the manufacturer's instructions.

3.3. Identification and characterization of repetitive elements

The repetitive elements were identified using RepeatMasker version 3.3.0 (http://www.repeatmasker.org/) with a custom repeat database made up of *Triticeae* repeat sequences from TREP release 10 (http://wheat.pw.usda.gov/ITMI/Repeats) with Repbase Update release 15.11 (Jurka *et al.*, 2005) and TIGR Plant Repeat Databases (Ouyang and Buell, 2004), which indicated that each chromosome arm contained >70% repetitive elements.

Additionally, 454 sequences were assembled using gsAssembler tool of the Newbler software v2.6 (454 Life Sciences). The assembly was carried out at default values with 'large and complex genome' and 'heterozygotic mode' options and the empirically determined minimum overlap identity of 95%. Assembled sequences revealed that contigs with low depths had <70% of known repeat content, while the repeat content rised to over 80% with contig depths 3 to 6. Therefore, all contigs with a depth of 5 or more were considered as collapsed repeats of unknown type, based on the above observations and the average sequence coverage of 1.34-1.61x. All sequences from these high depth contigs, together with the sequences masked by RepeatMasker, were excluded from subsequent analyses.

For comparative analyses with *Ae. tauschii* 5D data, unmasked reads were masked against a more recent repeat element database, MIPS Repeat Element Database (v. 9.3) for *Poaceae* (ftp://ftpmips.helmholtz-muenchen.de/plants/REdat/), using the RepeatMasker software (Akpinar *et al.*, 2014).

3.4. Genetic marker, gene, protein and assembled transcript sources

Cytogenetic map positions of 5D-mapped EST and SSR markers were retrieved from URGI Genetic and Genomic Information Center (GnpMap, map name: DEL_050308). EST sequences were retrieved from GrainGenes (http://wheat.pw.usda.gov/wEST/), while SSR and COS sequences were kindly provided by P. Sourdille and J. Salse, respectively.

Proteome annotations of fully sequenced grass genomes were retrieved from the following sources: *Brachypodium distachyon* genome annotation v1.2 (The International Brachypodium Initiative, 2010) and *Sorghum bicolor* genome assembly v1.4 (Paterson *et al.*, 2009) from MIPS PlantsDB (http://mips.helmholtz-muenchen.de/plant/genomes.jsp); *Oryza sativa* genome assembly IRGSP-1.0 (Tanaka *et al.*, 2008) from The Rice Annotation Project Database (http://rapdb.dna.affrc.go.jp/download/irgsp1.html); UniGene sequences for *Triticum aestivum*, *Hordeum vulgare*, *Panicum virgatum*, *Saccharum officinarum*, and *Zea mays* from NCBI UniGene Repository (ftp://ftp.ncbi.nih.gov/repository/UniGene/); UniProt sequences (The UniProt Consortium, 2012) from UniProt KnowledgeBase (http://www.uniprot.org/). The 2.2x coverage *Ae. tauschii* 5D chromosome survey sequences were retrieved from a recent study (Akpinar *et al.*, 2014).

3.5. Sequence similarity searches and gene modeling

In order to eliminate organellar DNA contaminations, non-repetitive 5D survey sequence reads were compared against *T. aestivum* mitochondrial and chloroplast genome sequences and all hits with ≥95% identity over ≥75% of the read length were discarded.

Similarity searches against the annotated proteins of *B. distachyon*, *O. sativa* and *S. bicolor* and UniProt sequences were performed using blastx and tblastn (-evalue 1E-6, -length 30, -ppos 75). Alignments with at least 75% similarity over 30 amino acids were considered significant at an e-value cutoff of 10^{-6} (Vitulo *et al.*, 2011). Only best reciprocal hits on blastx and tblastn searches were retained. For UniGene sequences, blastn searches (-evalue 1E-30, -length 90, -pident 75) were performed and only the best hits with at least 75% sequence identity over 90 nucleotides at an e-value cutoff of 10^{-30} were retained, except for *T. aestivum* UniGenes, where the sequence identity cutoff was raised to 95%. EST, SSR and gene-based marker sequences were identified using blastn (-evalue 1E-30, -length 90, -pident 95).

For all blast searches, redundant 5D sequence reads covering the exact same region on a protein or gene were removed to eliminate amplification bias (Wicker *et al.*, 2011). BLAST+ stand-alone toolkit, version 2.2.25 was utilized for all blast searches (Camacho *et al.*, 2009).

All positive blast hits from the grass genomes and UniGenes (the order of precedence: *Brachypodium*, rice, sorghum, UniGene) were used as references onto which all non-repetitive 5D reads were mapped using gsMapper tool of the Newbler software v2.6, with auto trimming on and a minimum overlap of 40 nucleotides (454 Life Sciences). Multiple sequence reads mapped on the same reference sequence were merged by filling non-aligned parts with strings of ‘N’ using an in-house Perl script. For UniProt hits that could not be associated with any other grass genes or UniGene sequences, matching reads were *de novo* assembled using gsAssembler tool Newbler software v2.6 (454 Life Sciences), as no reference DNA sequence could be obtained for mapping.

3.6. Visualization and annotation of genes

Genomic positions of annotated genes of model grass genomes were retrieved from MIPS PlantsDB (http://mips.helmholtz-muenchen.de/plant/genomes.jsp). Gene conservation patterns were visualized on heatmaps constructed in MATLAB R2010b with a sliding window approach of 50 kb step size. Circle plots were generated using Circos software, utilizing *binlinks* and *bundlelinks* (≥50 membership along 1Mb intervals) tools (Krzywinski *et al.*, 2009). The virtual gene order was constructed using the 'genome zipper' approach as described previously (Mayer *et al.*, 2009).

Gene Ontology (GO) annotations of the gene models were performed using Blast2GO software (Conesa and Götz, 2008). Initial blast step was run locally against all non-redundant *Viridiplantae* proteins (-evalue 1E-6, -outfmt 5, -max_target seqs 1). Blast results generated as .xml files were imported into the Blast2GO Software, where mapping, annotation and GO Slim steps were performed at default values for plants. Multilevel charts were generated for Biological Process, Cellular Component and Molecular Function terms. Fisher's exact test (two-tailed) was used to evaluate statistically significant differences, among the annotations for a given term, between conserved gene models and non-conserved gene-related sequences, compared to the total number of remaining annotations in the same category.

Putative tRNA genes were predicted using tRNAscan-SE software (Lowe and Eddy, 1997). The program was run locally at the default parameters for eukaryotic genomes.

As an exception, in order to reliably compare functional gene spaces of *T. aestivum* 5D chromosome and its wild progenitor, *Ae. tauschii* 5D chromosome, a different pipeline was followed. First, NGS sequences from both chromosomes were masked against the most recent and comprehensive repeat element database for *Poaceae* family of grasses, MIPS Repeat Element Database (v. 9.3) for *Poaceae* (ftp://ftpmips.helmholtz-muenchen.de/plants/REdat/), using the RepeatMasker software. Additionally, these sequences were compared against *Ae. tauschii* chloroplast genome (GenBank: JQ754651.1), *T. aestivum* chloroplast genome (GenBank: KC912694.1), *T. aestivum* mitochondrial genome (NCBI: NC_007579.1) (1E-15, -dust 'no') and all *Triticum* rRNA sequences (1E-5, -dust 'no') to eliminate organelle-associated sequences. To avoid redundancy, remaining non-repetitive sequences were assembled gsAssembler tool Newbler software v2.6 (454 Life Sciences) with the following parameters: large and complex genome, heterozygotic mode, extend low-depth overlaps and 98% minimum overlap identity. The sequence assemblies were compared against the fully annotated model grass genomes as detailed above and, also, against the high-confidence barley proteins retrieved from MIPS PlantsDB (http://mips.helmholtz-muenchen.de/plant/barley/). For barley high-confidence proteins, blast parameter for similarity was raised to 90%. Additionally, UniProt sequences of *Ae. tauschii* (http://www.uniprot.org/, a total of 34.639 sequences, last accessed on 15.09.2014) and UniGene sequences from *T. aestivum* (Build #63) were used for the identification of putative wheat-specific genes. For UniProt sequences, 96% and 98% similarity parameters were applied for filtering for *T. aestivum* and *Ae. tauschii* 5D sequences, respectively. Conversely, for UniGene sequences, 98% and 96% identity parameters were used for *T. aestivum* and *Ae. tauschii* 5D sequences, respectively (-evalue 1E-30, -length 100). All sequences from both assemblies that were associated with any of the above mentioned protein or UniGene/UniProt sequences were annotated using Blast2GO as detailed above.

3.7. 5DS-specific BAC library construction, fingerprinting and assembly

The short arm of chromosome 5D was flow sorted as described in Section 3.1. A total of 8,120,000 sorted chromosome arms were then embedded in agarose miniplugs, and the 5DS chromosome-specific BAC library was constructed according to Šimková *et al*. (Simková *et al.*, 2011). The 5DS-specific BAC library was composed of 36,864 BAC clones with an average insert size of 137 kb, giving 17x coverage of the 258 Mb-long chromosome arm (Safár *et al.*, 2010). This library was designated as TaaCsp5DShA.

For fingerprinting, 26,112 BAC clones, with an average insert size 143 kb, representing 12.5x coverage of the chromosome arm, were selected. These clones were fingerprinted using SNaPshot^{TM} High-Information Content Fingerprinting (HICF) procedure (Luo *et al.*, 2003). Prior to preliminary map construction, BAC fingerprints were processed using the FingerPrint Background removal (FPB) software (Scalabrin *et al.*, 2009) to eliminate the following: (1) bands derived from either the vector or the host gDNA, (2) bands generated by incomplete digestion or star activity, (3) bands of unexpected sizes, (4) background noise. Parameters used in FPB were as follows: Tolerance=0.4; Peak width=15; Size=50-500; Multiply factor=30; Min bands=40; Max sizes=250. True bands of 50-500 bp range were further analyzed with the GenoProfiler software to remove cross-contaminations and negative controls (You *et al.*, 2007).

3.8. Preliminary map construction

After processing with FPB and GenoProfiler, a total of 21,656 good-quality fingerprints were obtained. These fingerprints were used to construct two separate preliminary maps using FingerPrintedContig (FPC) and Linear Topology Contig (LTC) softwares.

FPC assembly was carried out using parameters optimized for complex and repetitive genomes (Nelson *et al.*, 2005). Initial build of contigs was performed under extremely stringent conditions, at a Sulston Score probability cutoff of $1e^{-75}$. The stringency was then gradually decreased by -5 in each step, until $1e^{-45}$, to incrementally extend the core of high-confidence contigs. Manual merging was performed by comparing contig ends at a more relaxed cutoff of $1e^{-25}$, according to the following criteria: (1) a unique and reciprocal relationship exists between the contig ends, (2) two clones from the end of each contig have significant matches at this stringency OR a single clone match is supported by marker data. Consensus Band (CB) map is calculated for the putative merged contig when the ends of two contigs are considered for merging. If the CB map reveals >10% questionable clones (Q-clones) or any other structural aberrations, merge is rejected. If a pair of contigs share the same molecular marker, these contigs are merged regardless of the presence of matching clones. Short contigs that contain 6 or less clones or that are smaller than 200kb are discarded as these contigs are considered uninformative. The FPC assembly constructed as detailed above contained 350 contigs with an N50 of 1141 kb.

LTC assembly was carried out as previously described (Lucas *et al.*, 2013). The initial net of significant clone overlaps was generated at a relatively liberal cutoff of 10^{-15} (same Sulston Scoring scheme). From this net of clone overlaps, Q-clones and Q-overlaps were eliminated at cutoffs of 10^{-15} and 10^{-25}, respectively. The first round of adaptive clustering was performed at the cutoff of 10^{-15}, and the stringency was increased by decreasing the cutoff to 10^{-33} at 6 consecutive steps, in order to split non-linear contigs. Persistent non-linear contigs were inspected individually by visualizing the net of clone overlaps. Thirteen clones suspected of causing the branching in non-linear contigs were identified and excluded from the second round of adaptive clustering. The second round of adaptive clustering resulted in 164 contigs, of which 44 were short contigs with <6 clones.

3.9. Minimum tiling path selection and BAC pooling strategy

The minimum tiling path (MTP) of the LTC assembly contained several buried clones. Therefore, the LTC map was imported into the FPC program and a new MTP was picked with the following parameters and a preference for large clones: Minimum overlap=30, Maximum overlap=250, FromEnd=0, Minimum shared bands=12. Clone overlaps of the MTP picked by FPC were evaluated using LTC. By definition, overlaps statistically significant under conditions less stringent than the conditions used to build the initial net of clone overlaps are considered unreliable. All such overlaps were supported by the manual addition of 210 clones to the MTP clones that cover potentially unreliable overlaps. Finally, a total of 163 Q-clones were added to the MTP, as these clones might act as bridges if supported by molecular markers. Overall, manually edited MTP representing the 5DS physical map contained 2528 clones.

The MTP clones were re-gridded on 7 x 384-well plates. A 3D pooling strategy was applied to facilitate MTP screening. BAC clones on each row, column and plate were pooled together, giving rise to 16 row, 24 column and 20 plate pools. Additionally, all clones were also combined into a single superpool, for positive control purposes. BAC DNA was isolated from each pool and amplified, then re-organized into one 96-well plate. DNA amplification was performed by the MDA method using random primers and phi29 DNA polymerase (GenomiPhi V2 DNA polymerase Kit, GE Healthcare). Pools were diluted 1:200 in PCR Grade water before screening.

3.10. MTP screening using molecular markers

MTP clones were screened using a variety of molecular markers. 2 gene-based markers, 63 EST, 23 SSR and 13 COS markers were retrieved as described in Section 3.4. Additionally, 16,727 high-confidence ISBP markers were designed from 1.34x coverage 5DS survey sequences using IsbpFinder.pl and IsbpSort.pl scripts (Paux *et al.*, 2010). Of these, 99 high-confidence ISBP markers were tested on MTP pools.

Screening of MTP pools was performed in a 10µl PCR reaction volume, using standard Taq polymerase (Fermentas) as follows: 1µl 10X KCl Buffer (-$MgCl_2$), 0.8µl 25 mM $MgCl_2$, 0.2µl 2.5mM each dNTP, 0.25µl 10µM Forward Primer, 0.25µl 10µM Reverse Primer, 1µl 1:200 diluted BAC pools, 0.05µl Taq Polymerase, 6.45µl dH_2O. Reaction conditions were as follows: Initial denaturation, 94^{o}C 5 min, 35 cycles of {Denaturation, 94^{o}C 30 sec, Annealing, variable 30 sec, Extension 72^{o}C 30 sec}, Final extension 72^{o}C 7 min. PCR products were analyzed on 1% agarose gel, run at 100V for 15 minutes. Multiple hits in row, column and/or plate pools were resolved through colony PCRs on original MTP clones, testing all possibilities.

3.11. Microarray design and hybridization

Three sources of sequences were used to design probes for an Agilent SurePrint G3 Gene Expression Custom Microarray, 8x60k format (Agilent Technologies). These were: 1) Genetically mapped gene/marker sequences, 2) Conserved 5DS sequence reads, 3) ISBP markers designed from 5DS survey sequences. For the first group of sequences, 7 genes mapped on 5DS (Pina-D1, Pinb-D1, Gsp-1, MdH-D3, Nor-D3, Pro2, 5S-RNA-D2), 13 COS markers, 122 EST markers and 20 SSR markers (a total of 162 gene/marker sequences), in addition to 3 SNP sequences mapped to 5DS (Allen *et al.*, 2013), were used to design probes with a Tm matching methodology, with the parameters as follows: probe length=60bp, probes per target=5, preferred probe Tm 85^{o}C. Additionally, 109 SNPs mapped to 5D by Illumina sequencing (Poland *et al.*, 2012) were included within this group; however, these sequences were too short to design probes, thus used directly used in 5 exact copies in the overall design. For the second group of sequences, 6,996 conserved gene associated 5DS survey sequences, identified by blast searches against fully annotated grass genomes as described in Section 3.5, were used to design probes with the same criteria given above. For the remaining features on the array, the third group of sequences made up of 5,120 ISBP marker sequences (amplicon size >150bp) designed from 5DS survey sequences were used. Probes were designed as explained above. Overall, the final design included 1,370 probes for genetically mapped genes/markers, 34,980 probes derived from conserved gene reads, and 25,600 probes for ISBP markers.

SureTag DNA Labeling Kit (Agilent Technologies, Cat. No. 5190-3400) was used to label MTP pools with Cy3 and Cy5, in a dye-swap design, using the following the manufacturer's instructions. Each pool was labeled, hybridized and analyzed separately using both Cy3 and Cy5 to detect reproducible results. Hybridization and wash steps were performed as instructed by the manufacturer, and NimbleGen MS 200 microarray scanner (Roche NimbleGen, Inc.) was used to scan the arrays with at 2 nm resolution with autogain. Fluorescence data from the scanned images was extracted using Agilent Feature Extraction Software (v. 11.5.1.1). Data normalization and deconvolution were performed independently for row, column and plate pools, as previously described, using slightly modified custom R scripts (Rustenholz *et al.*, 2010; Lucas *et al.*, 2013).

Among the 5 probes designed from each query sequence detailed above, outliers within each pool were discarded. Normalization was then carried out for each pool with respect to each other by subtracting the median and dividing by the standard deviation of all signal intensities within the pool. Two complementary statistical methods were utilized to identify positives within each pool type, on the normalized signal intensity data. In the first method, if the median intensity for all the probes for a single query in a single pool exceeded [Mean + C x Standard Deviation] of the intensities for that query across all the pools, for a given pool type, the signal was scored as positive. C value indicates a pre-defined threshold co-efficient and this value was determined separately for each pool type. High confidence C values were set as 2.8, 1.6 and 2.6 for column, plate and row pools, respectively. In the second method, Student's t-Test was used to decide, assuming equal variance, whether the intensities for a given query in one pool were significantly different from all the other pools at p-value<0.01. Gene associated sequences or markers, used to design microarray probes, were putatively assigned to specific BACs, only if they passed both statistical tests. For queries assigned to multiple pools passing both tests, only those found on overlapping BAC clones of the preliminary map were considered true positives.

3.12. Supercontig construction and contig elongation

The latest version of the LTC introduces a new feature to elongate contigs into supercontigs (Breen *et al.*, 2013). The net of clone networks for the clones located at the ends of each contig of the 5DS preliminary map was tested possible overlaps with other contigs at the cutoff of 10^{-15}. Elongations were accepted if they are detected reciprocally (if A elongates into B, B should elongate into A).

To aid in contig ordering, deletion bin mapping of ISBP markers were performed on homozygous deletion lines of 5DS, 5DS-2 and 5DS-5. These deletion lines contain 0.78 and 0.67 of the full length chromosome arm, respectively. Leaf tissues from 4-week old seedlings were frozen in liquid nitrogen. Genomic DNA (gDNA) was isolated using Wizard® Genomic DNA Purification Kit (Promega, Madison, WI, USA) from 200 mg frozen tissue according to the manufacturer’s instructions.

Deletion bin mapping was performed in a 10µl PCR reaction volume, using standard Taq polymerase (Fermentas) as follows: 1µl 10X KCl Buffer (-$MgCl_2$), 0.8µl 25 mM $MgCl_2$, 0.2µl 2.5mM each dNTP, 0.25µl 10µM Forward Primer, 0.25µl 10µM Reverse Primer, 1µl deletion line gDNA, 0.05µl Taq Polymerase, 6.45µl dH_2O. Reaction conditions were as follows: Initial denaturation, 94^{o}C 5 min, 35 cycles of {Denaturation, 94^{o}C 30 sec, Annealing, variable 30 sec, Extension 72^{o}C 30 sec}, Final extension 72^{o}C 7 min. PCR products were analyzed on 1% agarose gel, run at 100V for 15 minutes.

Genetic map positions of genetically mapped EST, SSR, COS markers and deletion-bin mapped ISBP markers, assigned to BAC clones from the 5DS physical map through PCR or microarray, were used to order 5DS contigs along the chromosome arm. Contigs assigned to each deletion bin were ordered within the bin using the order of orthologous sequences from the recently published physical map of the *Aegilops tauschii* genome (Luo *et al.*, 2013). Orthologous *Aegilops* sequences were identified through similarity searches against positive probe sequences using blastn (-evalue 1E-10) as described in Section 3.5. For contigs that could not be associated with an *Ae. tauschii* ortholog, the order of the *Brachypodium* orthologs on our 5DS genome zipper was used.

4. RESULTS

4.1. Survey sequencing of 5D chromosome using 454/Roche platform

4.1.1. Flow sorting, sequencing and repeatmasking of 5D chromosome arms

The double ditelosomic 5D line (2n = 40 + 2t5DS +2t5DL) of *Triticum aestivum* L. var. Chinese Spring was used to isolate short and long arms of 5D chromosome (5DS and 5DL, hereafter) from flow karyograms of DAPI-stained mitotic chromosomes, at 90.18% and 85.5% for purities, respectively (Fig. 7).

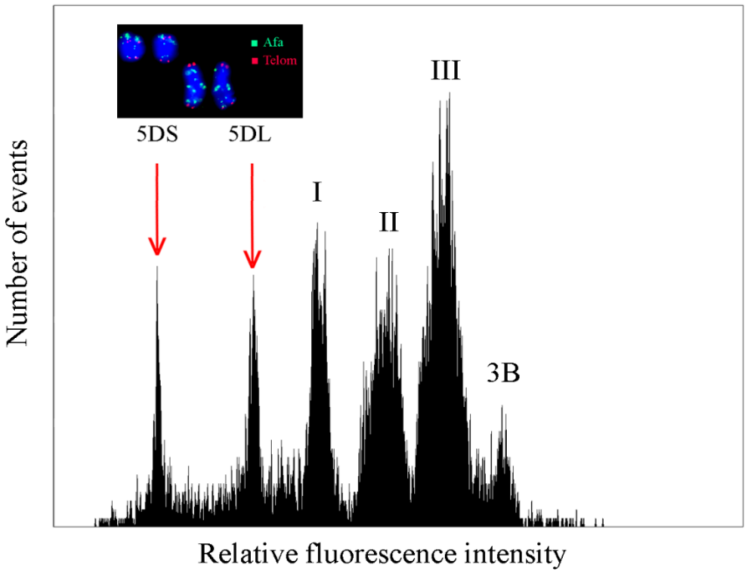


Figure 7. Flow karyogram of double ditelosomic line 5D of *T. aestivum* cv. Chinese Spring. Peaks representing telocentric 5DS and 5DL chromosomes are indicated. Inset: Flow-sorted 5DS stained by DAPI (blue). Afa repeats (green) and telomeric repeats (red) are labeled with FISH. x-axis: Relative fluorescence intensity, y-axis: Number of particles.

The flow-sorted telosomes were subsequently amplified by Multiple Displacement Amplification (MDA) yielding 15.81 µg (5DS) and 9.64 µg (5DL) DNA, as isolating sufficient amounts of DNA by flow-cytometry is prohibitively resource and time-intensive. Resulting amplified DNA was directly sequenced using GS FLX Titanium system of Roche/454 Platform (454 Life Sciences, Roche Applied Sciences, Basel, Switzerland). Good quality reads of 791 Mb and 347 Mb cumulative lengths were obtained for 5DS and 5DL, respectively, corresponding to 1.34x and 1.61x coverages (Table 1).

Table 1. Summary of sequencing data for 5D telosomes.

Arm	**Size[1]** **S** **(Mb)**	**No. of reads** **N**	**Mean read length** **L** **(bases)**	**Total read length** **(Mb)**	**Coverage[1]**	**Purity**	**Representation probability[2]**
5DL	490	2,271,366	347.25	791	1.61x	85.5%	0.684
5DS	258	937,264	370.28	347	1.34x	90.2%	0.667

[1] Calculated based on cytogenetic chromosome arm length estimates (Safár *et al.*, 2010)
[2] Calculated as: $P = [1 – (1 - L/S)^{N}]$ x Purity

The *Triticeae* tribe genomes are marked by high repetitive content of their genomes, exceeding 80% in most cases (Smith and Flavell, 1975), which interfere with the subsequent sequence assembly procedures. Therefore, the 1.34x and 1.61x survey sequences were masked against a custom repeat database, which included the TREP Database release 10 (http://wheat.pw.usda.gov/ITMI/Repeats) with the Repbase Update release 15.11 (Jurka *et al.*, 2005) and the TIGR Plant Repeat Database (Ouyang and Buell, 2004). Additionally, sequences collapsing into high depth contigs when assembled were eliminated as potential repeat elements of unknown type. Repeat masking of 5DS and 5DL survey sequences revealed that approximately 76% and 75% of 5DS and 5DL sequences, respectively, were comprised of repetitive elements. As expected, LTR retrotransposons were the most abundant type of repeat elements, making up over three-fourths of all repeat annotations. Repetitive element distributions of major repeat superfamilies were similar for both chromosome arms. However, CACTA superfamily of DNA transposons appeared to be more abundant among 5DS sequences, compared to 5DL, in contrast to the LTR retrotransposons (Fig. 8). A summary of all repeat annotations classified by repeat superfamilies is given in Appendix A.

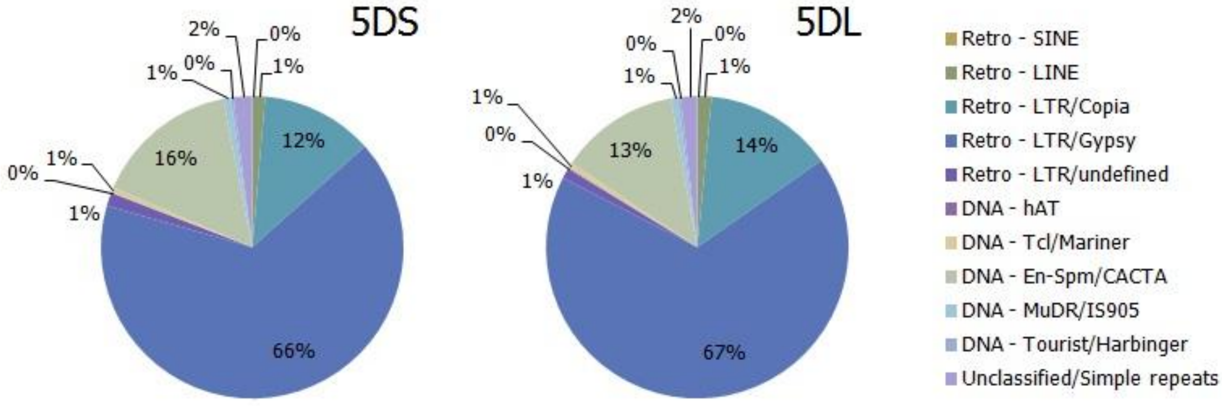


Figure 8. Distribution of repeat elements of 5DS and 5DL classified by superfamily.

The remaining non-repetitive sequences made up a total of 84.6 Mb for 5DS and 201 Mb for 5DL. These non-repetitive sequences for examined for the presence of seven genes previously cloned from the 5D chromosome (Table 2).

Table 2. Previously cloned genes identified in masked 5D sequences.

Gene	No. of matching reads	Total length	Matched length	Coverage (%)	Average depth
			5DS		
Pina-D1	6	447	447	100.00	2.48
Pinb-D1	6	828	447	53.99	2.29
Nor-D3*	165	887	887	100.00	50.54
5S-RNA-D2*	71	486	486	100.00	32.59
			5DL		
Vrn-D1	9	980	833	85.00	2.16
ADH1D	3	1140	235	20.61	1.60
VrnD3	1	1100	135	12.27	1.00
Lr1	90	4035	3958	98.09	7.00

*Repetitive reads were included in searches for these sequences.

The agronomically important *Pina-D1*, *Pinb-D1*, *VrnD1* and *Lr1* genes were evenly covered to 50-100% coverage of the respective gene, while *ADH1D* and *VrnD3* genes were only partially covered by the 5D sequence reads. Interestingly, the coverage of the *Lr1* gene was higher than expected, which may indicate the presence of multiple genes with high sequence similarity on 5D, assuming minimal to none amplification bias. rRNA genes usually exist in multiple copies and, therefore, may be misidentified as repetitive sequences. Unmasked 5D reads revealed the presence of *5S-RNA-D2* gene and the Nucleolus Organizing Region, *Nor-D3*, at high depths of coverage, although these were elusive among the masked 5D sequences.

4.1.2. Gene content and conservation of 5D chromosome

The non-repetitive sequences of 5DS and 5DL were compared against the fully annotated grass proteomes of *Brachypodium distachyon*, *Oryza sativa* (rice) and *Sorghum bicolor* (sorghum); UniProt sequences from related monocotyledonous plants; and UniGenes from *T. aestivum*, *Hordeum vulgare*, *Panicum virgatum*, *Saccharum officinarum*, and *Zea mays*, to explore potential protein-coding loci on 5D chromosome. A total of and 26,535 and 53,163 reads from 5DS and 5DL, respectively, retrieved significant matches from at least one of the above mentioned datasets. Among these, 18,771 reads (70.7%) from 5DS and 33,619 reads (59.9%) from 5DL yielded matches from only one of the query datasets, while 1,210 reads (4.5%) and 4,208 reads (7.5%) from 5DS and 5DL, respectively, yielded matches from all three query datasets, suggesting that 5DL may contain a higher proportion of highly conserved genes.

Table 3. The numbers of 5D survey sequence reads exhibiting homologies to model grass proteins, UniGene or UniProt sequences.

	Matching reads from 5DS					**Matching reads from 5DL**				
	Bdi	*Osa*	*Sbi*	UniG	UniP	*Bdi*	*Osa*	*Sbi*	UniG	UniP
# of total read	**5665**	**4035**	**4260**	**18521**	**8063**	**13413**	**9863**	**11054**	**39266**	**18844**
B. distachyon	*1303*	2737	1826	2628	2265	*5598*	6190	6852	7111	7117
O. sativa		*374*	1618	2195	1975		*4840*	5704	5767	6091
S. bicolor			*861*	2216	1905			*4988*	6364	6788
Unigene set				*13004*	3610				*12055*	10660
Uniprot set					*3229*					*6138*
Bdi+Osa+Sbi	1210					4208				

Impurities from chromosome sorting are unlikely to be represented by more than one unique read among the survey sequences (Wicker *et al.*, 2011). Thus, annotated grass proteins and UniGene/UniProt sequences covered by only a single 5D survey sequence read were excluded to avoid contaminants. A small number of UniGene/UniProt sequences with more than 50 matching 5D reads were also discarded based on the suspicion that these might correspond to novel or uncharacterized repetitive elements. Following this elimination step, 1,493 and 2,829 proteins from model grass genomes remained with significant matches from 5DS and 5DL survey sequences, respectively (Fig. 9).

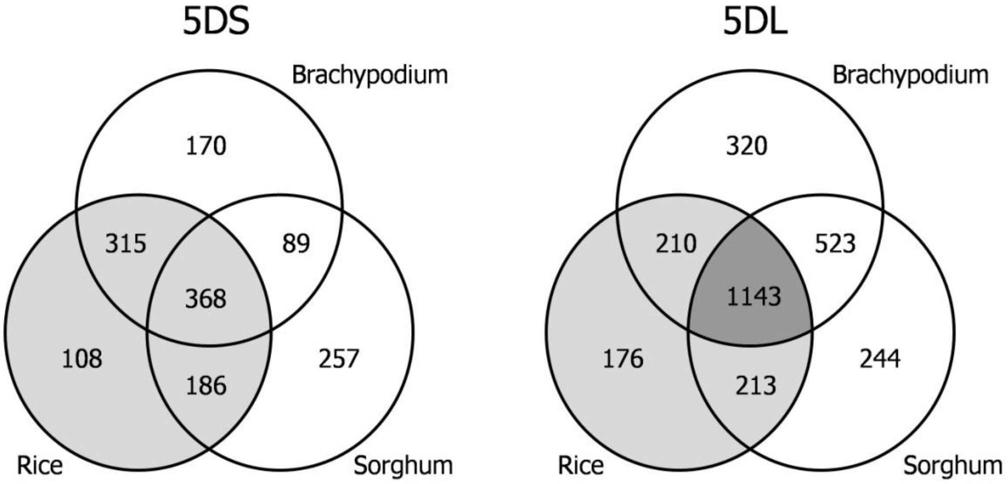


Figure 9. Gene conservation between the annotated model grass genomes and 5D. Light and dark gray shading indicate the presence of a UniGene/UniProt homolog for at least 70% and 90% of genes within the specified group, respectively.

UniGene/UniProt sequences provide evidence from related organisms without fully sequenced genomes; therefore, the presence of a UniGene/UniProt homolog increases the likelihood of a gene-associated sequence, conserved in model grasses, represents a functional gene (Fig. 9, gray shading). Gene putatively encoded by 5DL appears to have more UniGene/UniProt homologs, than that of 5DS, consistent with our previous observation. It can be argued that 5DS has accumulated more mutations or has undergone extensive neo- or subfunctionalization that resulted in the diversification of the conserved gene loci.

On the other hand, 1,812 and 4,500 UniGene/UniProt queries, majority of which were derived from *T. aestivum* and its close relative *H. vulgare*, were matched by 2 or more 5DS and 5DL sequence reads, respectively, which did not retrieve any matches from the model grass proteomes. A subset of these alignments is likely to correspond to *Triticeae* tribe-specific features; however, considering the prevalence of pseudogenes in the wheat genome (Wicker *et al.*, 2011), most of these matches should represent 5D genes that accumulated several mutations through the *Triticeae* evolution that impaired the functionality of the genes.

A total of 4289 gene models (3147 high-confidence and 1142 low-confidence) were predicted from all *T. aestivum* 5D reads and annotated based on all *Viridiplantae* proteins (Fig. 10). The gene models were classified as 'conserved' and 'non-conserved' depending on the sequence reads, incorporated into the gene models, which matched the model grasses for the former, and UniGene/UniProt sequences, for the latter.

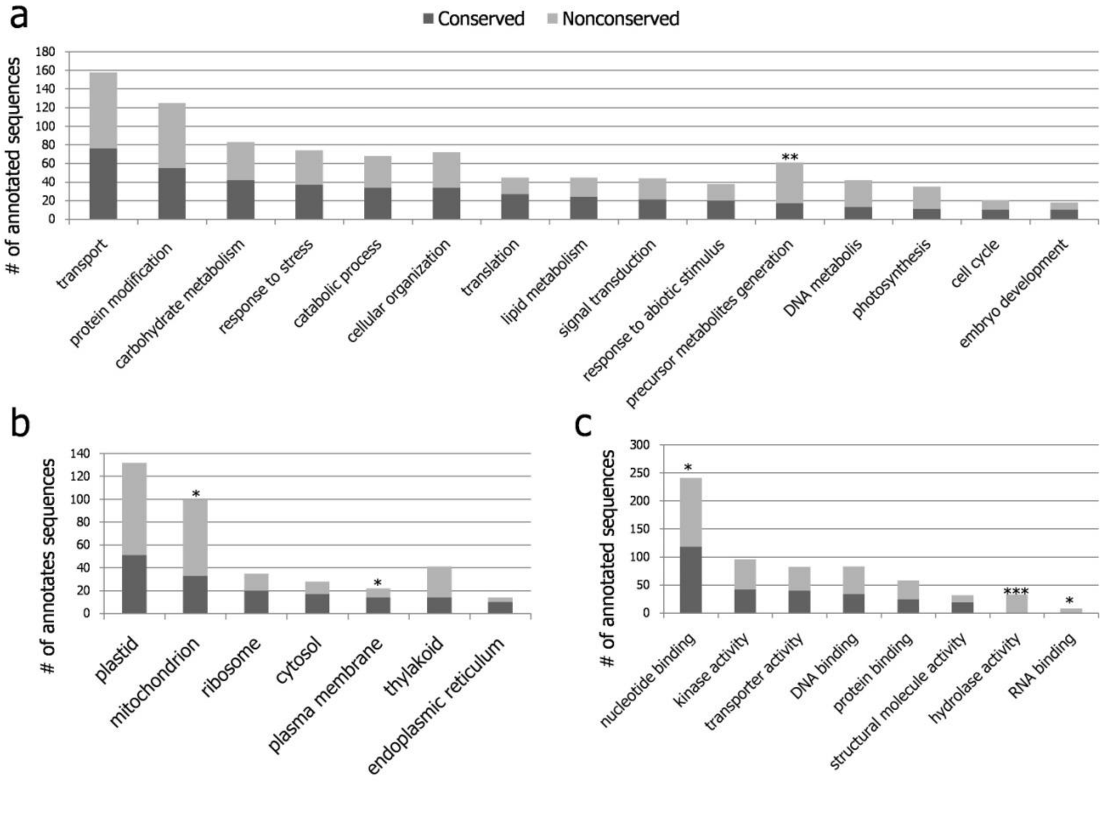


Figure 10. The most abundant GO annotations of 5D gene models for (a) Biological Process, (b) Cellular Component, and (c) Molecular Function terms. Significant differences between conserved and non-conserved gene model annotations for a given term are indicated by asterisks, deduced from Fisher's exact test for two-tailed probabilities (*p-value <0.05, **p < 0.01, ***p < 0.001).

As shown in Figure 10, certain annotations within BP, CC and MF terms were enriched among conserved and non-conserved 5D gene models. For instance, 'generation of precursor metabolites and energy' annotation of BP terms, 'mitochondrion' annotation of CC terms, and 'nucleotide binding, hydrolase activity and RNA binding' annotations of MF terms were significantly enriched amon non-conserved gene models; whereas, conserved gene models were enriched for ‘plasma membrane’ annotations of CC terms. Considering that some of the non-conserved gene models correspond to *Triticeae* specific genes, the enriched annotations may suggest novel genes related to energy related pathways might have evolved in the wheat genome after the divergence of the *Triticeae* tribe. Hydrolase activity annotations predicted exclusively from non-conserved gene models at high statistical significance (p-value = 1.01×10^{-8}, Fisher’s exact test) were also intriguing.

4.1.3. Putative tRNA genes encoded by 5D chromosome

Putative tRNA genes were explored using 5D sequences, unmasked and masked against repetitive elements, following the observation of an unusual abundance for tRNA^{Lys} species predicted from the survey sequence of wheat chromosome 6B (Tanaka *et al.*, 2013). Similarly, unmasked 5D sequences revealed a striking abundance for tRNA^{Lys} species, followed by tRNA^{Met} (Fig. 11). The same trend was not observed among the masked 5D sequences, which suggests that either some repetitive sequences are similar to tRNA coding sequences or some tRNA genes are located within the repetitive sequences. In particular, the striking abundance for tRNA^{Lys} species among repetitive sequences might have resulted from a Transposable Element(TE)-driven capture and subsequent proliferation through TE expansion. The prevalence of tRNA^{Met} species was retained among non-repetitive sequences. This is not surprising as majority of proteins start with a Met residue. In general, putative tRNA genes of 5DS and 5DL followed a similar distribution; however, 5DL had a higher content of putative tRNA genes per Mb of its size (0.59 tRNAs/Mb vs 0.29 tRNA/Mb, Appendix B).

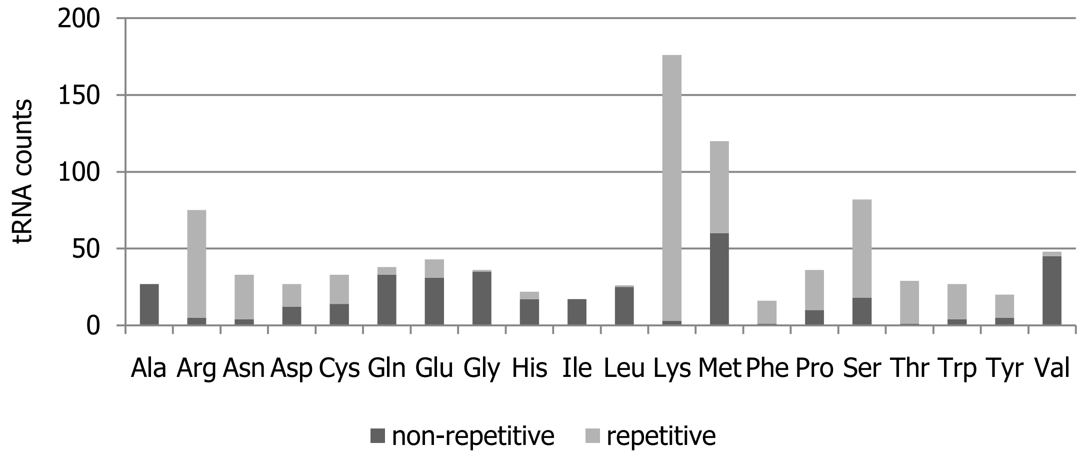


Figure 11. Putative tRNA genes predicted from repetitive and non-repetitive 5D survey sequences.

The origin of the abundance for the tRNA^{Lys} species is intriguing. In addition to wheat chromosomes 5D and 6B sequenced by Roche/454 technology, putative tRNA repertoires predicted from Illumina sequence contigs from *T. aestivum* group 5 chromosomes exhibited the same pattern for tRNA^{Lys} (Tanaka *et al.*, 2013; Mayer *et al.*, 2014). This observation indicated that tRNA^{Lys} abundance is common to different homeologous groups, independent of the sequencing technology (Fig. 12).

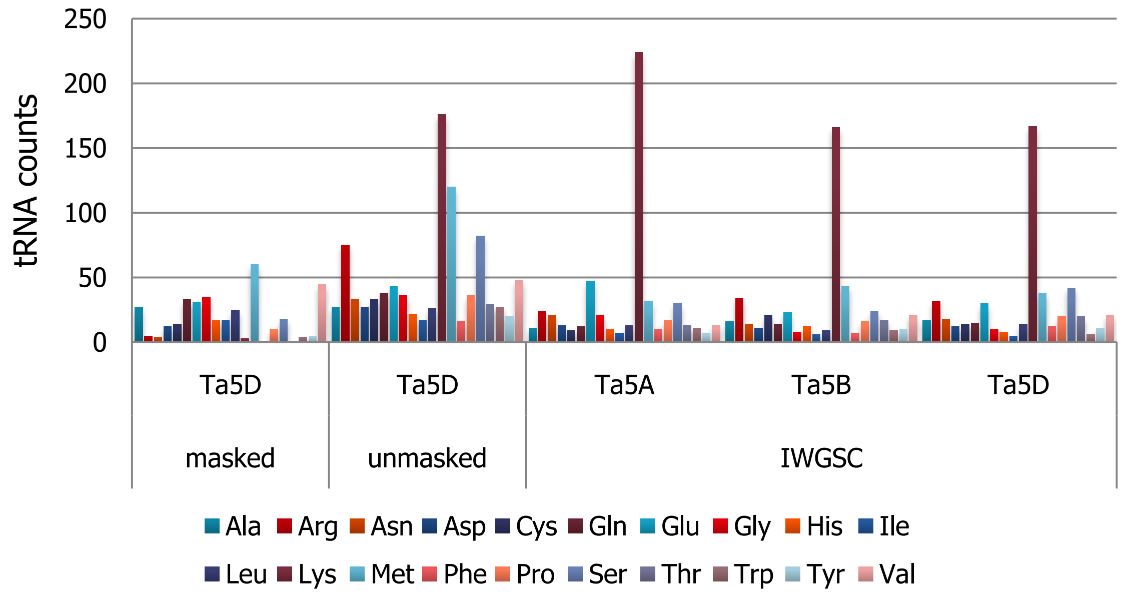


Figure 12. Putative tRNA counts predicted from chromosome 5D sequences, compared to the IWGSC Illumina contigs from homeologous group 5 chromosomes.

Bread wheat chromosomes 5D and 6B are believed to originate from different ancestral grass chromosomes (A12, A9 and A2, respectively). Therefore, a TE-driven capture of tRNA^{Lys} genes might have occured in the ancestral genome. Through wheat evolution, the expansion of TEs could lead to the proliferation of nested tRNA^{Lys} genes, greatly expanding this family of tRNA species as a genome-wide pattern in the modern bread wheat genome.

4.1.4. Syntenic relationships with model grasses

The chromosomal locations of all protein-coding loci from *Brachypodium*, rice and sorghum were used to determine conserved genomic regions by mapping orthologous 5D sequence reads onto the model grass chromosomes. As seen in Figure 13, 5DS reads identified a conserved block on the proximal end of *Brachypodium* chromosome 4 (Bd4) where several orthologous reads were clustered in that region, in contrast to orthologous reads scattered along the chromosome. Similarly, 5DL reads identified two regions of high conservation, at the proximal and distal ends of *Brachypodium* chromosomes 1 and 4, respectively (Bd1 and Bd4), as expected from the previous observations (The International Brachypodium Initiative, 2010). Comparison of *T. aestivum* 5D heatmaps of *Brachypodium* chromosomes 1 and 4 with *T. aestivum* 5A heatmaps, constructed by the same procedure using Roche/454 sequences of this chromosome, demonstrated that the two homeologous chromosomes share a similar structure. The 5A chromosome had a secondary conserved region, however, at the distal end of Bd1, which corresponds to the well documented 4AL/5AL translocation (Nelson *et al.*, 1995; Vitulo *et al.*, 2011).

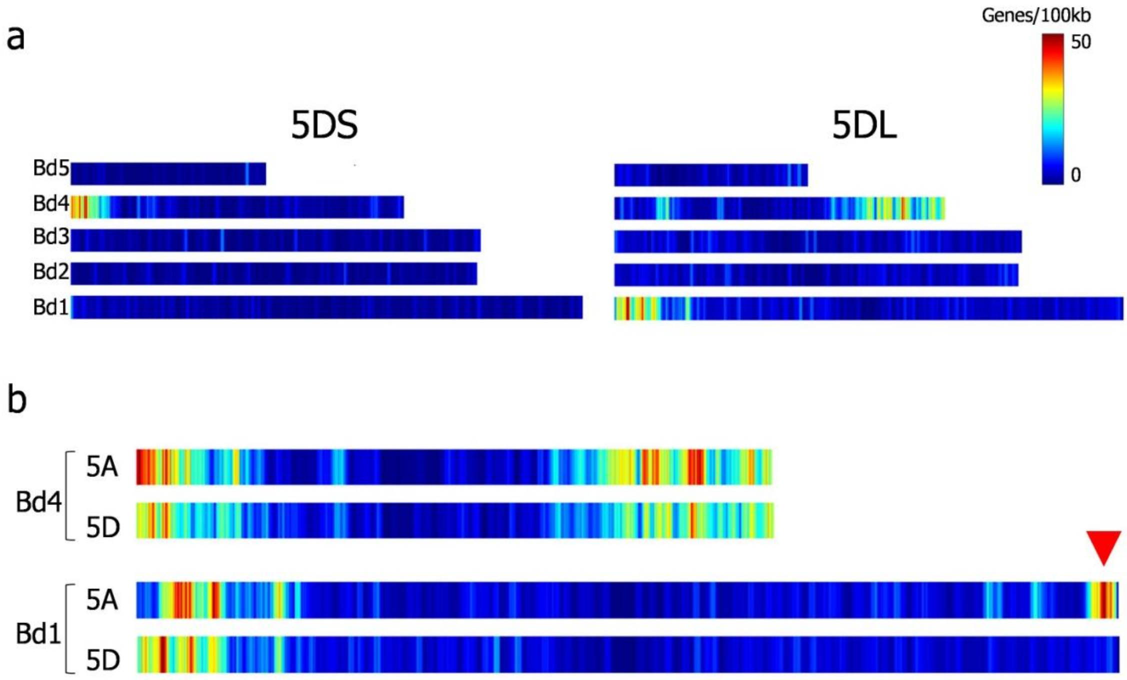


Figure 13. Heatmaps demonstrating conserved regions across the *Brachypodium* genome. (a) Conserved blocks on *Brachypodium* genome with 5DS and 5DL, (b) Conserved blocks on *Brachypodium* chromosomes 1 and 4 with *T. aestivum* 5A and 5D. The effect of previously documented 4AL/5AL translocation is indicated by a red arrowhead. Bd1-5: *Brachypodium* chromosomes 1-5.

Additionally, orthologous 5DS and 5DL reads identified conserved regions on *Oryza sativa* chromosomes 12 and 3 & 9 (Os12, Os3 & Os9), respectively. These regions were spread along the rice chromosomes, rather than concentrating at the chromosome ends. The 5DS chromosome arm exhibited a small region of homology at the distal end of *Sorghum bicolor* chromosome 8 (Sb8), whereas the 5DL chromosome arm identified two clear conserved blocks on sorghum chromosomes 1 and 2 (Sb1 and Sb2). These observations define large scale conservation patterns between the wheat 5D chromosome and the model grasses, consistent with the previous findings (The International Brachypodium Initiative, 2010) (Fig. 14).

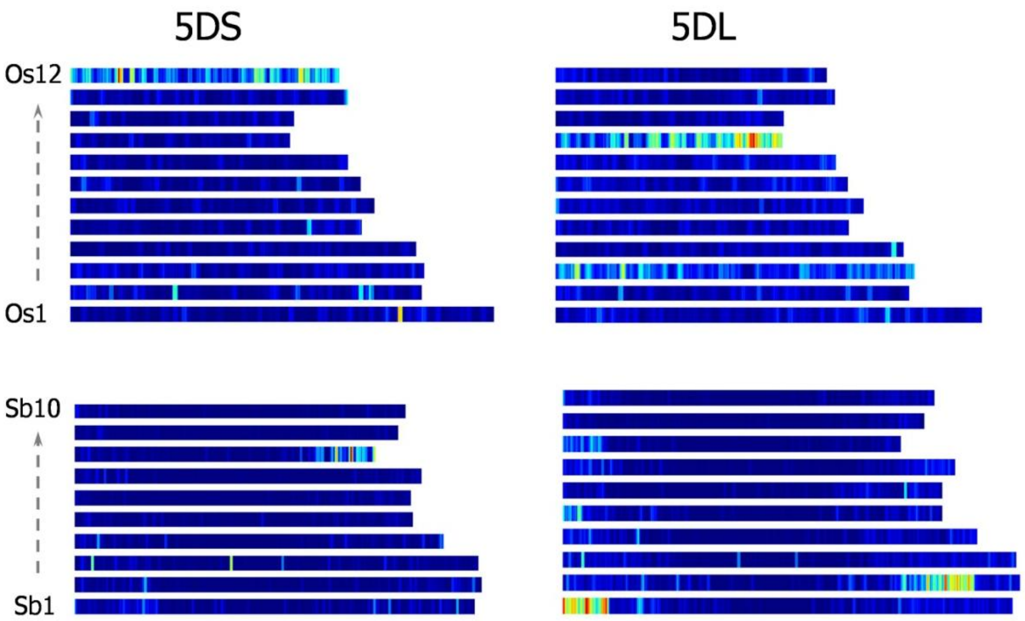


Figure 14. Heatmaps demonstrating conserved regions on *Oryza sativa* (top) and *Sorghum bicolor* (bottom) orthologous to 5DS and 5DL sequences Os1---Os12: *Oryza sativa* chromosome 1---12; Sb1---Sb10: *Sorghum bicolor* chromosome 1---10.

Orthologous 5D reads matching two or more model grass genomes were used to visualize syntenic relationships between these genomes (Fig. 15). These syntenic relationships were largely consistent with the gene conservation patterns. However, additional regions of synteny were also observed on non-orthologous *Brachypodium* chromosomes 2 & 5 and sorghum chromosomes 3, 5 & 6. These regions were composed of few orthologous genes conserved as blocks, maintaining micro-colineraity (Fig. 15, histograms). Thus, these minor syntenic regions likely reflect genome rearrangements, where small groups of genes are moved from orthologous regions to non-orthologous locations during the wheat genome evolution.

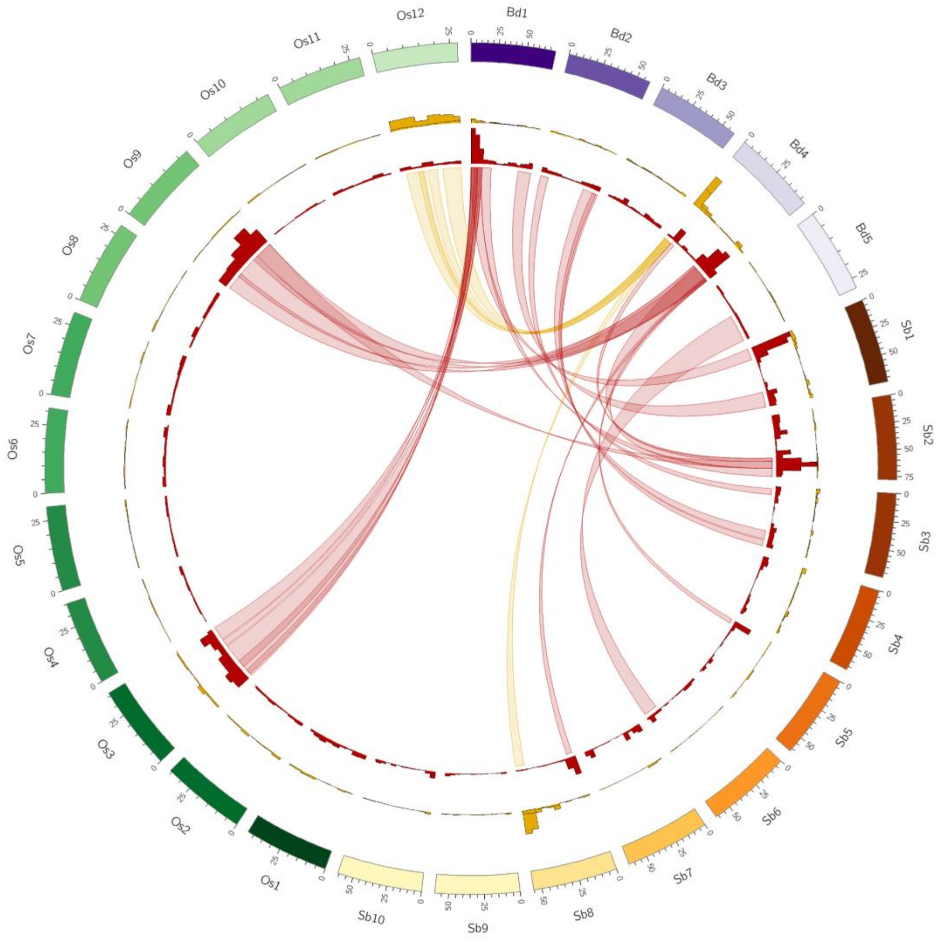


Figure 15. Syntenic relationships among model grass genomes assessed by orthologous 5DS (yellow ribbons) and 5DL (red ribbons) sequence reads. Histograms indicate the gene counts within the ribbons.

4.1.5. Virtual gene order and the 5D 'genome zipper'

In the absence of a reference genome sequence, Mayer and his colleagues described a powerful approach that utilizes genetically mapped molecular markers and synteny to define a virtual gene order for barley chromosome 1H (Mayer *et al.*, 2009). This 'genome zipper' approach was subsequently applied to all 7 barley chromosomes (Mayer *et al.*, 2011). The 5D genome zipper was constructed by mapping 518 deletion bin-mapped wheat EST and SSR markers onto the syntenic gene reads. Using the genetic mapping data for bin-mapped ESTs and SSRs, some of which also had positions on the International *Triticeae* Mapping Initiative (ITMI) wheat reference genetic map, co-linear genes were ordered, keeping the order on the *Brachypodium* genome wherever a genetic map data was not available. The 5DS genome zipper indicated a that Bradi4g00200 – Bradi4g07997 interval on *Brachypodium* chromosome 4 (Bd4) was mostly colinear with 5DS, except an inversed section between Bradi4g02840 – Bradi4g03750 (Fig. 16). Similarly, the telomeric region of 5DL revealed an inverse colinearity between Bradi1g15730 – Bradi1g00227 genes on the short arm of Bd1, despite a few genes apparently translocated to other deletion bins. The rest of the 5DL exhibited extensive perturbations and fragmented patterns of colinearity as depicted in Figure 16. Delineated by Bradi4g23910 – Bradi4g45397, three separate regions (1, 2, 3) on Bd4 were colinear, within these regions, with 5DL in a rearranged fashion (3, 2, 1). Within these colinear segments, however, several small-scale rearrangements were also evident. The boundary between the regions 1 and 2 on Bd4 could not be precisely determined with the current information; therefore, any further small-scale rearrangements between Bradi4g38980 – Bradi4g39020 would require increased resolution through more molecular markers mapped to this region. The centromeric region of 5DL was colinear with Bradi4g08180 – Bradi4g08900 segment from Bd4 (Fig. 16). It is important to note that the virtual gene order of chromosome 5D is not absolute; the wheat reference genome sequence will ultimately define the positions of all genomic features, revealing all small-scale rearrangements and breaks in micro-colinearity.

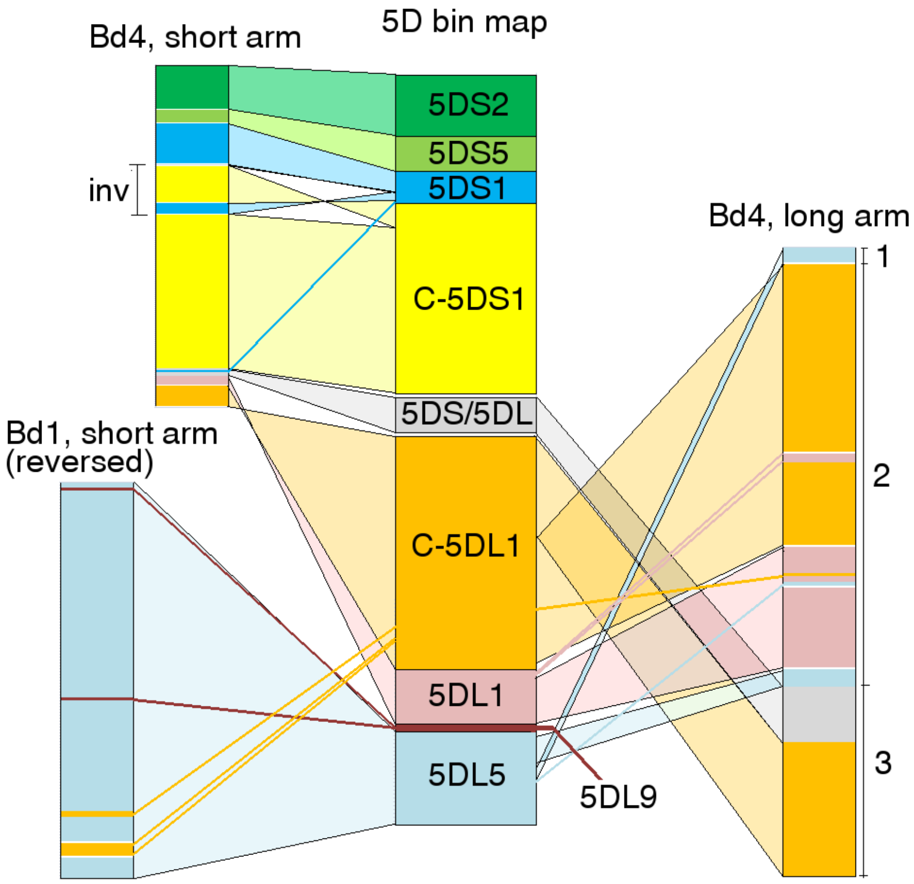


Figure 16. A virtual gene order for 5D chromosome constructed using the genome zipper approach.

4.1.6. Wheat specific genome rearrangements

Recently, a three-way comparison of the model grass genomes *Brachypodium*, rice and sorghum provided clues into genome specific rearrangements. Genome sequences of these model grasses indicate that sorghum diverged from rice and *Brachypodium* ~50 Mya, while rice and *Brachypodium* diverged from each other ~40 Mya (The International Brachypodium Initiative, 2010). Considering the evolutionary history of the three model grasses, a gene that is non-coliner in *Brachypodium*, but colinear in the other two suggests that the gene may have been ‘moved’ specifically in the *Brachypodium* genome (Wicker *et al.*, 2010). Following the same rationale, othologous genes that are found in colinear positions on the three model grass genomes, but in non-colinear positions on wheat chromosome 5D were explored to identify genes possibly moved in the wheat lineage. Contaminants from the impure fraction of the sorted chromosomes are unlikely to be represented by more than one sequence read; thus, all orthologous genes covered by a single 5D sequence read were excluded from this analysis.

Of the the remaining non-syntenic genes, 86 (5DS) and 309 (5DL) were conserved across the three grass genomes, covered by 294 5DS and 905 5DL sequence reads. Sequence reads derived from pseudogenes or gene fragments, in general, exhibit uneven coverage of the functional gene. Therefore, candidate non-syntenic genes were examined for coverage. Genes covered by at least 4 sequence reads were visually evaluated for even coverage (see Appendix C, for an example). Genes covered by 2 or 3 sequence reads were divided into 2 or 3 equal parts, respectively, and accepted as genuine copies only if all parts are covered. After the eliminatiın of pseudogenes and gene fragments, 32 and 129 putative non-syntenic genes remained for 5DS and 5DL, respectively. Of these, 22 and 36 orthologous genes were colinear in *Brachypodium*, rice and sorghum, but, found on 5DS and 5DL, respectively (Appendix D). For instance, Bradi1g17710, Os02t0167700 and Sb04g004540 are located on syntenic chromosomes in model grasses, covered by 19 sequence reads from the non-syntenic 5DS chromosome, which implies that the wheat ortholog may have been moved to the non-syntenic 5DS chromosome after wheat and *Brachypodium* lineages diverged from each other.

4.2. The 5DS Physical Map

4.2.1. Construction of the preliminary physical map

The double ditelosomic line 5D of *Triticum aestivum* cv. Chinese Spring (Sears and Sears, 1978) was used to isolate individual arms of chromosome 5D, as outlined in Section 4.1.1. The purity of the flow-sorted 5DS fractions was 88% as indicated by Fluorescence In Situ Hybridization (FISH) of the characteristic repeat families. The contaminating particles included a random mixture of fragments from various other chromosomes and chromatids.

A 5DS-specific BAC library, composed of 36,864 BAC clones, was generated from 8,120,000 sorted chromosome arms and designated as TaaCsp5DShA. The average insert size of the library was 137 kb, yielding 17x coverage of the 258 Mb-long chromosome arm (Safár *et al.*, 2010). Of this library, 26,112 BAC clones, giving 12.5x coverage of the chromosome arm, with an average insert size of 143 kb were picked for fingerprinting. Following SNaPshot^{TM} HICF procedure, good-quality fingerprints were obtained for 21,656 clones (Luo *et al.*, 2003). These fingerprints were used to construct the preliminary physical map of 5DS.

Traditionally, the construction of physical maps from fingerprinted BAC clones has been achieved by the FingerPrinted Contig (FPC) software, which was also implemented in the initial physical mapping studies carried out under the framework established by the International Wheat Genome Sequencing Consortium (IWGSC) (Nelson *et al.*, 2005). Recently, however, an alternative software, Linear Topology Contig (LTC), has been reported to build longer contigs, resulting in more reliable maps comprised of fewer contigs. LTC also enables the visual control of the contig topology to improve or eliminate problematic contigs with local disruptions of contig linearity (Frenkel *et al.*, 2010). Therefore, LTC has been increasingly adopted in recent studies and the comparison of the FPC- and LTC-constructed maps favors the latter approach (Lucas *et al.*, 2013; Philippe *et al.*, 2013). Initially, both software programs, FPC and LTC, were independently used to construct the 5DS physical map and the resulting preliminary maps were compared to choose the most promising map to progress with (Table 4).

Table 4. The comparison of FPC and LTC constructed 5DS preliminary maps.

	FPC assembly	**LTC assembly**
Total no. of clones	21656	21656
Number of contigs (>5 clones)	350	120
MTP clones	1894	2155*
Assembly length	202728 kb	176838 kb
Average contig size	579	1078
Largest contig size	4053 kb	6649 kb
N50	1141 kb	2173 kb
L50	53	27
Contigs>1Mb	63	58

*Picked by FPC software.

As seen in Table 4, the number of LTC-constructed contigs was considerably less than that of FPC, with the average contig size almost twice the size of the average FPC-contigs (1078 kb vs. 579). Consistently, the N50 value, regarded as the quality measure of the LTC physical map, was almost twice as large as the N50 value of the FPC physical map. Therefore, despite the lower coverage of the LTC-constructed map compared to the FPC-constructed map (78% versus 68%), LTC map was concluded to be more reliable and informative than the more fragmented FPC map, and adopted as the method of choice for the map construction.

The evaluation of the Minimum Tiling Path (MTP) picked by LTC in the LTC-constructed map revealed several buried clones that were smaller subsets of longer clones. Hence, the MTP of the LTC map was discarded, and, a new MTP comprised of 2,155 clones was picked from the LTC map using the FPC software to select longer clones. The FPC-selected MTP was further evaluated by the LTC in terms of clone overlaps, as statistically insignificant clone overlaps may lead to gaps at the sequence level. Any clone overlaps that were not significant below a cutoff of 10^{-14} was reinforced by the manual addition of 210 supplementary clones to cover the overlap region. This significance cutoff was picked relative to the significance cutoff of 10^{-15} used by LTC to build the contigs (Frenkel *et al.*, 2010). Also, 163 clones deemed as Questionable clones (Q-clones) by LTC were added to the MTP. These Q-clones can cluster into 2 or more contigs. While some Q-clones are chimaeric clones that lead to problematic contigs, some are bridge clones that can be used to merge separate contigs of low coverage, if supported by molecular marker data. In total, the manually edited FPC-picked MTP of LTC-constructed preliminary 5DS physical map was comprised of 2,528 clones.

4.2.2. Assessment of the 5DS preliminary map

The LTC-constructed 5DS preliminary map had an assembly length of 176 Mb, covering over 68% of 258 Mb-long chromosome arm, organized into 164 contigs of which 120 had 6 or more clones. The N50 value of the preliminary map indicated that half of the assembly was covered with contigs longer than 2173 kb. The L50 value, denoting the number of such contigs, was 27. Additionally, 58 contigs were larger than

1 Mb in the 5DS physical map, the largest being over 6.6 Mb (Table 4). Compared to the N50 values of the previously published chromosome-specific wheat physical maps of 1AL (1166 kb), 1BL (961 kb), 1AS (798 kb), 1BS (2430 kb), 6AS (1106 kb) and 6AL (921 kb), the N50 value of 5DS map at 2173 kb, including the short contigs with 5 or less clones, indicates a high quality map for this chromosome arm (Lucas *et al.*, 2013; Philippe *et al.*, 2013; Breen *et al.*, 2013; Raats *et al.*, 2013; Poursarebani *et al.*, 2014).

To assess the contig length distribution of the preliminary map, contig sizes were plotted against the number of contigs and megabases of the assembly covered by the contigs in the respective size range. For visualization purposes contig sizes are distributed in size ranges of 100 kb for contigs smaller than 200 kb, 500 kb for contigs larger than 3000 kb, and 200 kb for the remaining contigs (Fig. 17).

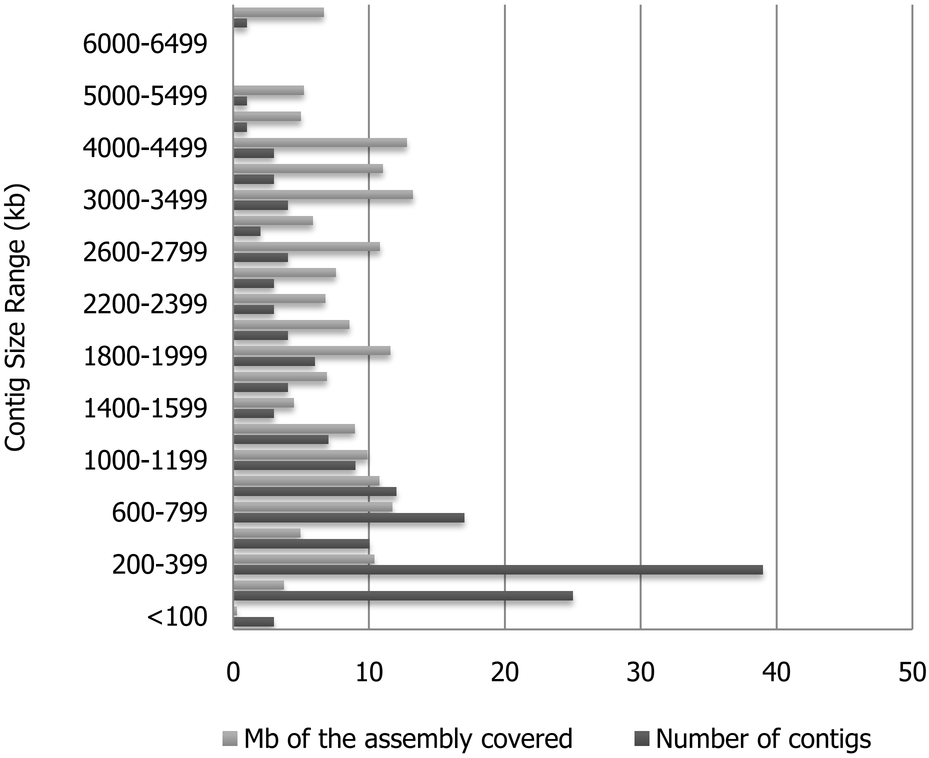


Figure 17. Distribution of contig lengths and corresponding size of the assembly covered across different size ranges.

As seen in Figure 17, many contigs fell in the 100 – 400 kb range; however, the fraction of the assembly covered by each size group was similar across all ranges. The assembly depth of the contigs was calculated by dividing the estimated contig length by the actual length of the contig length, and then plotted against the contig lengths (Fig. 18). This revealed that while smaller contigs were clustered around a depth of 1 – 5x, large contigs clustered around higher depths around 12 – 20x, closer to the average assembly depth, 14x (estimated assembly length/actual assembly length). In particular, 65 contigs had 14x or higher depth. The preliminary map details are given in Appendix E.

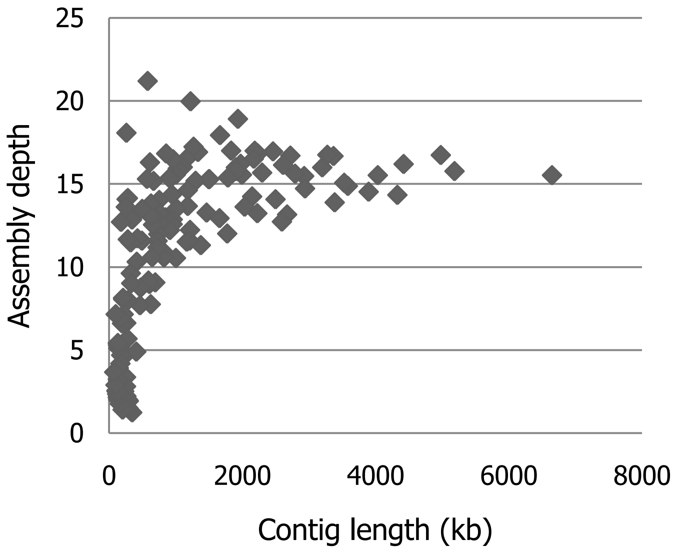


Figure 18. Depth of assembly by contig length.

4.2.3. MTP screening with molecular markers

In order to refine the preliminary 5DS physical map and order the contigs along the chromosome arm, a variety of molecular markers, including Simple Sequence Repeat (SSR), Conserved Orthologous Set (COS), Expressed Sequence Tag (EST) and Insertion Site-Based Polymorphism (ISBP), were used to anchor the contigs of the physical map through Polymerase Chain Reaction (PCR). A 3D pool approach was adopted to screen the MTP clones, in which, instead of screening each clone one by one, MTP clones are rearranged in 3 dimensional pools as row, column and plate Appendix F). Initially, 23 genetically mapped SSR markers (6 BARC, 9 CFD, 3 WMC, 4 WMS and 1 GPW), 13 COS markers, 10 EST markers and 2 gene-based markers (*Pina-D1*, *Pinb-D1*) were used to screen the MTP pools. Of these, 13 SSR markers

(56%), 12 COS markers (92%) and 2 gene-based markers (100%) were successfully assigned to specific MTP clones. In contrast, only 1 out of 15 EST markers could be assigned to a specific clone, whereas the remaining markers failed to amplify on multiple occasions. Intronic sequences found within the MTP clones were concluded to interfere with the amplification of EST-based markers in which intronic sequences are spliced out. To resolve this issue, the sequences of genetically mapped EST markers retrieved from Grain Genes (http://wheat.pw.usda.gov/GG3/) were blasted against 1.34x coverage genomic survey sequences of 5DS, in an attempt to determine intron-exon boundaries. Using these alignments, new markers were designed from sequences belonging to a single or two closely adjacent exons. A total of 43 out of 50 newly designed EST markers (86%) could be assigned to MTP clones. Validating this approach, two EST markers, namely BF483719 and CD882766 that failed to amplify from any of the MTP pools previously, were anchored to contigs 115 and 134 through the newly designed primer sets. For the remaining 12 EST markers from the initial set, the above mentioned approach could not find favorable sequences to be amplified.

In addition to the genetically mapped markers, 47 ISBP markers (out of a total of 16,727) designed from 1.34x survey sequences were screened on the MTP pools, anchoring 30 of them to specific MTP clones. In total, out of the 164 contigs of the 5DS physical map, 48 were physically anchored by at least 1 molecular marker via PCR (Appendix G).

Recently, a customized microarray hybridization approach was proposed and validated to screen BAC-based MTP pools by large numbers of markers, in a dye-swap design (Rustenholz *et al.*, 2010). Accordingly, an 8x60k Agilent SurePrint G3 Gene Expression Custom Microarray (Agilent, Santa Clara, CA, USA) was designed from the genetically mapped gene(7), SSR(20), COS(13), SNP(112) and EST(122) markers, along with the ISBP markers(5,120) and conserved gene-associated sequences(6,996) deriving from 1.34x 5DS survey sequences. Of these, 25% of SSR (5 out of 20), 23% of COS (3 out of 13), 12% of EST (15 out of 122), 15% of SNP (17 of the 112), 18% of conserved reads (1,306 out of 6,996), and 8% of ISBP (416 out of 5120) were putatively assigned to specific 5DS MTP clones through microarray hybridization, following data normalization and deconvolution. Processing of probes from 3 conserved genes and 2 ISBPs resulted in ambiguous assignments, and, thus, these assignments were discarded.

In the end, 1,762 unique gene or marker associated sequences were putatively assigned to 3,066 MTP clones of the 5DS physical map at high stringency (Appendix H).

In summary, a total of 1,865 molecular markers were confidently anchored to 105 of the 164 contigs of the 5DS physical map using the two approached mentioned above corresponding to a marker density of ~10.5 markers per Mb. The total length of the anchored contigs comprised approximately 91% of the total assembly length (161Mb of the total 176Mb).

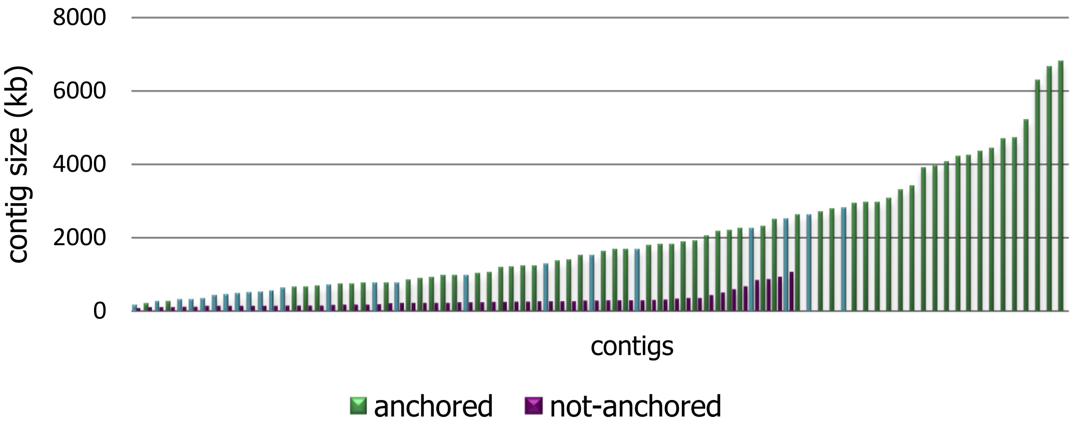


Figure 19. Size distribution of the anchored and not-anchored contigs. Both green and blue bars indicate anchored contigs; blue bars indicate anchored contigs without defined locations on the 5DS chromosome arm (see the following section). Purple bars represent contigs that could not be anchored by either of the PCR or microarray approaches.

The contigs that could not be anchored by the molecular markers using either of the approaches were mostly short contigs that were of little informative value, as demonstrated in Figure 19 (purple bars). In fact, when short contigs composed of 5 or less clones were excluded, the assembly length of the anchored contigs (100 out of 120) made up approximately 95% of the total assembly length (160Mb of the total 168Mb, excluding short contigs).

4.2.4. Ordering 5DS contigs and supercontigs along the chromosome arm

Following the construction of the 5DS preliminary map, 5DS contigs were re-evaluated to build 'supercontigs' to aid in the mapping of contigs along the 5DS chromosome. Both ends of each contig were checked on a one-by-one basis on LTC for possible elongation into other contigs. The elongate function of LTC searches for statistically less significant overlaps (cutoff>10^{-15}) between the clones of different contigs, once the contigs are built at a stringent cutoff (cutoff <10^{-15}) (Breen *et al.*, 2013). This procedure allowed the construction of 21 supercontigs involving 45 contigs.

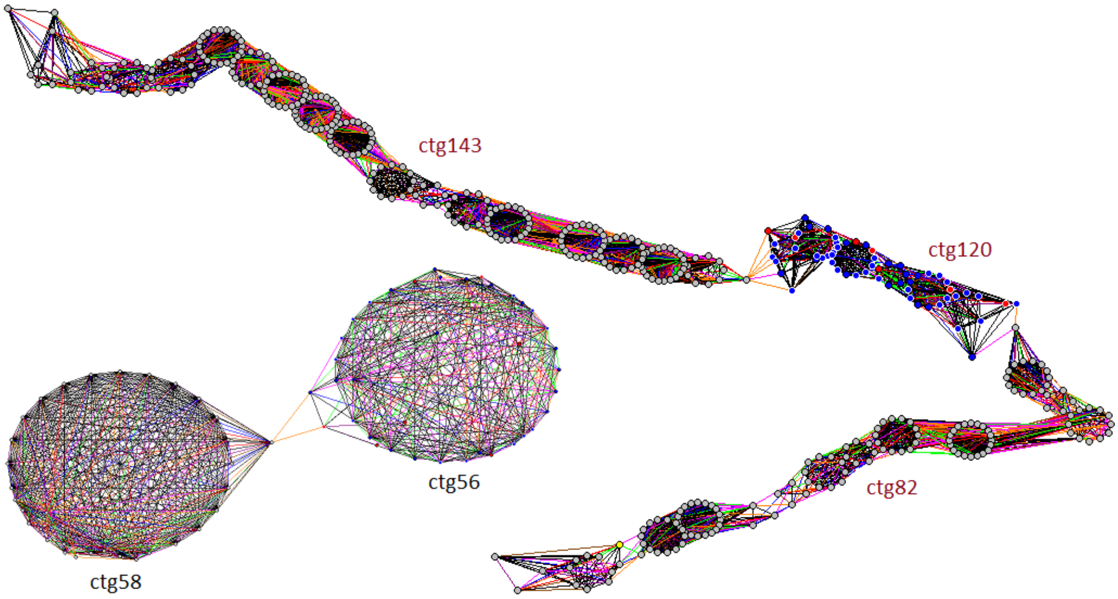


Figure 20. Network of clone overlaps for two representative supercontigs, SC2 and SC10 (Table 5). Vertices indicate individual BAC clones and edges with different colouring indicate clone overlaps with different levels of significance.

Two representative supercontigs are shown in Figure 20, and all supercontigs are listed in Table 5. A supercontig is accepted only when elongation from all participating contigs yield the same structure.

Table 5. All 21 supercontigs of the 5DS physical map.

Supercontig no.	contigs	# of clones	total # of clones	status
SC1	[CTG98-CTG54-CTG68]	86, 21, 583	690	mapped
SC2	[CTG56-CTG58]	38, 33	71	mapped
SC3	[CTG57-CTG162]	74, 101	175	anchored
SC4	[CTG66-CTG122]	34, 135	169	mapped
SC5	[CTG70-CTG145-CTG146]	231, 14, 104	349	mapped
SC6	[CTG71-CTG100]	26, 318	344	mapped
SC7	[CTG74-CTG109]	9, 9	18	anchored
SC8	[CTG77-CTG127]	138, 393	531	mapped
SC9	[CTG79-CTG80]	11, 3	14	anchored
SC10	[CTG143-CTG120-CTG82]	251, 61, 136	448	mapped
SC11	[CTG88-CTG90]	38, 96	134	mapped
SC12	[CTG92-CTG64]	99, 2	101	Notanchored
SC13	[CTG111-CTG112]	211, 62	273	mapped
SC14	[CTG158-CTG118]	224, 70	294	mapped
SC15	[CTG144-CTG121]	372, 113	485	mapped
SC16	[CTG136-CTG148]	218, 124	342	anchored
SC17	[CTG140-CTG149]	217, 256	473	mapped
SC18	[CTG159-CTG150]	76, 62	138	mapped
SC19	[CTG131-CTG151]	48, 16	64	mapped
SC20	[CTG157-CTG156]	358, 373	731	mapped
SC21	[CTG105-CTG108]	28, 26	54	anchored

In order to order the contigs and supercontigs from the 5DS physical map along the chromosome arm, molecular marker data and syntenic relationships with the model grass *Brachypodium distachyon* or the D-genome progenitor *Aegilops tauschii* were utilized. In the case of molecular markers, for the previously mapped SSR, EST and COS markers, genetic map positions were retrieved (Quraishi *et al.*, 2009, http://wheat.pw.usda.gov/GG3/), and for the ISBP markers, deletion bin mapping was performed using the 5DS deletion stocks of wheat (Endo and Gill, 1996).

The D-genomes of *Ae. tauschii* and *T. aestivum* are highly similar due to the relatively recent hybridization of *Ae. tauschii* with *T. turgidum*, creating the hexaploid *T. aestivum*. The physical map and the draft genome of *Ae. tauschii* were published very recently, providing invaluable resources for the wheat community (Luo *et al.*, 2013; Jia *et al.*, 2013). Based on the extensive similarities between *Ae. tauschii* and *T. aestivum* D-genomes, within a deletion bin, allocated 5DS contigs, which were linked to an orthologous sequence from *Ae. tauschii* 5D chromosome, were further ordered within the bin according to the order of *Ae. tauschii* orthologs on its 5D chromosome. In cases where such an orthologous sequence from *Ae. tauschii* could not be detected, the order on the 5DS genome zipper, built by the syntenic relationships with the model grass *Brachypodium distachyon*, was retained to locate an associated contig. Anchored contigs that could not be associated with any orthologous sequences from *Ae. tauschii* or *B. distachyon* were only assigned to the relatively large deletion bin intervals, with the order undetermined.

A total of 80 contigs (39 contigs in 18 supercontigs and 41 contigs) and 79 (39 contigs in 18 supercontigs and 40 contigs) of the 105 and 100 anchored contigs of the 5DS physical map with or without short contigs, respectively, were allocated to 4 cytogenetically defined deletion bins using the approaches explained above (https://www.ksu.edu/wgrc/Germplasm/Deletions/group5.html, 5DS bins: 1.00-0.78, 0.78-0.67, 0.67-0.63, 0.63-0, Appendix I). Of these, 63 were further ordered within the deletion bins utilizing syntenic relationships primarily with *Ae. tauschii* or, secondarily with *B. distachyon* (Fig. 21).

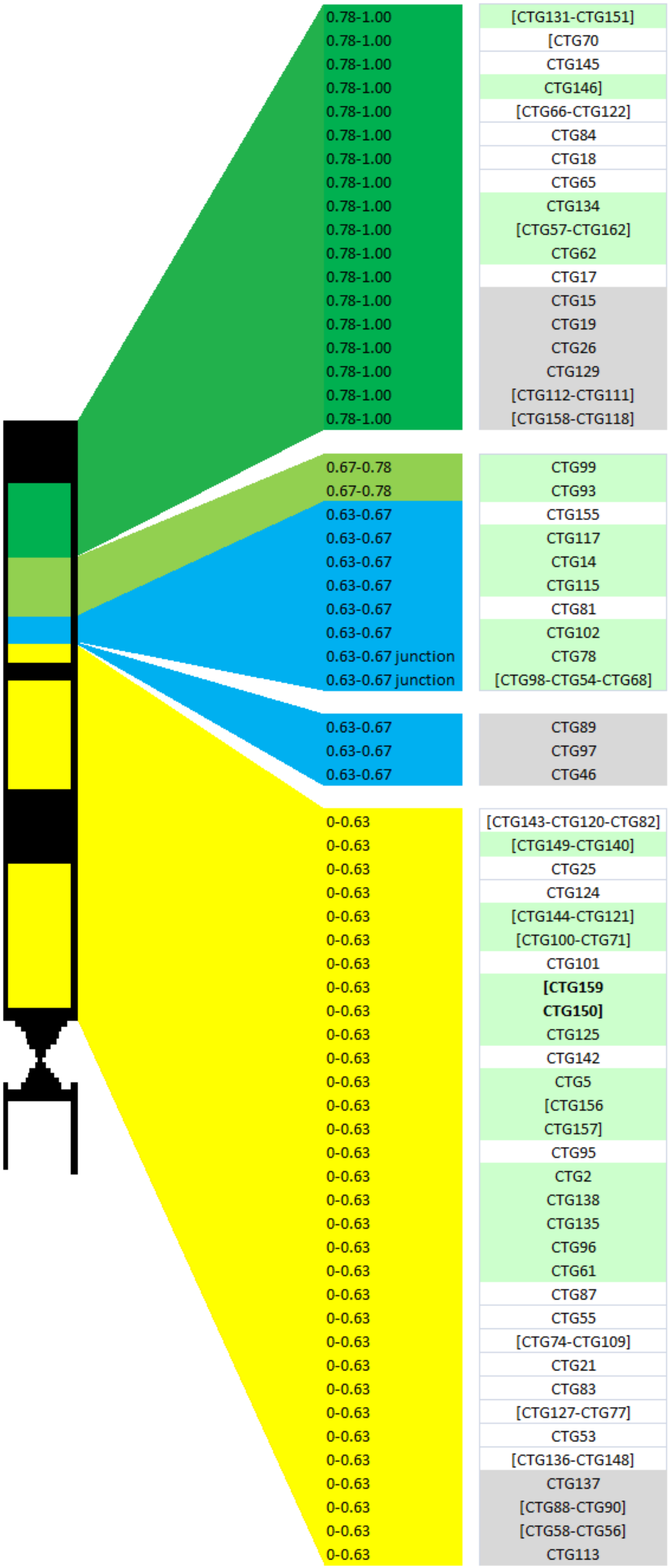


Figure 21. Final bin-map of the 5DS chromosome with ordered contigs or supercontigs. Contigs highlighted in light-green indicate ordering based on *Ae. tauschii* ortholog; contigs that are not highlighted indicate ordering based on *B. distachyon* ortholog; contigs in grey indicate unknown order.

According to the mapping and ordering of contigs and supercontigs on 5DS chromosome, the most distal bin, 0.78-1.00, and the most proximal bin, 0.63-0, contained 23 contigs and 42 contigs, respectively. Intriguingly, the narrow 0.63-0.67 interval, representing only 4% of the chromosome arm based on cytogenetic estimates, was allocated 13 contigs, whereas the relatively larger bin, 0.67-0.78, had 2 contigs allocated. Some contigs appeared to be located at or close to the bin junctions. For instance, CTG78 and SC1[CTG98-CTG54-CTG68] were assigned to 0.63-0.67 deletion bin; however, the relative positioning of these contigs based on the 5DS genome zipper implied that these contigs might be located at the junction of the deletion bins 0-0.63 and 0.63-0.67. Curiously, CTG93 was assigned to the 0.67-0.78 bin by synteny, but CFD81 SSR marker anchoring this contig indicated to the most proximal 0-0.63 which may indicate a break in micro-colinearity.

4.2.5. Analysis of the 5DS genome structure

Contigs and supercontigs that were assigned to specific intervals were used to estimate the physical sizes of the cytogenetically determined deletion bins. The size estimates were corrected by the 54% chromosome coverage of the mapped contigs. The physical size estimate of the most distal bin, 0.78-1.00, was close to the cytogenetic estimates (22% of the chromosome arm), at a little over 49 Mb, comprising 19.2% of the 5DS. The relatively narrow deletion bin, 0.67-0.78, which cytogenetically represented 11% of the chromosome arm, had an estimated physical size of almost 15 Mb, corresponding to 5.7% of the chromosome arm. Strikingly, the physical size of the much smaller deletion bin, 0.63-0.67, was over 55 Mb, roughly 21% of the entire arm, which suggested that either one of the estimates was highly inaccurate, or, this deletion bin was overrepresented in our BAC library due to some unknown artifact. The estimated physical size of the most proximal deletion bin, 0-0.63, was 138 Mb. The difference between the estimated physical (53.5%) and cytogenetic (63%) sizes of this deletion bin may reflect an inaccuracy of the cytogenetic estimates of the two consecutive deletion bins. However, it is equally likely that some deletion bins are over- or underrepresented in our dataset as only 54% of the chromosome arm was covered with mapped contigs. A third explanation could be the unequal representation of the deletion bins by the genetically mapped molecular markers.

The 1BS physical map reported a bias in the number of clones in mapped contigs. The telomeric contigs contained fewer clones on average, in contrast to the centromeric contigs (Raats *et al.*, 2013). Such a bias was not observed in the 5DS physical map, although the cumulative lengths of the contigs tended to decrease towards the telomere. While the number of clones per Mb was slightly lower in the most distal deletion bin, at 93.1 clones/Mb, these values were more consistent across the remaining bins (105.2, 105.7 and 106.6 clones/Mb for 0.67-0.78, 0.63-0.67 and 0-0.63 intervals, respectively). This observation suggests that mapping of the contigs or supercontigs across the deletion bins were generally uniform, regardless of the bin size.

Despite the lack of an apparent bias in the number of the clones in contigs mapped to different deletion bins, contig lengths allocated to deletion bins revealed an interesting situation. While over than half of the contigs assigned to the most distal bin were smaller than 1 Mb, much longer contigs, including the longest, CTG138, were assigned to the most proximal bin, 0-0.63 (Fig. 22).

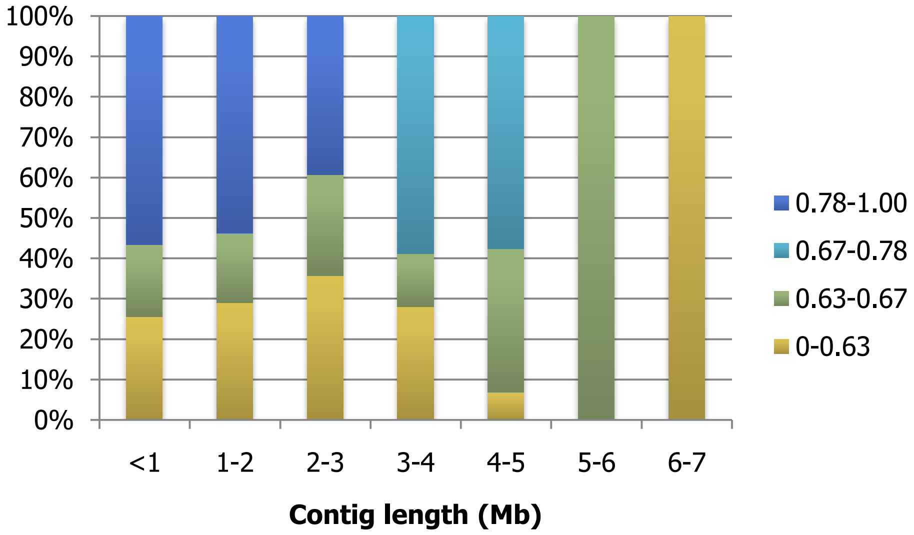


Figure 22. Distribution of contig lengths assigned to different deletion bins. Contigs are grouped into 1 Mb-intervals.

4.2.6. Small-scale genome rearrangements and perturbations in micro-colinearity

The refined 5DS physical map, incorporating conserved genomic features from the *Ae. tauschii* physical map the draft genome (Luo *et al.*, 2013; Jia *et al.*, 2013), enabled further improvements to the 5DS genome zipper, constructed from 1.34x survey sequences. The improved genome zipper was compared against the 5DS genome zipper constructed by the IWGSC from Illumina contigs, revealing several small scale inconsistencies (Mayer *et al.*, 2014). The 'Genome Zipper' approach was first demonstrated by Mayer and his colleagues for the barley chromosome 1H, which was subsequently extended to all seven barley chromosomes (Mayer *et al.*, 2009; Mayer *et al.*, 2011). In the absence of extensive genomic resources, these genome zippers had been highly useful in building a virtual gene order. Although genome zipper approach is a powerful tool to explore the gene space, particularly for species that lack a reference genome sequence, the inconsistencies observed between the genome zippers by IWGSC and by our group indicates that these tools should be utilized cautiously (Fig. 23a).

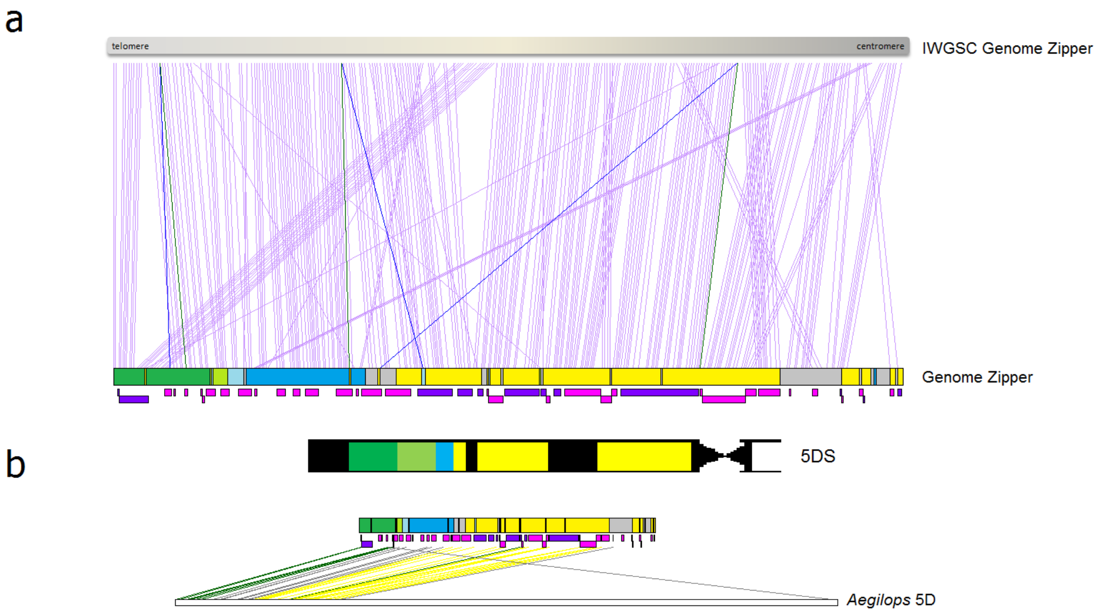


Figure 23. Comparisons of the 5DS genome zippers of (a) by IWGSC**, (b)** by Lucas *et al.*, 2014 (refined by the 5DS physical map). Deletion bins are color-coded, where dark green, light green, blue and yellow corresponds to 0.78-1.00, 0.67-0.78, 0.63-0.67 and 0-0.63 bins, respectively. Gray shaded regions on the lower genome zipper indicate locations of uncertainty. Contigs and supercontigs matching *Brachypodium* orthologs

are indicated as pink or purple boxes, respectively, below the genome zipper. Sizes of the boxes do not necessarily reflect corresponding contig/supercontig sizes.

Among the inconsistencies between the genome zippers (Fig. 23a), one region involving a number of genes was intriguing. Orthologous *Brachypodium* genes delineated by Bradi4g00450-Bradi4g00790 were assigned to the 0.78-1.00 deletion bin of the 5DS physical map, while this region was located much closer to the 0-0.63 deletion bin in the IWGSC genome zipper. Strikingly, Bradi4g05880, within this region, was previously located to the 0-0.63 bin in our genome zipper, but was relocated to the distal deletion bin 0.78-1.00, based on the *Ae. tauschii* physical map. Additionally, three putative duplications involving *Brachypodium* orthologs, Bradi4g00980, Bradi4g02450, Bradi4g06000, were detected, which were not previously reported, through the refinements by the 5DS physical map. These putative duplications are indicated by blue and green lines in Figure 23a, where blue lines correspond to the suspected duplicated copies.

Despite the extensive homology between *Ae. tauschii* and *T. aestivum*, the improved genome zipper revealed putative small-scale rearrangements between the two genomes. For instance, the EST marker BE443751, relocated from the 1.00-0.78 distal bin to 0.63-0 proximal bin, based on additional information by CTG100 also anchored by this marker, suggested small-scale rearrangement (Fig. 23b, green single line among the yellow lines). In addition, CTG134 revealed significant similarities to two separate locations on *Ae. tauschii* 5D chromosome, one of which is found close to the telomere region, suggestive of a putative duplication event (Fig. 23b, isolated gray line). The borders of the deletion bins largely remained uncertain between the most distal and proximal bins (Fig. 23b, multiple gray lines). These regions may carry additional small-scale rearrangements, which remains elusive at this time. Small-scale rearrangements are common among the highly dynamic *Triticeae* genomes (Wicker *et al.*, 2011).

4.2.7. Gene space organization of chromosome 5DS

Putative assignment of conserved gene-associated probes to MTP clones through microarray hybridization allowed the exploration of the gene space and organization of

5DS. The 1.34x 5DS survey sequences were blasted against three fully annotated grass genomes, *Brachypodium*, rice and sorghum, from which reciprocal best hits were retained as ‘conserved gene-associated sequences’, which were later used to design microarray probes. A total of 1,306 sequences were associated with specific MTP clones of the 5DS physical map; of these 95, 41, 105 and 231 were found within deletion bins 0.78-1.00, 0.67-0.78, 0.63-0.67 and 0-0.63, respectively. Considering the estimated physical sizes of the deletion bins, these numbers indicate gene densities ranging from 3.17 genes/Mb to 5.17 genes/Mb, with the highest gene density observed for the 0.67-0.78 bin (Table 6).

Table 6. The gene content and organization of 5DS across deletion bins, as assessed by the conserved gene-associated probe hybridizations.

	Syntenic		**Non-syntenic**		**Total**		**In islands**		**Isolated**		
Interval	N	D	N	D	**N**	**D**	N	D	N	D	**Cumulative length (Mb)**
0-0.63	131	1.80	100	1.37	**231**	**3.17**	192	2.63	39	0.53	72.92
0.63-0.67	44	1.48	61	2.05	**105**	**3.53**	83	2.79	22	0.74	29.71
0.67-0.78	13	1.64	28	3.53	**41**	**5.17**	33	4.16	8	1.01	7.93
0.78-1.00	20	0.75	75	2.83	**95**	**3.58**	79	2.98	16	0.60	26.54

* N: Number, D: Density.

** Cumulative lengths are based on physical size estimates.

The conserved gene-associated sequences assigned to 5DS MTP clones were classified as ‘syntenic’ if the sequence was derived from syntenic regions across 5DS, or ‘non-syntenic’ otherwise. Despite the positive gradient of gene densities towards the telomere, syntenic gene densities were not significantly correlated with the overall gene density gradient (Pearson’s correlation coefficient r = 0.16, p-value = 0.84), in contrast to the non-syntenic gene density (r = 0.87, p-value=0.13), although the correlation was relatively weak.

Two or more genes located on the same or overlapping BAC clones can be assumed to form “islands” of genes and the gene space of the wheat genome has been suggested to be dominated by gene islands, compared to the isolated genes (Choulet *et al.*, 2010; Raats *et al.*, 2013; Philippe *et al.*, 2013). The same trend was observed for the 5DS; genes assigned to MTP clones tended to cluster together (Fig. 24, Table 6).

Furthermore, the density of the genes forming gene islands was highly correlated with the overall gene gradient across the deletion bins (r = 0.9956, p-value = 0.0044); such a correlation was not observed for the isolated genes (r = 0.9509, p-value = 0.0491).

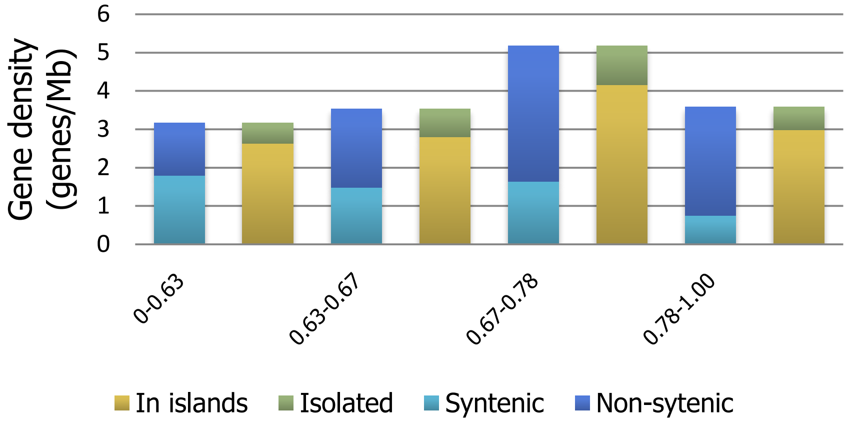


Figure 24. Gene space organization of 5DS chromosome.

To functionally characterize the 5DS gene space, 1,306 conserved gene-associated sequences assigned to the 5DS physical map were annotated against all *Viridiplantae* proteins. The annotation of these sequences revealed the top species in terms of sequence similarity as *Ae. tauschii*, as expected from the close evolutionary relationships between wheat and *Aegilops* (Fig. 25a).

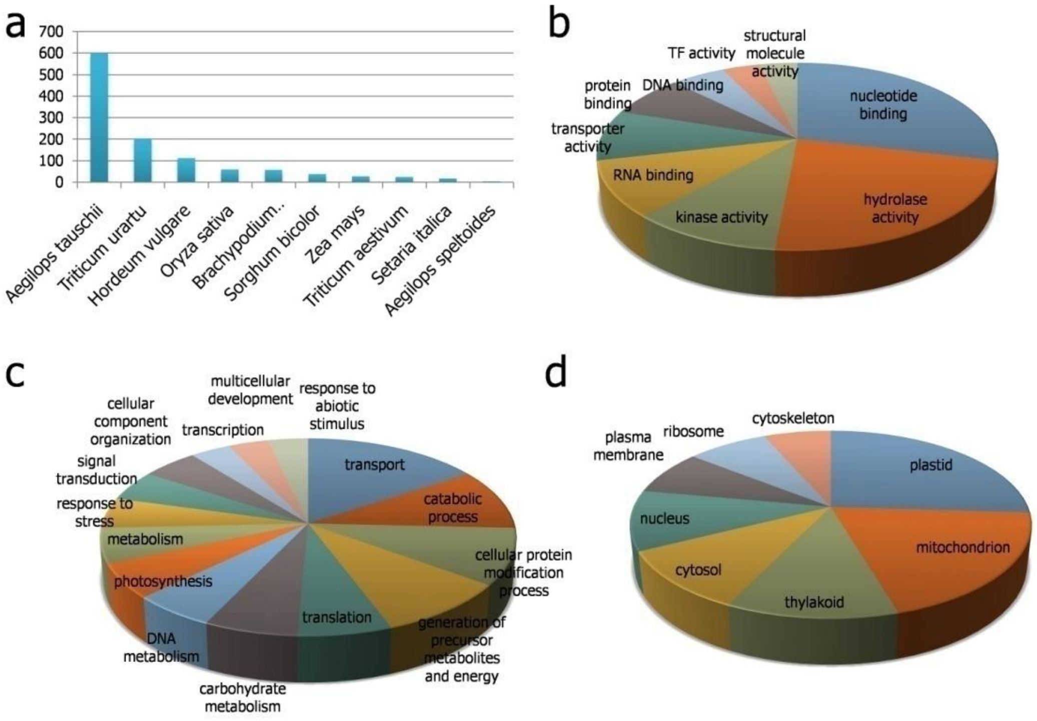


Figure 25. Gene ontology annotations of conserved gene-associated sequences assigned to the 5DS physical map.

The Gene Ontology (GO) terms assigned for Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) terms suggested a variety of processes and functions involving several cellular compartments (Fig. 25b-d), which would be consistent with the transcriptional autonomy of wheat sub-genomes (Mayer *et al.*, 2014). BP terms were enriched for transport, catabolic process and protein modification, among others, while MF terms were highly enriched for nucleotide binding, hydrolase activity and kinase activity, together comprising over 60% of all MF terms. On the other hand, plastid or mitochondrion annotations dominated CC terms, pointing out to energy-related pathways, although related processes of functions were not prominent among BP or MF terms, respectively.

4.3. Comparative analysis of *T. aestivum* 5D with its wild progenitor

4.3.1. Repeat contents of *T. aestivum* and *Ae. tauschii* 5D chromosomes

Although the recent hybridization of the *Ae. tauschii* with *T. turgidum* giving rise to *T. aestivum* occurred relatively recently, repeat contents of *Ae. tauschii* and *T. aestivum* 5D chromosomes exhibited significant differences in terms of repeat families (Fig. 26). While Gypsy superfamily of LTR retrotransposons were the most abundant repeat type for both chromosomes, this superfamily made up almost 46% of all repeats in *T. aestivum* 5D and only 40% that of *Ae. tauschii* 5D chromosome. Conversely, CACTA superfamily of DNA transposons comprised a larger fraction of all repeats in *Ae. tauschii* 5D chromosome (~27%), compared to *T. aestivum* 5D chromosome (~21%). This observation suggested an expansion in retroelements, coupled with the shrinking of DNA transposons.

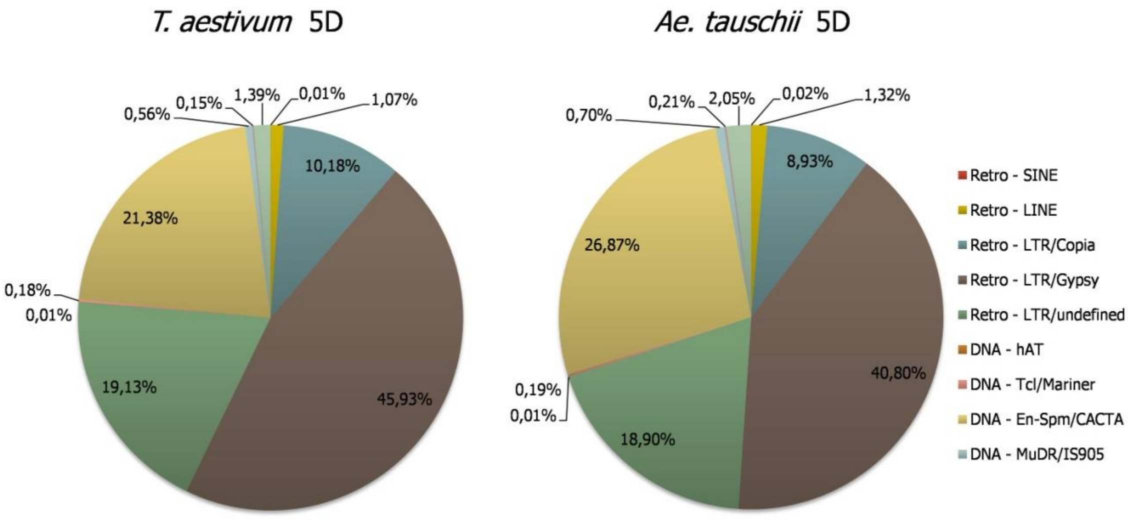


Figure 26. The relative percentages of major repeat superfamilies assessed from survey sequences of *T. aestivum* and *Ae. tauschii* 5D.

The estimated physical sizes of the most abundant repeat families revealed that while Jorge family of DNA transposons, represented by CACTA elements, were roughly same in size, certain Gypsy family of LTR retrotransposons has expanded in *T. aestivum* 5D chromosome (Fig. 27). In particular, Sabrina, Wilma and Sakura elements had >75% growth, accounting for approximately 40 Mb of sequences in total. The 748 Mb-long *T. aestivum* 5D chromosome is 30% larger than its wild progenitor, 577 Mb-long *Ae. tauschii* 5D chromosome. Specific LTR retroelements rather than an overall increase in TE activity, appear to contribute to the genome expansion observed in the polyploidy wheat, consistent with the previous observations of grass genomes differing in repeat composition and abundance (Middleton *et al.*, 2012).

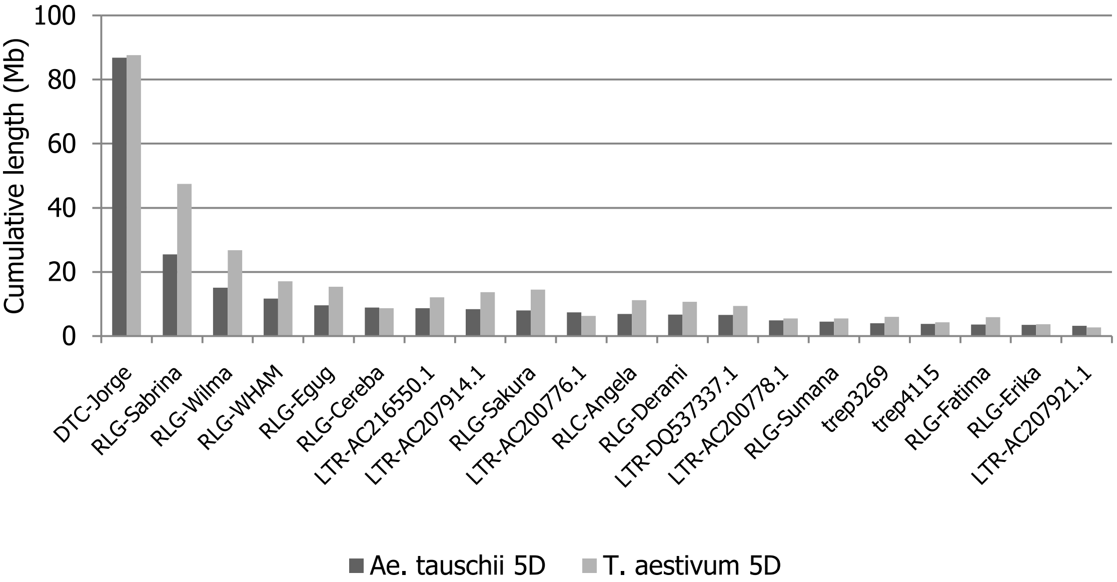


Figure 27. Physical size estimates of the 20 most abundant repeat families in megabases, calculated by multiplying the percentages with the respective chromosome sizes. DTC = DNA transposon, CACTA; RLG = retroelement, LTR, Gypsy; RLC = retroelement, LTR, Copia.

4.3.2. Putative tRNA repertoires of *T. aestivum* and *Ae. tauschii* 5D chromosomes

Putative tRNA repertoires of the two 5D chromosomes, predicted from masked and unmasked survey sequences, revealed a similar distribution (Fig. 28). The unusual abundance for the tRNA^{Lys} species among unmasked sequences was observed for both chromosomes, consistent with an ancient TE capture scenario of some tRNA^{Lys} genes leading to their extensive proliferation in the modern genomes. Whether those tRNA genes captured by and co-expanded with TEs remained functional is elusive at this time. The total number of putative tRNA genes encoded by non-repetitive portion of *T.*

aestivum and *Ae. tauschii* 5D chromosomes, masked against the most recent and comprehensive repeat element database for *Poaceae*, were comparable (153 vs. 142, respectively), despite the considerable size difference between the two chromosomes. In fact, orthologous *Brachypodium* chromosomes 1 and 4, Bd1 and Bd4, 74 and 48 Mb in size, respectively, encoded 166 and 109 putative tRNA genes. Considering the sizes of the respective chromosomes, *T. aestivum* and *Ae. tauschii* 5D chromosomes had putative tRNA gene densities of 1.24 and 2.16 tRNAs/Mb for unmasked, and, 0.20 and 0.25 tRNAs/Mb for masked sequences; while the relatively small *Brachypodium* genome had putative tRNA gene densities of 2.24 and 2.27 for Bd1 and Bd4, respectively. The repetitive content of the *Brachypodium* genome is much lower (The International Brachypodium Initiative, 2010); therefore, wheat genomes are likely to encode far fewer tRNA genes than their close relative. Compared to its wild progenitor, *T. aestivum* 5D had a lower putative tRNA gene content which suggests that tRNA genes had not been major components of genome expansion in wheat.

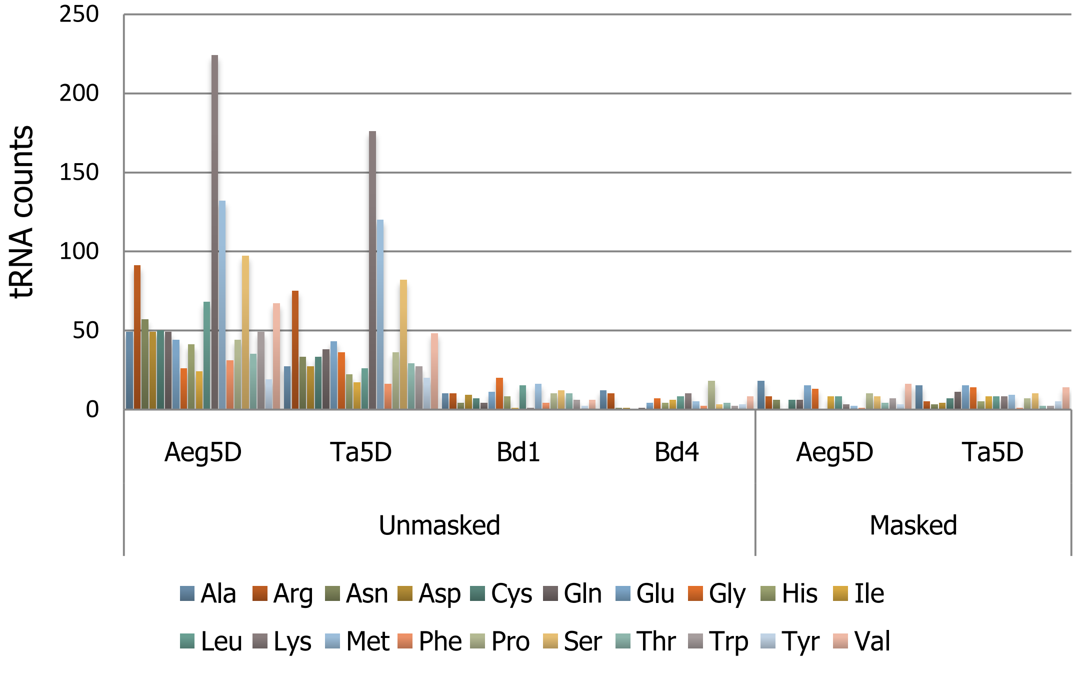


Figure 28. Putative tRNA gene predictions from masked and unmasked *T. aestivum* and *Ae. tauschii* 5D chromosomes, along with orthologous *Brachypodium* chromosomes 1 (Bd1) and 4 (Bd4).

4.3.3. Gene conservation and organization of *T. aestivum* and *Ae. tauschii* 5D chromosomes

Wild progenitors and landraces provide a rich source of genetic diversity for wheat improvement. Recently, draft chromosome sequences of *T. aestivum* suggested 133,090 high-confidence protein coding loci for the entire 17 Gb bread wheat genome (Mayer *et al.*, 2014), while draft genome sequences of its A-genome and D-genome progenitors, *T. urartu* and *Ae. tauschii*, reported 34879 and 34498 protein-coding loci for the estimated 4.94 and 4.36 Gb of entire genomes, respectively, at this level of confidence (Jia *et al.*, 2013; Ling *et al.*, 2013). Considering the average gene lengths of 2000, 3207 and 2772 bases estimated for *T. aestivum*, *T. urartu* and *Ae. tauschii*, respectively, these predicted gene loci correspond to genic fractions of 1.57% for the bread wheat genome, and 2.19% and 2.26% for the wild progenitors (Vitulo *et al.*, 2011; Ling *et al.*, 2013; Luo *et al.*, 2013).

A total of 4289 gene models (3147 high-confidence) were constructed from *T. aestivum* 5D survey sequences with homologies to annotated proteins of model grasses *Brachypodium*, rice and sorghum, as well as to related grass UniGene/Uniprot sequences. These suggest a genic fraction of 0.84-1.15% assuming an average coding sequence length of 2000 bases (Vitulo *et al.*, 2011). Although *T. aestivum* 5D survey sequence may have overlooked a fraction of genuine genes due to the coverage, it is likely that *T. aestivum* 5D chromosome harbors fewer protein-coding loci than its wild progenitor, *Ae. tauschii* 5D chromosome.

Gene conservation patterns also revealed significant differences between *T. aestivum* and *Ae. tauschii* 5D chromosomes (Fig. 29). Consistently, both chromosomes revealed orthologous relationships with *Brachypodium* chromosomes 1 & 4, rice chromosomes 3, 9 & 12, and sorghum chromosomes 1 and 2; although, gene conservation was observed to a lesser extent for *Ae. tauschii* 5D chromosome. The most striking difference, however, was the orthologous regions along rice chromosome 3 (Fig. 29). While *T. aestivum* 5D orthologous sequences were dispersed along the chromosome, *Ae. tauschii* 5D orthologous sequences were concentrated at the distal region of this chromosome. This observation suggest extensive rearrangements in *T.*

aestivum 5D chromosome, involving genes orthologous to rice chromosome 3, which might have led to several breaks in synteny between *T. aestivum* and rice.

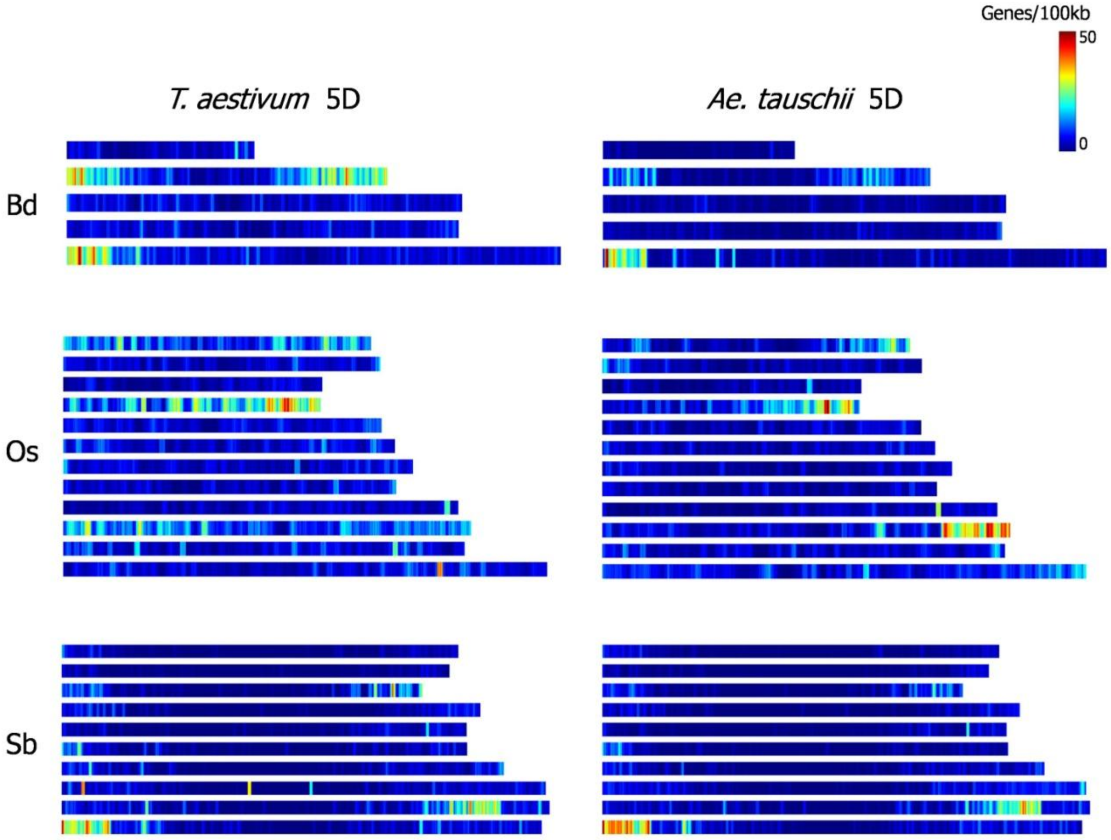


Figure 29. Orthologous relationships of *T. aestivum* and *Ae. tauschii* 5D chromosomes with the model grasses, *Brachypodium distachyon* (Bd), *Orzya sativa* (Os) and *Sorghum bicolor* (Sb). Bd (5), Os (12) and Sb (10) chromosomes were ordered, from bottom to top, in an ascending order.

Genome zippers of *T. aestivum* and *Ae. tauschii* 5D chromosomes were compared, along with the published gene order for orthologous barley chromosome 5H (Mayer *et al.*, 2011), to explore genome rearrangements (Fig. 30). In accordance with the close evolutionary relationships between *T. aestivum* and *Ae. tauschii* 5D chromosomes, the virtual gene orders on both chromosomes were mostly colinear, with few putative translocations (Fig. 30, pink lines). This colinearity was largely preserved with the virtual gene order of orthologous 5H chromosome of barley, the close relative of wheat from the *Triticeae* tribe. Interestingly, a group of genes involved in an apparent inversion between barley 5H chromosome and *Ae. tauschii* 5D chromosome (Fig. 30, inversed bundle of cyan links) was found in colinear order between *Ae. tauschii* 5D and

T. aestivum 5D chromosomes, which may point out to a species-specific inversion, occured after wheat and barley had diverged from their last common ancestor.

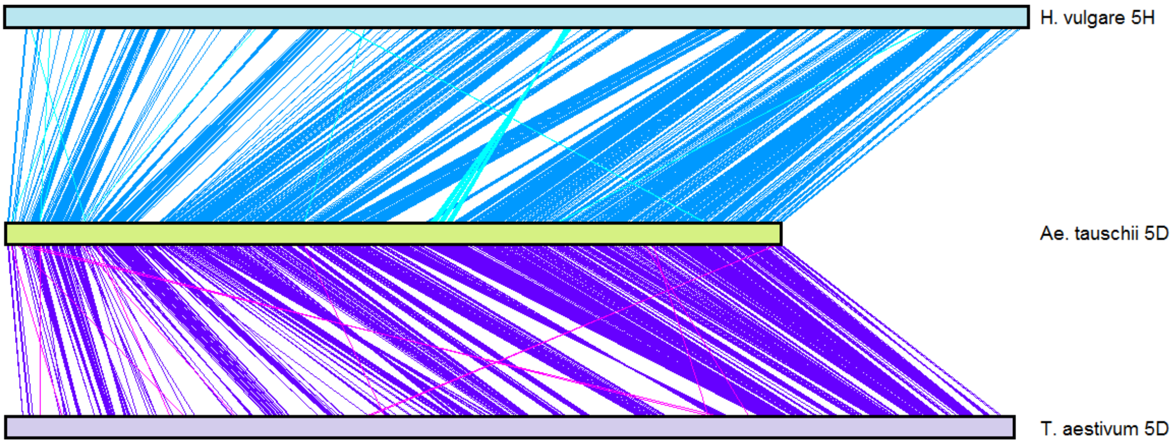


Figure 30. Comparative maps of *H. vulgare* 5H, *Ae. tauschii* 5D and *T. aestivum* 5D chromosomes.

It should be noted that the genome zippers do not represent the ultimate order of the entire gene repertoires. The reference genome sequences of wheat and barley may add up to these putative rearrangements and define precise relationships between these genomes.

4.3.4. Functional gene spaces of *T. aestivum* and *Ae. tauschii* 5D chromosomes

In order to compare functional gene spaces of *T. aestivum* and *Ae. tauschii* 5D chromosomes, survey sequences masked against *Poaceae* repeat elements and assembled with the same parameters using gsAssembler tool of Newbler software 2.6 (454 Life Sciences). Both assemblies were compared against annotated proteins from *Brachypodium*, rice, sorghum; high-confidence barley proteins; *Ae. tauschii* UniProt sequences; and *T. aestivum* UniGene sequences to reveal gene-associated sequences with adjusted parameters. Finally, these sequences were annotated based on all *Viridiplantae* proteins for Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) terms (Fig. 31).

Functional annotation of gene-associated survey sequences from *T. aestivum* and *Ae. tauschii* 5D chromosomes revealed enrichments for different BP, MF and CC terms. Energy related BP terms, ‘generation of precursor metabolites and energy’ and ‘photosynthesis’, were enriched among *T. aestivum* 5D sequences. These enrichments were also reflected in CC terms, where ‘plastid’ and ‘mitochondrion’ associated sequences were leading among *T. aestivum* 5D sequences. In MF terms, ‘structural molecular activity’ and ‘protein binding’ were more prominent for *T. aestivum* 5D chromosome, while *Ae. tauschii* 5D chromosome was enriched in ‘chromatin binding’. These observations suggest that *T. aestivum* 5D chromosome may harbour genes related to photosynthetic machinery and energy metabolism to a greater extent, compared to its wild progenitor.

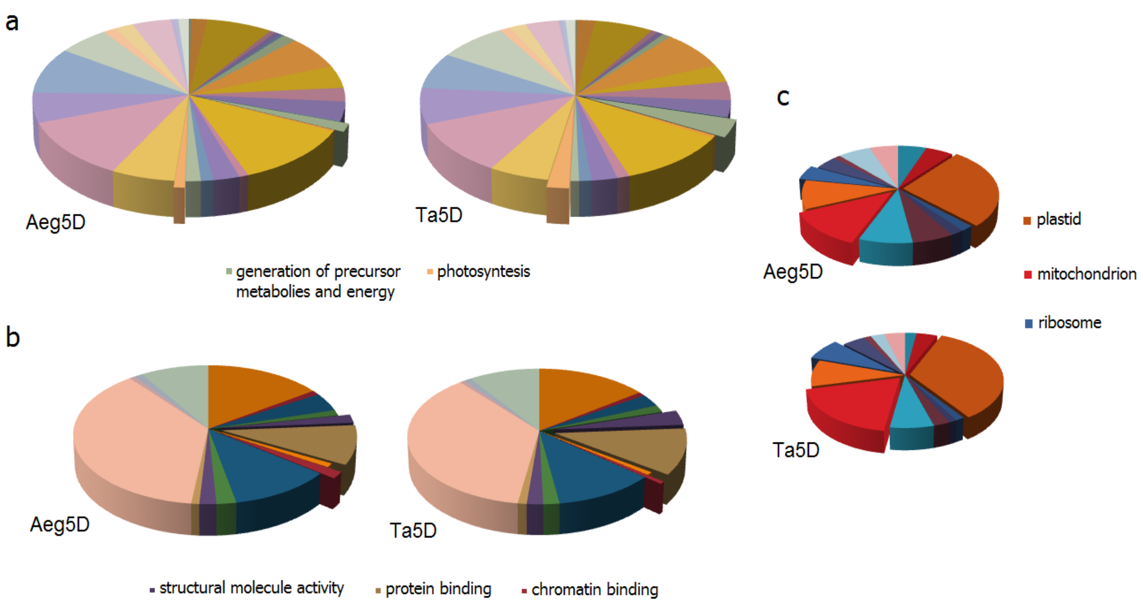


Figure 31. Gene-ontology annotations of *T. aestivum* and *Ae. tauschii* 5D gene-associated sequences.(a) Biological Process, (b) Molecular Function, and, (c) Cellular Component terms. Only GO terms with differential enrichments are emphasized.

5. DISCUSSION

Wheat is a hardy, cereal grain crop that is capable to grow across a wide range of environments. Accordingly, wheat is the most extensively harvested crop worldwide and ranks the third, after rice and maize, in terms of production, placing this crop as an essential component of human and animal nutrition. Growing world population, global climate changes and the increasing use of crops in biofuel industry demand significant increases in crop production in the upcoming decades. Genetic diversity forms the basis for crop improvement for better production; however, domestication and subsequently, thousands of years of agricultural practices have considerably narrowed the gene pools of modern cultivars. Genome sequencing and genomics research are promising tools to explore and exploit natural genetic diversity found within wild populations, to create diversity within elite populations or to elucidate genome biology and functioning on a grand scale.

Despite its agronomic importance, wheat genomics has been largely hampered by the wheat genome size and complexity. Bread wheat, *Triticum aestivum*, accounting for >95% of all wheat production, has a hexaploid genome, composed of >80% repetitive elements (Feuillet *et al.*, 2007). The bread wheat genome is 17 Gb in size, almost three times as large as the human genome. Tackling this huge and repetitive genome had been considered intractable until recently. The International Wheat Genome Sequencing Consortium has been established to handle this daunting task and employed a chromosome-by-chromosome approach to elucidate the genome sequence and structure of bread wheat. In this study, the structure and organization of 5D chromosome of bread wheat were investigated, through a combination of genome sequencing and physical mapping approaches, and 5D chromosome was compared to its wild progenitor in order to gain insights into wheat genome evolution.

Survey sequencing of *T. aestivum* L. cv. Chinese Spring chromosome 5D indicated that up to 82% of the chromosome was made up of repetitive elements, as assessed by the similarity searches against known *Poaceae* repeat elements, consistent with previous observations from both BAC-based and chromosome-specific sequencing studies from wheat (Choulet *et al.*, 2010; Vitulo *et al.*, 2011; Brenchley *et al.*, 2012; Hernandez *et al.*, 2012; Lucas *et al.*, 2014; Mayer *et al.*, 2014). As expected, LTR retrotransposons were the most abundant repeat type, accounting for over three fourths of all repeat annotations. In fact compared to the 5D chromosome of its wild progenitor, *Ae. tauschii*, LTR elements appeared to have expanded in *T. aestivum* 5D chromosome (Fig. 25) (Akpinar *et al.*, 2014; Lucas *et al.*, 2014). Nested insertions within retroelements do not interfere with the proliferation of repetitive elements and are implicated as a major component of wheat genome expansion (Li *et al.*, 2004). It is likely that the hexaploid bread wheat genome could have tolerated nested insertion to a greater extent than its diploid progenitor, leading to the expansion of certain retroelements in *T. aestivum* 5D chromosome, but not in *Ae. tauschii* 5D chromosome.

Non-repetitive survey sequences from 5DS and 5DL chromosome arms revealed significant matches to 1,493 and 2,829 annotated proteins, respectively, from model grasses *Brachypodium distachyon*, rice and sorghum, a subset of which were also supported by UniGene/UniProt sequences from related monocot species. Interestingly, 5DS survey sequences matched a high proportion of genes conserved between *B. distachyon* and rice, but not sorghum, while 5DL survey sequences matched a high proportion of genes conserved between *B. distachyon* and sorghum, but not rice (Fig. 9). This observation can be explained by variable mutation rates at different regions on chromosome 5D acted at different stages of evolutionary history, as *Panicoidae* family, to which sorghum belongs, has diverged from the *Pooidae* family that includes *Brachypodium* and wheat, earlier than the *Ehrhartoideae* family of rice (The International Brachypodium Initiative, 2010). Comparative analyses between *T. aestivum* 5D chromosome and the model grasses revealed syntenic regions on *Brachypodium* chromosomes 1 & 4, rice chromosomes 3, 9 & 12 and sorghum chromosomes 1, 2 & 8, consistent with large-scale patterns of synteny observed previously (The International Brachypodium Initiative, 2010). Intriguingly, while sequences from *T. aestivum* 5D chromosome orthologous to rice chromosome 3 were dispersed across the chromosome, orthologous sequences from its wild progenitor *Ae.*

tauschii 5D chromosome revealed a more concentrated conserved block on the distal region (Fig. 29). This observation suggested extensive rearrangements on *T. aestivum* 5D chromosome, causing several breaks into this syntenic region. Diploidization following polyploidization may have resulted in such extensive rearrangements leading to gene losses in the hexaploid wheat genome, in contrast to its diploid progenitor (Murat *et al.*, 2014). Conserved blocks of few genes were also observed on non-orthologous *Brachypodium* chromosomes 2 & 5 and sorghum chromosomes 3, 5 & 6, pointing out to small scale genome rearrangements (Fig. 15). The genome zipper approach used to construct a virtual gene order for *T. aestivum* 5D chromosome revealed further small scale rearrangements particularly on 5DL, in addition to an apparent inversed block on 5DS (Fig. 16).

Functional characterization of 5D gene space through both gene models predicted from survey sequences or gene-associated microarray probes revealed a wide array of biological processes, molecular functions and cellular components (Fig. 10, 26), consistent with the transcriptional autonomy of the sub-genomes (Mayer *et al.*, 2014). Hydrolase activity, intriguingly, had a marked abundance among both functional characterization attempts. Curiously, hydrolase activity was also prominent in the secretome of an apple pathogen *Venturia inaequalis* which is closely related to the wheat pathogen *Pyrenophora tritici-repentis* (Thakur *et al.*, 2013), and reported from the transcriptome of the wheat pathogen *Heterodera avenae* (Kumar *et al.*, 2014). Whether these hydrolase activity related 5D sequences are associated with disease responses remain unclear at the moment.

Putative tRNA predictions revealed an unusual abundance for the tRNA^{Lys} species among repetitive sequences, speculated to result from co-proliferation following an ancient TE-capture, as observed previously for 6B chromosome (Tanaka *et al.*, 2013). Among the non-repetitive sequences, however, the most abundant tRNA species were tRNA^{Met}, tRNA^{Val}, tRNA^{Gly}, tRNA^{Gln} and tRNA^{Glu} species. A species-specific preference for the second position in a protein sequence is speculated to exist for the majority of proteins starting with a Methionine (Met) residue, with implicated effects on the translation, and thus regulation, of the protein. A strong preference was reported in *Arabidopsis thaliana* that favoured Alanine (Ala) residues, followed by Serine (Ser) residues, at the second position (Shemesh *et al.*, 2010). Putative tRNA gene abundances may reflect a similar preference of amino acids in *T. aestivum* 5D proteome (Fig. 11).

In addition to survey sequences, physical maps provide valuable insights into the wheat genome structure and organization and serve as a framework for reference sequencing studies. In this study, a physical map of chromosome 5DS was constructed which contained 164 contigs, of which 120 had 6 or more clones, with an average contig size of 1078 kb. The longest contig was 6649 kb, comparable to the LTC-maps of 1AL, 1BL and 1BS, for which the longest contig size ranged between 5.8 Mb to over 7 Mb (Lucas *et al.*, 2013; Philippe *et al.*, 2013; Raats *et al.*, 2013). A variety of molecular markers were used to anchor 105 of the 164 contigs through PCR screening or microarray hybridization. PCR screening yielded success rates of 56% for SSR, 92% for COS, and 86% for EST markers that were previously mapped the 5DS chromosome. The relatively low success rate for the SSR markers could stem from the sequence divergence between *T. aestivum* L. cv. Chinese Spring used to isolate the 5DS chromosome and the cultivars which the SSR markers were designed from. Deriving from conserved genic sequences, COS and EST markers were relatively easy to anchor to the 5DS map, as expected. As a relatively recent approach, microarray hybridization was utilized to 647, 1,122, 1,615 and 3,878 UniGenes to the physical maps of 1AS, 1AL, 1BL and 1BS, respectively (Breen *et al.*, 2013; Lucas *et al.*, 2013; Philippe *et al.*, 2013; Raats *et al.*, 2013). While all these studies used the same pre-designed NimbleGen 40k UniGene microarray (Choulet *et al.*, 2010), differences in chromosome arm sizes and the stringency levels applied to the microarray data might have resulted in the differences in the number of UniGenes assigned to each map through this approach. In addition to the NimbleGen 40k UniGene array, 1BL physical map also included putatively assigned 3912 ISBP markers, through a custom-design 17k ISBP NimbleGen array (Philippe *et al.*, 2013). A custom-design Agilent array incorporating probes from a variety of gene- and marker-associated sequences enabled 1762 markers to be putatively assigned to the 5DS map unambiguously. Seventeen of the 18 markers physically anchored to the 5DS map and also included in the microarray design were in complete agreement, indicating that the stringency levels used in the interpretation of the microarray data gave reliable results.

The percentage of anchored contigs (105 out of 164, 64%) was improved compared to the 1AL physical map constructed by our group previously (Lucas *et al.*, 2013). More recent wheat physical maps reported anchored contig percentages of 74-79%, similar to the 5DS physical map (83%, excluding the non-informative short contigs). The anchored molecular markers yielded a marker density of ~10.5 markers/Mb of the chromosome arm. This marker density was close to that of the 1BL physical map (11 markers/Mb), and exceeded the 1BS physical map (10.1 markers/Mb), indicating a high-quality map saturated with molecular markers. This highly saturated 5DS map is likely to provide a useful resource for future map-based cloning or marker-assisted genomics studies.

Of the anchored contigs, 80 were ordered along the 4 deletion bins of 5DS, covering 53.6% of the chromosome arm at a cumulative length of 138.3 Mb. The coverage of the chromosome arm by the mapped contigs exceeded that of 1BL, at 48% (Philippe *et al.*, 2013), despite lower coverage of the chromosome arm by the physical map.

The 1BL physical map revealed a bias in the number of clones in mapped contigs, where telomeric contigs contained far fewer clones (Philippe *et al.*, 2013). Such a bias was not observed for the 5DS physical map, however, telomeric contigs tended to have smaller cumulative lengths. The most proximal bin, 0-0.63, contained relatively larger contigs, including the longest contig (CTG138), whereas, more than half of the contigs smaller than 1 Mb were assigned to the most distal deletion bin, 0.78-1.00. Gene densities are observed to increase towards the telomeric ends of *Triticeae* chromosomes (Choulet *et al.*, 2010). Therefore, telomeric ends are more likely to be populated by genetically mapped markers, which may have facilitated the mapping of shorter contigs to the telomeric bins of the 5DS physical map.

The physical sizes of 5DS deletion bins estimated from the cumulative lengths of the contigs mapped to each respective bin suggested discrepancies with the cytogenetic estimates for the proximal deletion bins. While the physical size estimate was close to the cytogenetic estimate for the most distal deletion bin, 0.78-1.00 (19.2% vs. 22%), physical size estimates were considerable underestimated for 0.67-0.78 and 0-0.63 bins, at 15 Mb (5.7%) and 138 Mb (53.5%), respectively, compared to cytogenetic estimates.

Strikingly, the cytogenetically smallest deletion bin, 0.63-0.67, estimated to contain only 4% of the 5DS was estimated to cover 55 Mb (21%) based on mapped contigs.

The inconsistencies between the cytogenetic and physical size estimates may be, in part, explained by the low coverage of the chromosome arm by the mapped contigs (54%). Additionally, unequal representation of the deletion bins by the genetically mapped molecular markers may have led to an underestimation of proximal deletion bins in physical size estimates. However, gene densities assessed by positively assigned gene-associated microarray probes paint an intriguing picture on this issue. Of the 1306 gene-associated markers putatively assigned to the contigs of the 5DS physical map, 95, 41, 105 and 231 markers were allocated to 0.78-1.00, 0.67-0.78, 0.63-0.67 and 0-0.63 bins, respectively. These correspond to gene density estimates between 3.17-5.17 genes/Mb along the chromosome arm, assuming estimated physical sizes. A similar high gene density estimate was reported for the 1AS physical map, at 5.1 genes/Mb (Breen *et al.*, 2013). However, the gene density rises up to 19 genes/Mb for the 0.63-0.67 bin, if the cytogenetic size estimates are taken into account, which is highly unlikely. Additionally, the 0.67-0.78 bin, represented by only two contigs, revealed the highest gene density, further corroborating the adequacy of the physical size estimates, regardless of the unequal distribution of genetically mapped molecular markers across different deletion bins. Therefore, the physical size estimates is concluded to reflect the actual sizes of the deletion bins, albeit with moderate accuracy due to the low chromosome coverage by the mapped contigs.

6. CONCLUSIONS

Advances in next-generation sequencing technologies have greatly accelerated genomics research in wheat. Combined with the genetic stocks, these genomics resources are anticipated to provide efficient tools for molecular breeding studies for crop improvement. As our understading of its genome expands, we are able to dissect wheat genome evolution to greater extents, which provides critical clues into its domestication at the molecular level, which, in turn, can be utilized in wheat breeding. Large-scale sequences generated by ongoing efforts present a rich resource from which polymorphisms can be readily identified and screened through molecular markers, enabling the utilization of natural genetic diversity found within wild germplasms. Additionally, BAC-based physical maps provide resources that can be readily utilized; physical segments of chromosome harboring a gene-of-interest can be identified using linked molecular markers and a chromosome-walking approach can be initiated for positional cloning of genes for functional characterization. In the long term, these genomics studies will pave the way for the ultimate goal of unraveling the bread wheat genome to a reference quality. Accordingly, the genomics resources presented in this study will be an integral part of this long term goal.

APPENDIX A

Summary of all known repeat annotations, classified by superfamily, for 5DS and 5DL.

Chromosome arm			**5DS**		**5DL**	
Total no. of reads			937266		2271393	
			bp	%	bp	%
Total read length			347752702	100	791451651	100
Total length masked as repeats			245325262	70.55	553392605	69.92
Repeat type	Order	Superfamily				
Retroelements			**197996023**	**56.94**	**462267357**	**58.41**
	LTR		194642140	55.97	453806906	57.34
		Copia	29371613	8.45	75960597	9.6
		Gypsy	161788197	46.52	371664487	46.96
		Unclassified	3482330	1.0	6181822	0.78
	SINEs		33240	0.01	85271	0.01
	LINEs		3320643	0.95	8375180	1.06
		CRE/SLACS	0	0	0	0
		L2/CR1/Rex	0	0	0	0
		R1/LOA/Jockey	0	0	0	0
		R2/R4/NeSL	0	0	0	0
		RTE/Bov-B	106	0	1212	0
		L1/CIN4	503852	0.14	1649518	0.21
DNA transposons			**42420333**	**12.2**	**81486264**	**10.3**
	TIR (Terminal Inverted Repeats)					
		hAT (hobo-Activator)	156157	0.04	651834	0.08
		CACTA (En-Spm)	37881304	10.89	70737844	8.94
		Tc1/Mariner	1338606	0.38	3139514	0.40
		Mutator	1903993	0.55	4125581	0.52
		PIF/Harbinger	699077	0.20	1751131	0.22
	Helitron (Rolling Circles)		261914	0.07	862142	0.11
Unclassified interspersed repeats			**3837426**	**1.1**	**7075768**	**0.89**
Other recurring elements						
		Small RNAs	16963	0.005	49379	0.006
		Satellites	14109	0.004	27081	0.003
		Simple repeats	561033	0.16	1166830	0.15
		Low complexity	496338	0.14	1369305	0.17

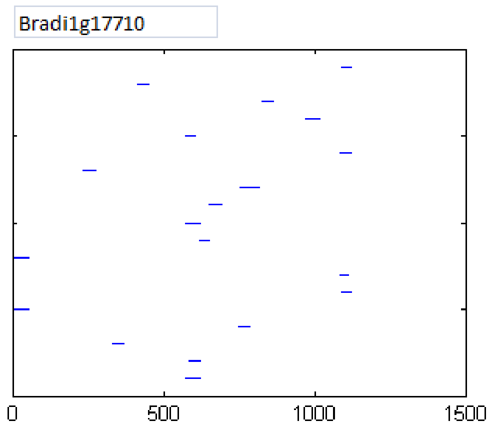
APPENDIX B

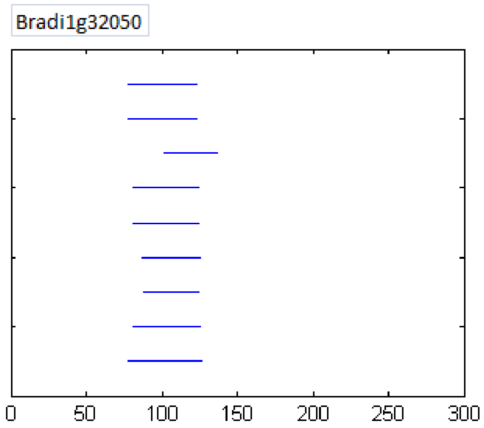
All tRNA counts for the masked 5DS and 5DL survey sequences and for all unmasked 5D sequences. Ta5DS: *T. aestivum* 5DS; Ta5DL: *T. aestivum* 5DL.

tRNA type	**Masked**			**Unmasked**
	Ta5DS	**Ta5DL**	**Ta5D**	**Ta5D**
Ala	8	19	27	27
Arg	0	5	5	75
Asn	0	4	4	33
Asp	0	12	12	27
Cys	2	12	14	33
Gln	12	21	33	38
Glu	5	26	31	43
Gly	10	25	35	36
His	3	14	17	22
Ile	1	16	17	17
Leu	2	23	25	26
Lys	0	3	3	176
Met	9	51	60	120
Phe	0	1	1	16
Pro	1	9	10	36
Ser	3	15	18	82
Thr	1	0	1	29
Trp	1	3	4	27
Tyr	1	4	5	20
Val	17	28	45	48

APPENDIX C

Coverages of two non-syntenic genes by 5D survey sequences, given as examples. Two *Brachypodium* genes, Bradi1g17710 and Bradi1g32050 were matched by 19 5DS and 9 5DL survey sequence reads, respectively. Bradi1g17710 was covered evenly along its length, indicating that its wheat ortholog is possibly a genuine gene. In contrast, matching reads were clustered close to the 5' end of Bradi1g32050, indicating a truncated ortholog on wheat chromosome 5DL that is likely a pseudogene.





APPENDIX D

List of all orthologous genes putatively rearranged after the divergence of *Triticeae*.

5DS

Brachypodium ortholog	Rice ortholog	Sorghum ortholog	# of reads matching	Predicted function	Source organism
Bradi1g02740	Os03t0833700	Sb02g027580.1	2	putative RNA 3'-terminal phosphate cyclase-like protein	*Aegilops tauschii*
Bradi1g11490	Os03t0713000	Sb01g010250.1	2	Threonine dehydratase biosynthetic, chloroplastic	*Aegilops tauschii*
Bradi1g11640	Os03t0701000	Sb01g011010.1	1	Importin-5	*Aegilops tauschii*
Bradi1g15330	Os03t0587200	Sb01g015490.1	4	Kinesin-like protein KIF15	*Aegilops tauschii*
Bradi1g17710	Os02t0167700	Sb04g004540.1	19	Cullin-associated NEDD8-dissociated protein 1	*Aegilops tauschii*
Bradi1g21030	Os03t0583900	Sb01g015670.1	1	Endoribonuclease Dicer-1-like protein	*Aegilops tauschii*
Bradi1g23880	Os12t0597400	Sb08g020230.1	2	hypothetical protein F775_30608	*Aegilops tauschii*
Bradi1g25780	Os07t0531700	Sb02g035270.1	7	hypothetical protein F775_21310	*Aegilops tauschii*
Bradi1g28790	Os07t0447800	Sb02g010840.2	2	Phosphomannomutase/phosphoglucomutase	*Aegilops tauschii*
Bradi1g35630	Os12t0568800	Sb08g018670.1	3	Importin subunit beta-1	*Aegilops tauschii*
Bradi1g48880	Os06t0156900	Sb10g003920.1	2	hypothetical protein F775_52471	*Aegilops tauschii*
Bradi1g59970	Os06t0498500	Sb01g032360.1	7	Nucleolar complex 3-like protein	*Aegilops tauschii*
Bradi1g61680	Os03t0364500	Sb01g034550.1	1	hypothetical protein F775_29736	*Aegilops tauschii*
Bradi2g13137	Os01t0374200	Sb03g001460.1	2	hypothetical protein F775_07206	*Aegilops tauschii*
Bradi2g44510	Os01t0634900	Sb03g028940.1	2	hypothetical protein F775_01416	*Aegilops tauschii*
Bradi2g61830	Os01t0966700	Sb03g047060.1	2	Beta-fructofuranosidase, insoluble isoenzyme 4	*Aegilops tauschii*
Bradi3g05950	Os12t0541500	Sb02g041940.1	5	Elongation factor Ts	*Aegilops tauschii*
Bradi3g11460	Os02t0312700	Sb04g010120.1	3	ATP synthase mitochondrial F1 complex assembly factor 1	*Aegilops tauschii*
Bradi3g43010	Os01t0974000	Sb03g047410.1	4	hypothetical protein F775_21918	*Aegilops tauschii*
Bradi3g54980	Os02t0812400	Sb04g036790.1	10	Translation initiation factor eIF-2B subunit epsilon	*Aegilops tauschii*
Bradi3g55920	Os02t0824000	Sb04g037860.1	4	hypothetical protein F775_16833	*Aegilops tauschii*
Bradi4g41960	Os12t0160900	Sb08g003780.1	3	Protein FAR1-RELATED SEQUENCE 5	*Aegilops tauschii*

5DL

Brachypodium ortholog	Rice ortholog	Sorghum ortholog	# of reads matching	Predicted function	Source organism
Bradi1g16770	Os07g0691200	Sb02g002970.1	1	D-alanine--D-alanine ligase	*Aegilops tauschii*
Bradi1g19280	Os07g0296200	Sb02g032170.1	2	TATA-binding protein-associated factor 2N	*Triticum urartu*
Bradi1g35960	Os05g0370600	Sb10g024200.1	2	Multiple C2 and transmembrane domain-containing protein 1	*Aegilops tauschii*

Bradi1g36730	Os06g0602400	Sb10g023440.1	2	DEAD-box ATP-dependent RNA helicase 52A	*Aegilops tauschii*
Bradi1g53060	Os07g0264100	Sb02g008660.1	3	Glyoxylate reductase	*Aegilops tauschii*
Bradi1g53085	Os07g0260400	Sb02g008130.1	9	Phospholipase D delta	*Aegilops tauschii*
Bradi1g54250	Os07g0208500	Sb02g006290.1	1	Putative cellulose synthase A catalytic subunit 8 (UDP-forming)	*Aegilops tauschii*
Bradi1g54900	Os06g0151600	Sb10g003620.1	2	DNA repair radA-like protein	*Aegilops tauschii*
Bradi1g59910	Os03g0417900	Sb01g032420.1	3	hypothetical protein F775_28852	*Aegilops tauschii*
Bradi1g61130	Os03g0379100	Sb01g034010.1	3	hypothetical protein F775_20122	*Aegilops tauschii*
Bradi1g61750	Os03g0363600	Sb01g034610.1	2	Sugar transporter ERD6-like protein 16	*Aegilops tauschii*
Bradi1g63160	Os03g0336300	Sb01g036110.1	3	Insulin-degrading enzyme	*Aegilops tauschii*
Bradi1g67790	Os03g0254800	Sb01g040790.1	2	Chorismate synthase 1, chloroplastic	*Aegilops tauschii*
Bradi1g69900	Os03g0215200	Sb01g042850.1	2	DL related protein	*Triticum aestivum*
Bradi1g70057	Os03g0213400	Sb01g042980.1	1	Activating signal cointegrator 1 complex subunit 3	*Aegilops tauschii*
Bradi1g71600	Os08g0550100	Sb07g024800.1	1	Putative 26S proteasome non-ATPase regulatory subunit 3	*Aegilops tauschii*
Bradi1g72490	Os03g0182400	Sb01g045110.1	2	hypothetical protein F775_06981	*Aegilops tauschii*
Bradi1g72620	Os03g0180700	Sb01g045200.1	4	Chitobiosyldiphosphodolichol beta-mannosyltransferase	*Aegilops tauschii*
Bradi1g74580	Os03g0158200	Sb01g046570.1	6	DEAD-box ATP-dependent RNA helicase 38	*Aegilops tauschii*
Bradi2g18270	Os09g0241100	Sb10g019730.1	3	hypothetical protein F775_04652	*Aegilops tauschii*
Bradi2g62760	Os01g0977600	Sb05g002400.1	9	RING finger protein 160	*Aegilops tauschii*
Bradi3g03990	Os02g0150100	Sb01g035760.1	3	DEAD-box ATP-dependent RNA helicase 35A	*Aegilops tauschii*
Bradi3g17770	Os10g0141900	Sb01g026340.1	6	Queuine tRNA-ribosyltransferase subunit qtrtd1	*Aegilops tauschii*
Bradi3g19890	Os06g0255700	Sb07g002945.1	6	DNA repair protein rhp54	*Aegilops tauschii*
Bradi3g32160	Os10g0537600	Sb01g030480.1	2	hypothetical protein F775_01806	*Aegilops tauschii*
Bradi3g38690	Os09g0515800	Sb04g037270.1	4	GTPase-activating protein gyp1	*Triticum urartu*
Bradi3g43050	Os04g0492300	Sb06g021120.1	2	DNA-directed RNA polymerase III subunit RPC1	*Aegilops tauschii*
Bradi3g49150	Os02g0623500	Sb01g002140.1	2	Lysyl-tRNA synthetase	*Aegilops tauschii*
Bradi4g02010	Os03g0850100	Sb06g001010.1	2	CTD small phosphatase-like protein 2	*Aegilops tauschii*
Bradi4g07541	Os12g0477700	Sb08g014300.1	1	hypothetical protein F775_02086	*Aegilops tauschii*
Bradi4g08907	Os09g0266400	Sb02g015540.1	2	GPI inositol-deacylase	*Aegilops tauschii*
Bradi4g13940	Os11g0610900	Sb05g024270.1	4	Seryl-tRNA synthetase	*Aegilops tauschii*
Bradi4g15450	Os03g0804800	Sb05g022470.1	2	chaperonin family theta subunit	*Triticum aestivum*
Bradi4g17270	Os11g0528400	Sb05g019790.1	1	hypothetical protein F775_14529	*Aegilops tauschii*
Bradi4g19937	Os07g0280200	Sb02g020270.1	2	ABP-1	*Triticum aestivum*
Bradi5g09500	Os01g0896800	Sb01g048270.1	1	60S ribosomal protein L5	*Triticum aestivum*

APPENDIX E

The details of the preliminary 5DS physical map, constructed by LTC. Contigs are sorted by their lengths. N50noshort*: N50 value excluding short contigs (<6 clones).

Contig	# of clones	Length (kb)	Length (CB)	Cumulative kb	
Ctg138	722	6649	5541	6649	
Ctg78	572	5186	4322	11835	
Ctg68	583	4980	4150	16815	
Ctg135	501	4422	3685	21237	
Ctg102	434	4328	3607	25565	
Ctg99	438	4034	3362	29599	
Ctg93	396	3897	3248	33496	
Ctg156	373	3585	2988	37081	
Ctg144	372	3530	2942	40611	
Ctg87	329	3387	2823	43998	
Ctg127	393	3369	2808	47367	
Ctg115	384	3277	2731	50644	
Ctg157	358	3200	2667	53844	
Ctg96	303	2942	2452	56786	
Ctg125	317	2929	2441	59715	
Ctg128	305	2793	2328	62508	
Ctg100	318	2722	2269	65230	
Ctg155	246	2673	2228	67903	
Ctg123	295	2613	2178	70516	
Ctg70	231	2595	2163	73111	
Ctg126	246	2499	2083	75610	
Ctg124	292	2462	2052	78072	
Ctg89	252	2298	1915	80370	
Ctg86	263	2234	1862	82604	
Ctg65	206	2226	1855	84830	N50noshort*
Ctg142	260	2187	1823	87017	
Ctg143	251	2173	1811	89190	N50
Ctg134	214	2148	1790	91338	
Ctg53	194	2036	1697	93374	
Ctg140	217	1995	1663	95369	
Ctg158	224	1976	1647	97345	
Ctg149	256	1935	1613	99280	
Ctg111	211	1914	1595	101194	
Ctg113	213	1904	1587	103098	
Ctg136	218	1833	1528	104931	
Ctg97	192	1785	1488	106716	
Ctg84	149	1774	1479	108490	
Ctg116	209	1666	1389	110156	

Ctg2	150	1658	1382	111814	
Ctg91	161	1506	1255	113320	
Ctg14	160	1497	1248	114817	
Ctg82	136	1465	1221	116282	
Ctg5	109	1378	1149	117660	
Ctg137	158	1335	1113	118995	
Ctg77	138	1299	1083	120294	
Ctg3	153	1269	1058	121563	
Ctg83	171	1224	1020	122787	
Ctg146	104	1216	1014	124003	
Ctg129	98	1210	1009	125213	
Ctg81	121	1183	986	126396	
Ctg121	113	1182	985	127578	
Ctg122	135	1167	973	128745	
Ctg19	94	1167	973	129912	
Ctg148	124	1107	923	131019	
Ctg55	115	1032	860	132051	
Ctg90	96	1015	846	133066	
Ctg57	74	1004	837	134070	
Ctg62	109	1002	835	135072	
Ctg1	111	963	803	136035	
Ctg98	86	956	797	136991	
Ctg117	84	955	796	137946	
Ctg17	95	946	789	138892	
Ctg92	99	919	766	139811	
Ctg95	98	915	763	140726	
Ctg22	78	913	761	141639	
Ctg162	101	858	715	142497	
Ctg8	77	848	707	143345	
Ctg159	76	844	704	144189	
Ctg16	61	825	688	145014	
Ctg150	62	810	675	145824	
Ctg118	70	782	652	146606	
Ctg20	67	757	631	147363	
Ctg61	70	756	630	148119	
Ctg12	74	754	629	148873	
Ctg18	62	740	617	149613	
Ctg101	59	729	608	150342	
Ctg21	57	729	608	151071	
Ctg11	44	694	579	151765	
Ctg112	62	682	569	152447	
Ctg25	71	669	558	153116	
Ctg9	58	661	551	153777	
Ctg131	48	648	540	154425	
Ctg15	58	637	531	155062	

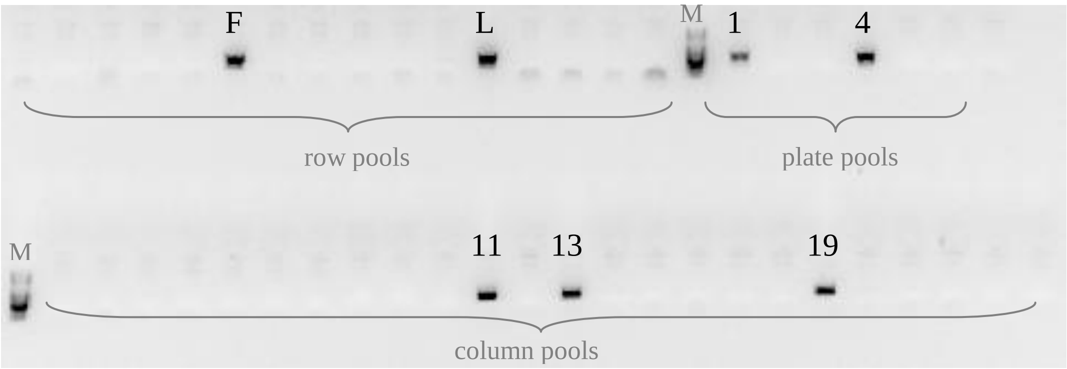
Ctg26	61	631	526	155693	
Ctg66	34	627	523	156320	
Ctg4	70	614	512	156934	
Ctg56	38	606	505	157540	
Ctg88	38	592	494	158132	
Ctg164	86	580	484	158712	
Ctg120	61	570	475	159282	
Ctg63	47	498	415	159780	
Ctg13	40	494	412	160274	
Ctg6	29	474	395	160748	
Ctg7	25	465	388	161213	
Ctg160	35	427	356	161640	
Ctg34	30	416	347	162056	
Ctg145	14	409	341	162465	
Ctg51	3	348	290	162813	
Ctg10	31	345	288	163158	
Ctg54	21	333	278	163491	
Ctg85	22	327	273	163818	
Ctg29	3	325	271	164143	
Ctg71	26	324	270	164467	
Ctg43	3	300	250	164767	
Ctg52	4	297	248	165064	
Ctg94	3	295	246	165359	
Ctg147	3	290	242	165649	
Ctg103	16	286	239	165935	
Ctg45	4	286	239	166221	
Ctg105	28	283	236	166504	
Ctg42	3	283	236	166787	
Ctg107	23	282	235	167069	
Ctg39	11	277	231	167346	
Ctg30	3	274	229	167620	
Ctg108	26	264	220	167884	
Ctg58	33	261	218	168145	
Ctg33	4	260	217	168405	
Ctg28	12	259	216	168664	
Ctg46	4	256	214	168920	
Ctg48	6	255	213	169175	
Ctg76	5	255	213	169430	
Ctg106	24	252	210	169682	
Ctg24	3	249	208	169931	
Ctg49	3	247	206	170178	
Ctg104	11	238	199	170416	
Ctg41	3	237	198	170653	
Ctg23	3	234	195	170887	
Ctg44	3	232	194	171119	

Ctg50	4	228	190	171347	
Ctg79	11	220	184	171567	
Ctg67	7	218	182	171785	
Ctg32	3	216	180	172001	
Ctg27	12	213	178	172214	
Ctg139	4	213	178	172427	
Ctg161	12	211	176	172638	
Ctg141	2	204	170	172842	
Ctg80	3	199	166	173041	
Ctg109	9	195	163	173236	
Ctg74	9	186	155	173422	
Ctg38	6	183	153	173605	
Ctg151	16	180	150	173785	
Ctg75	5	171	143	173956	
Ctg35	3	169	141	174125	
Ctg72	6	168	140	174293	
Ctg37	2	163	136	174456	
Ctg31	4	146	122	174602	
Ctg40	3	144	120	174746	
Ctg110	2	144	120	174890	
Ctg47	3	142	119	175032	
Ctg114	5	140	117	175172	
Ctg152	2	140	117	175312	
Ctg133	3	136	114	175448	
Ctg73	5	135	113	175583	
Ctg153	2	135	113	175718	
Ctg154	5	132	110	175850	
Ctg36	3	132	110	175982	
Ctg163	2	132	110	176114	
Ctg64	2	122	102	176236	
Ctg69	2	114	95	176350	
Ctg60	2	112	94	176462	
Ctg59	5	100	84	176562	
Ctg119	2	99	83	176661	
Ctg132	2	99	83	176760	
Ctg130	2	78	65	176838	

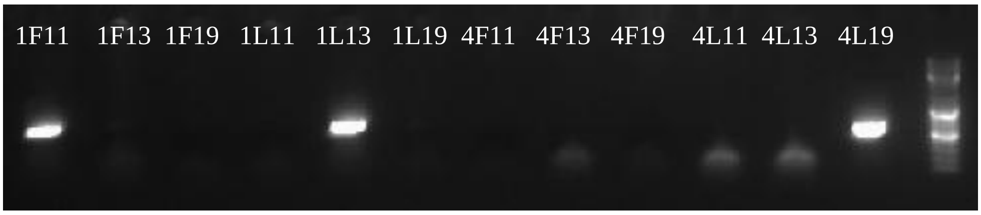
APPENDIX F

The 3-Dimensional pool strategy for MTP screening. MTP clones are combined into 16 row (A-P), 24 column (1-24) and 7 plate (1-7) pools. Positive signals from 3D pools indicate the coordinates (plate, row, column) of the original BAC clones. In the below example, COS marker GPI:C:758029 yielded 2 positive signals from row and plate pools, and 3 positive signals from the column pool, giving rise to 12 possibilities (2x2x3). In such cases, all possibilities were checked through colony PCR on original MTP clones. Colony PCR revealed (1F11), (1L13) and (4L19) as true positives. These coordinates corresponds to, TaeCsp5DShA_0036_K04, TaeCsp5DShA_0036_O01 and TaeCsp5DShA_0068_M22 clones, respectively. All PCR reactions were carried out at 56^{o}C.

Pool screening:



Colony PCR:



APPENDIX G

List of all contigs anchored by molecular markers through PCR.

No.	Marker name	Marker type	Contig
1	Pina-D1	Gene	CTG70
2	Pinb-D1	Gene	CTG70
3	CFD78	SSR	CTG159
4	BARC130	SSR	CTG146
5	CFD18	SSR	CTG66
6	CFD81	SSR	CTG93
7	WMC233	SSR	CTG66
8	WMS190	SSR	Bridge
9	WMS205	SSR	Bridge
10	WMC608	SSR	CTG99
11	CFD189	SSR	CTG113
12	CFD74	SSR	CTG144
13	GPW326	SSR	CTG70
14	WMS16	SSR	CTG99
15	WMS358	SSR	CTG157
16	GPI:C:726959	COS	CTG102
17	GPI:C:728036	COS	CTG149
18	GPI:C:728956	COS	CTG68
19	GPI:C:729592	COS	CTG156
20	GPI:C:739811	COS	CTG117
21	GPI:C:741009	COS	CTG124
22	GPI:C:743567	COS	CTG88
23	GPI:C:744654	COS	CTG124
24	GPI:C:746971	COS	CTG100
25	GPI:C:758029	COS	CTG89
26	GPI:C:758334	COS	CTG102
27	GPI:C:762599	COS	CTG102
28	A1Z95	ISBP	CTG144
29	A78ID	ISBP	CTG89
30	AFQ6M	ISBP	CTG2
31	AMEBF	ISBP	CTG129
32	B04N2	ISBP	CTG15
33	B2LTL	ISBP	CTG19
34	B5PZZ	ISBP	CTG14
35	B7QNM	ISBP	CTG95
36	BA0XF	ISBP	CTG98
37	BCLS8	ISBP	CTG93
38	BIC1N	ISBP	CTG159

39	BKSSP	ISBP	CTG99
40	BQDJS	ISBP	CTG115
41	BRAMT	ISBP	CTG26
42	BT6NA	ISBP	CTG58
43	BTCOT	ISBP	CTG96
44	BV62B	ISBP	CTG99
45	BUH5R	ISBP	CTG102
46	BXQWE	ISBP	CTG100
47	BYIFL	ISBP	CTG156
48	BZB42	ISBP	CTG138
49	C0NIF	ISBP	CTG131
50	EXX6A	ISBP	CTG156
51	E1EKW	ISBP	CTG129
52	D8U1H	ISBP	CTG102
53	CRSMO	ISBP	CTG145
54	CAJFV	ISBP	CTG68
55	CAGIE	ISBP	CTG138
56	C7J3U	ISBP	CTG78
57	C0VNS	ISBP	CTG125
58	BE403618	EST	CTG127
59	BE403785	EST	CTG70
60	BE404135	EST	CTG15
61	BE404490	EST	CTG2
62	BE405667	EST	CTG70
63	BE405839	EST	CTG53
64	BE422471	EST	CTG137
65	BE424775	EST	CTG137
66	BE591461	EST	CTG156
67	BE591734	EST	CTG99
68	BE591974	EST	CTG156
69	BE604729	EST	CTG135
70	BE606535	EST	CTG65
71	BE606637	EST	CTG146
72	BE606654	EST	CTG127
73	BE636795	EST	CTG144
74	BE444113	EST	CTG70
75	BE444644	EST	CTG138
76	BE585732	EST	CTG146
77	BE499257	EST	CTG120
78	BE471016	EST	CTG83
79	BE444720	EST	CTG102
80	BE497093	EST	CTG124
81	BE470750	EST	CTG78

82	BE490781	EST	CTG156
83	BE498768	EST	CTG125
84	BF202632	EST	CTG149
85	BF291319	EST	CTG88
86	BF473658	EST	CTG125
87	BF474606	EST	CTG26
88	BF474953	EST	CTG97
89	BF483719	EST	CTG115
90	BF484212	EST	CTG115
91	BF485220	EST	CTG88
92	BG262914	EST	CTG112
93	BG263391	EST	CTG26
94	BG604740	EST	CTG5
95	BG607041	EST	CTG70
96	BG607697	EST	CTG46
97	GH722882	EST	CTG96
98	CD882766	EST	CTG134
99	AX462334	EST	CTG159
100	BF485261	EST	CTG100
101	BE403373	EST	Bridge

APPENDIX H

List of all contigs anchored by molecular markers through microarray hybridization. Conserved: Conserved gene-associated sequences; COS: Conserved orthologous sequence; EST: Expressed sequence tag; ISBP: Insertion site-based polymorphism; SNP: Single nucleotide polymorphism.

Contig	# of matching clones	# of matching probes	Probe sources
CTG1	3	7	Conserved, ISBP
CTG100	14	36	Conserved, COS, EST, ISBP
CTG101	4	28	Conserved, ISBP
CTG102	22	114	Conserved, ISBP
CTG105	2	2	ISBP
CTG106	2	10	Conserved, ISBP
CTG11	3	7	Conserved, ISBP
CTG111	6	8	Conserved, ISBP
CTG112	7	14	Conserved, ISBP
CTG113	2	2	ISBP
CTG114	2	17	Conserved, ISBP
CTG115	25	94	Conserved, ISBP
CTG116	4	4	Conserved, ISBP
CTG117	6	54	Conserved, COS, ISBP
CTG12	7	22	Conserved, ISBP
CTG120	4	11	Conserved, ISBP
CTG121	4	6	ISBP
CTG122	7	24	Conserved, ISBP
CTG123	10	21	Conserved, ISBP
CTG124	17	77	Conserved, COS, ISBP
CTG125	22	119	Conserved, ISBP
CTG126	4	12	Conserved, ISBP
CTG127	8	16	Conserved, ISBP
CTG128	13	28	Conserved, ISBP
CTG129	4	5	Conserved, ISBP
CTG131	5	14	Conserved, ISBP
CTG134	9	44	Conserved, ISBP, SNP
CTG135	18	82	Conserved, ISBP
CTG136	6	24	Conserved, ISBP
CTG137	12	40	Conserved, ISBP, SNP
CTG138	33	170	Conserved, EST, ISBP
CTG14	9	49	Conserved, ISBP
CTG140	1	2	Conserved
CTG142	7	29	Conserved, EST, ISBP
CTG143	6	13	Conserved, ISBP
CTG144	14	37	Conserved, ISBP
CTG145	1	1	Conserved

CTG146	7	58	Conserved, EST, ISBP, SNP, SSR
CTG147	2	52	Conserved
CTG149	13	71	Conserved, ISBP
CTG15	5	40	Conserved, ISBP
CTG150	4	38	Conserved, EST, ISBP
CTG155	8	67	Conserved, ISBP
CTG156	26	163	Conserved, EST, ISBP
CTG157	11	50	Conserved, ISBP
CTG158	8	23	Conserved, EST, ISBP
CTG159	3	26	Conserved, ISBP, SSR
CTG17	7	11	Conserved
CTG18	4	15	Conserved, ISBP
CTG19	7	25	Conserved, ISBP
CTG2	4	4	EST, ISBP
CTG20	3	3	Conserved
CTG21	4	23	Conserved, ISBP
CTG25	4	16	Conserved
CTG26	4	16	Conserved, ISBP, SNP
CTG3	6	22	Conserved, ISBP
CTG34	4	10	Conserved, ISBP
CTG4	2	2	ISBP
CTG46	1	1	ISBP
CTG5	9	29	Conserved, ISBP
CTG53	8	25	Conserved, ISBP
CTG55	1	1	Conserved
CTG56	2	2	Conserved
CTG57	8	11	Conserved, ISBP
CTG6	6	10	Conserved, ISBP
CTG61	4	8	Conserved
CTG62	11	55	Conserved, EST, ISBP, SNP
CTG63	2	2	ISBP
CTG65	12	62	Conserved, ISBP, SNP
CTG66	5	12	Conserved, SSR
CTG68	20	59	Conserved, EST, ISBP
CTG7	3	10	Conserved, ISBP
CTG70	12	69	Conserved, EST, ISBP, SNP, SSR
CTG71	1	1	ISBP
CTG74	3	13	Conserved, ISBP
CTG77	5	9	Conserved, ISBP
CTG78	29	98	Conserved, ISBP
CTG79	1	1	ISBP
CTG81	3	6	Conserved, ISBP
CTG82	4	17	Conserved, ISBP
CTG83	8	34	Conserved, ISBP
CTG84	11	47	Conserved, ISBP, SNP
CTG85	4	5	Conserved, ISBP

CTG86	4	4	ISBP
CTG87	15	70	Conserved, ISBP
CTG89	8	24	Conserved, ISBP
CTG90	2	4	Conserved
CTG91	6	34	Conserved, ISBP
CTG93	28	126	Conserved, ISBP, SNP, SSR
CTG94	1	1	Conserved
CTG95	5	12	Conserved, ISBP
CTG96	10	43	Conserved, ISBP
CTG97	14	45	Conserved, ISBP, SNP
CTG98	2	2	Conserved
CTG99	25	77	Conserved, ISBP, SNP

APPENDIX I

Final version of the 5DS physical map. Markers in black are physically anchored; markers in blue are putatively assigned via microarray; markers in green are anchored by both approaches. Contigs highlighted in green indicate the presence of an orthologous *Ae. tauschii* sequence.

Deletion bin		Anchored marker/probe	Contig
0.78-1.00		C0NIF, HJKAX1S01A3G1S, HJKAX1S01A8XG6_1, HJKAX1S01AFTIY, HJKAX1S01AICQU_1, HJKAX1S01B05PQ, HJKAX1S01BRYKN, HJKAX1S01C0NIF_1, HJKAX1S01DEGP7, HJKAX1S01EHUX6	[CTG131-CTG151]
0.78-1.00		Pina-D1, Pinb-D1, BE403785, BE405667, BG607041, BE444113, BE637485, BF293305, gpw326, HJKAX1S01A70V1,HJKAX1S01A7347_1, HJKAX1S01A8B3O_1, HJKAX1S01A91G5, HJKAX1S01AF8UU, HJKAX1S01AG15D_1, HJKAX1S01AH62T, HJKAX1S01AI42B, HJKAX1S01AO2IM_1, HJKAX1S01ATT3Z_1, HJKAX1S01AWF2H_1, HJKAX1S01AXL6T, HJKAX1S01B6SGQ, HJKAX1S01B7OOQ_1, HJKAX1S01BG97Y_1, HJKAX1S01BHDPO, HJKAX1S01BKUGM, HJKAX1S01BVWHU, HJKAX1S01BXMEO, HJKAX1S01C4ZI8, HJKAX1S01CD7FZ, HJKAX1S01CGJMM_2, HJKAX1S01CW75S_1, HJKAX1S01D23GR_3, HJKAX1S01D6HM2, HJKAX1S01DAQ12, HJKAX1S01DDBY8, HJKAX1S01DMNTP, HJKAX1S01DR64Q, HJKAX1S01DRQWU, HJKAX1S01DYQMY, HJKAX1S01E0BTK, HJKAX1S01E3VOE, HJKAX1S01EKNNU, HJKAX1S01EX52F, synopGBS108640, synopGBS110356_7_8_A	[CTG70
0.78-1.00		CRSMO, HJKAX1S01B90WE	CTG145
0.78-1.00		BE606637, BARC130, BE585732, BE636954, HJKAX1S01A3NTM, HJKAX1S01A727K_1, HJKAX1S01AIFSZ, HJKAX1S01AP5C7, HJKAX1S01AREGM, HJKAX1S01AWAZM, HJKAX1S01B6LRX_1, HJKAX1S01B8HN9, HJKAX1S01B8PPZ, HJKAX1S01BERCR_1, HJKAX1S01BPL1Q, HJKAX1S01BQ6NP, HJKAX1S01BR82E, HJKAX1S01C54TY, HJKAX1S01C5H7F_2, HJKAX1S01C868X_1, HJKAX1S01CFT8W, HJKAX1S01CVP7I, HJKAX1S01D1IAU_1, HJKAX1S01D56LR, HJKAX1S01DHW82, HJKAX1S01DOMI4, HJKAX1S01DXYG2, HJKAX1S01DY2CB, HJKAX1S01E3U0J, HJKAX1S01EBAQH, HJKAX1S01EDGY7, HJKAX1S01EJV5W, HJKAX1S01EQ080, HJKAX1S01ER6UG, HJKAX1S01EXJN7, synopGBS102655_B	CTG146]
0.78-1.00		CFD18, WMC233, cfd165, HJKAX1S01A7MLJ, HJKAX1S01CM6WO, HJKAX1S01CPX97, HJKAX1S01D5GB5, HJKAX1S01DMXJ8, HJKAX1S01EAEME, HJKAX1S01EECDM	[CTG66-CTG122]
0.78-1.00		HJKAX1S01A8DE2, HJKAX1S01AK1FS_2, HJKAX1S01AK8JS, HJKAX1S01AMZPY_2, HJKAX1S01AZIKV_1, HJKAX1S01B2F31, HJKAX1S01B2QMF, HJKAX1S01B63H0, HJKAX1S01B74GC, HJKAX1S01BC19Q_1, HJKAX1S01BGEN2, HJKAX1S01BQS5H, HJKAX1S01BTO6R, HJKAX1S01BVKLW, HJKAX1S01BZLKT, HJKAX1S01C0CL3_1, HJKAX1S01C62IA, HJKAX1S01CKCGK, HJKAX1S01CTC87, HJKAX1S01D0JR9_1, HJKAX1S01DL5WK, HJKAX1S01DMUZH, HJKAX1S01DSRPF, HJKAX1S01DYHJ2, HJKAX1S01ESQAC, HJKAX1S01EVML2, synopGBS112823, synopGBS124013	CTG84
0.78-1.00		HJKAX1S01A0PW6, HJKAX1S01AH52G_2, HJKAX1S01AJSEB, HJKAX1S01APUII, HJKAX1S01AZP0H, HJKAX1S01B1ERL_1, HJKAX1S01B4MEI, HJKAX1S01BHESQ, HJKAX1S01BL3SB, HJKAX1S01CMBCB, HJKAX1S01CZ1TS, HJKAX1S01EVS1K	CTG18
0.78-1.00		BE606535, HJKAX1S01A44MH_1, HJKAX1S01A8XA8_1, HJKAX1S01AF8F9_1, HJKAX1S01AG9B5, HJKAX1S01AINB7_2, HJKAX1S01AIUC6, HJKAX1S01AO571, HJKAX1S01AR9ZC, HJKAX1S01AXP70_1, HJKAX1S01B27FE, HJKAX1S01BBAIZ_1, HJKAX1S01BGWSX, HJKAX1S01BI0P3, HJKAX1S01BOD1G_1, HJKAX1S01BVEYG, HJKAX1S01BXDC4, HJKAX1S01C2HR3_1, HJKAX1S01C9GI2_1, HJKAX1S01CBUEU, HJKAX1S01CP1IP, HJKAX1S01CVZFQ, HJKAX1S01CXXIW, HJKAX1S01CYJJA_1, HJKAX1S01DA1RU_1, HJKAX1S01DE6WM, HJKAX1S01DEBBX, HJKAX1S01DIWPI, HJKAX1S01DSST4, HJKAX1S01E0P00, HJKAX1S01EI56T, synopGBS108676, synopGBS114788, synopGBS119432	CTG65
0.78-1.00		CD882766, HJKAX1S01A8B8T, HJKAX1S01A8XC0_1, HJKAX1S01AF0UY_1, HJKAX1S01AIG75_1, HJKAX1S01AMP1U, HJKAX1S01ARY5F_1, HJKAX1S01BBTKY, HJKAX1S01BIB29, HJKAX1S01BM04B_1, HJKAX1S01BUU8Q, HJKAX1S01BWCKO, HJKAX1S01CBM6D, HJKAX1S01CCYOO, HJKAX1S01CGBGZ, HJKAX1S01CGDI4, HJKAX1S01CIRZG, HJKAX1S01CZ7BB, HJKAX1S01DC0GZ, HJKAX1S01DCDZJ, HJKAX1S01DDZ7Y, HJKAX1S01DW9ZP, HJKAX1S01E0CLV, synopGBS118957	CTG134
0.78-1.00		HJKAX1S01A2WM0, HJKAX1S01A3M7E, HJKAX1S01AQDW6_1, HJKAX1S01ASO8W_1, HJKAX1S01AUUUP_1, HJKAX1S01B2FXF, HJKAX1S01D549R	[CTG57-CTG162]
0.78-1.00		BF292091, HJKAX1S01A4D9I, HJKAX1S01AD6GX, HJKAX1S01AF8WE_1, HJKAX1S01ATJOE, HJKAX1S01ATL6K_1, HJKAX1S01AV9J8_1, HJKAX1S01AXA1K, HJKAX1S01B0FUF, HJKAX1S01B1PLD, HJKAX1S01B5EBV, HJKAX1S01BDTPL_1, HJKAX1S01BIIUA_1, HJKAX1S01BMQEZ, HJKAX1S01BY8WA_1, HJKAX1S01BZCMK_1, HJKAX1S01CE3OL, HJKAX1S01CEQNK_1, HJKAX1S01CHEQ9_1, HJKAX1S01CMW98_1, HJKAX1S01COE5P, HJKAX1S01CX4AZ_1, HJKAX1S01D32OU, HJKAX1S01DAWNA, HJKAX1S01DG5DP, HJKAX1S01DG6FQ, HJKAX1S01DJB7O, HJKAX1S01DOAKR, HJKAX1S01DW3ZR, synopGBS105314_15_16_A	CTG62
0.78-1.00		HJKAX1S01B2HMT, HJKAX1S01CRQII, HJKAX1S01CTHT5, HJKAX1S01CZILL, HJKAX1S01D1GT8, HJKAX1S01D3VJB, HJKAX1S01D8BED	CTG17
0.78-1.00	mapped to 0.78-1.00 but order is uncertain	B04N2, BE404135, HJKAX1S01A4F6S, HJKAX1S01A62CD_1, HJKAX1S01AJJTQ, HJKAX1S01ART7U, HJKAX1S01AVTBI, HJKAX1S01AXOWN, HJKAX1S01BCQAY, HJKAX1S01BQ7XQ_1, HJKAX1S01BUP39, HJKAX1S01C8PDJ_1, HJKAX1S01CBZ94_1, HJKAX1S01CGBAE, HJKAX1S01CJAA7, HJKAX1S01CKGF6_1, HJKAX1S01CL8CQ, HJKAX1S01CUHYM, HJKAX1S01CV459_1, HJKAX1S01D5EDW, HJKAX1S01DALSX, HJKAX1S01DD922, HJKAX1S01DIQXD, HJKAX1S01DIX1W, HJKAX1S01DJUUD, HJKAX1S01DL8H3, HJKAX1S01E050A, HJKAX1S01ET89S	CTG15
0.78-1.00		B2LTL, HJKAX1S01AOIHY_1, HJKAX1S01AUXDJ, HJKAX1S01AVMN5_1, HJKAX1S01B20WK, HJKAX1S01B8UBF, HJKAX1S01B9FXW_1, HJKAX1S01BBF8R_1, HJKAX1S01BPS6P_1, HJKAX1S01CUA9Q, HJKAX1S01DJEM9, HJKAX1S01EL57B, HJKAX1S01ELPFZ, HJKAX1S01ENR0Q, HJKAX1S01EW3TX	CTG19
0.78-1.00		BF474606, BG263391, BRAMT, HJKAX1S01A9C1T_2, HJKAX1S01AUOGV, HJKAX1S01AZ9E5, HJKAX1S01B6N64, HJKAX1S01BR51Q, HJKAX1S01BY6YM, HJKAX1S01EPNYP, synopGBS129672_A	CTG26
0.78-1.00		AMEBF, E1EKW, HJKAX1S01B29MO_1, HJKAX1S01B2TE7, HJKAX1S01D55YG, HJKAX1S01EORLM	CTG129
0.78-1.00		BG262914, HJKAX1S01C23SU_1, HJKAX1S01D5TB8_1, HJKAX1S01D8U82, HJKAX1S01DZW2V, HJKAX1S01E2M1Z, HJKAX1S01EW1Y8	[CTG112-CTG111]
0.78-1.00		BM137384, HJKAX1S01BGKTI, HJKAX1S01BUQEX_1, HJKAX1S01CDPOC_2, HJKAX1S01CE09M, HJKAX1S01CJZR8, HJKAX1S01CPF3S, HJKAX1S01CRAQQ, HJKAX1S01E1JGI, HJKAX1S01EAMZN, HJKAX1S01EX3XK	[CTG158-CTG118]

0.67-0.78		BV62B, BKSSP, WMC608, WMS16, BE591734, HJKAX1S01A1AL5_1, HJKAX1S01A8M7O_1, HJKAX1S01A8WLN_1, HJKAX1S01A9MKQ_1, HJKAX1S01AIE76, HJKAX1S01AMR6I_1, HJKAX1S01ANEAI_1, HJKAX1S01APUXE, HJKAX1S01ASKML_1, HJKAX1S01B0CI9, HJKAX1S01B0DFD_1, HJKAX1S01B3SEH_1, HJKAX1S01B8MNI_1, HJKAX1S01BAV84, HJKAX1S01BCHF3_2 HJKAX1S01BJN2Y, HJKAX1S01BMC7J, HJKAX1S01BROWL, HJKAX1S01BTC15, HJKAX1S01BYN71_1, HJKAX1S01C7C36, HJKAX1S01CA4L9, HJKAX1S01CFPH3_1, HJKAX1S01CQT5O, HJKAX1S01CRQP3, HJKAX1S01CWXCN_1, HJKAX1S01D0S5U, HJKAX1S01D20J0, HJKAX1S01D372T_1, HJKAX1S01DFKLE, HJKAX1S01DJMPS, HJKAX1S01DK80M, HJKAX1S01DQ3DA, HJKAX1S01DQ497, HJKAX1S01DQWG0, HJKAX1S01DXGDB, HJKAX1S01DZV05, HJKAX1S01E3TZA, HJKAX1S01EFSPO, HJKAX1S01EW0MA, HJKAX1S01EWJWA, synopGBS101007_08_B, synopGBS101137	CTG99
0.67-0.78		BCLS8, CFD81, HJKAX1S01A19BM, HJKAX1S01A2R5Z, HJKAX1S01A37N1_1, HJKAX1S01A92WI_1, HJKAX1S01AE2X2, HJKAX1S01AF1XS, HJKAX1S01AF6LJ, HJKAX1S01AFTY0_1, HJKAX1S01AGTJU_1, HJKAX1S01AIVBR, HJKAX1S01AP5M4_1, HJKAX1S01APRPE_2, HJKAX1S01ASVPD, HJKAX1S01AT1GT, HJKAX1S01AXE9R_1, HJKAX1S01AY22M, HJKAX1S01AYADL, HJKAX1S01B0GUX_1, HJKAX1S01B0IJA_1, HJKAX1S01B1GIU, HJKAX1S01B1GOW_3, HJKAX1S01B2VNV_1, HJKAX1S01B6RXS, HJKAX1S01B777M_1, HJKAX1S01B7Q0X, HJKAX1S01BHO2A, HJKAX1S01BHO2A_2, HJKAX1S01BKH6A_1, HJKAX1S01BL8EE_1, HJKAX1S01BLL46, HJKAX1S01BLR8D, HJKAX1S01BM6YA, HJKAX1S01BMSG6, HJKAX1S01BU1IA, HJKAX1S01BW7L9, HJKAX1S01C203M_1, HJKAX1S01C6QO0, HJKAX1S01C7RIR_1, HJKAX1S01CAJOL_1, HJKAX1S01CBIA1, HJKAX1S01CBN1C, HJKAX1S01CHTNQ, HJKAX1S01CIO1T_2, HJKAX1S01CIWDN, HJKAX1S01CK3YH, HJKAX1S01CKE67, HJKAX1S01CP56D, HJKAX1S01CQNPH, HJKAX1S01CXFXT, HJKAX1S01CYVJ0, HJKAX1S01D3OK5_1, HJKAX1S01D5SB7, HJKAX1S01D662F, HJKAX1S01D8EDE, HJKAX1S01D8LYB, HJKAX1S01DA8NJ, HJKAX1S01DDRSP, HJKAX1S01DE1IL, HJKAX1S01DJ5JX, HJKAX1S01DUISU, HJKAX1S01DVCQG, HJKAX1S01DVNGU, HJKAX1S01DX51P, HJKAX1S01EDTQL, HJKAX1S01EIO8D, HJKAX1S01EJ7UI, HJKAX1S01EMLNR, HJKAX1S01ES7F6, HJKAX1S01ESQ3H, HJKAX1S01EVT17, HJKAX1S01EXCVX, synopGBS122594	CTG93
0.63-0.67		HJKAX1S01A0AAJ_1, HJKAX1S01A3F5V_2, HJKAX1S01A6MNI, HJKAX1S01AGS4J, HJKAX1S01AJKII_1, HJKAX1S01AKBI9, HJKAX1S01AQYBU, HJKAX1S01ASI9T, HJKAX1S01B1890_1, HJKAX1S01BC8XL, HJKAX1S01BF4XX_1, HJKAX1S01BN4CY, HJKAX1S01BP9ZD, HJKAX1S01BREO3, HJKAX1S01BSDN3, HJKAX1S01C5OGS, HJKAX1S01C7M9W, HJKAX1S01CDL76_1, HJKAX1S01CH2FX, HJKAX1S01CH3EZ, HJKAX1S01CHNE5_2, HJKAX1S01CISK2_1, HJKAX1S01CX21Y, HJKAX1S01CXDQ1, HJKAX1S01CXDQ1_2, HJKAX1S01D0G7W, HJKAX1S01D1JN4, HJKAX1S01D1RRN, HJKAX1S01D4TLP, HJKAX1S01D83O4, HJKAX1S01DAHRI_2, HJKAX1S01DAMQJ_1, HJKAX1S01DO78I, HJKAX1S01E2M6P, HJKAX1S01ECL74, HJKAX1S01EPOUD, HJKAX1S01EQTN5, HJKAX1S01EWUBH, HJKAX1S01EWVZR	CTG155
0.63-0.67		GPI:C:739811, HJKAX1S01A75RI, HJKAX1S01A81A2, HJKAX1S01AGRHC_1, HJKAX1S01ANBIF_1, HJKAX1S01AR8LO_1, HJKAX1S01ASP8I_2, HJKAX1S01ATD00, HJKAX1S01AYV7P, HJKAX1S01B02EY, HJKAX1S01B6N8I, HJKAX1S01BGYC8, HJKAX1S01BW5VS, HJKAX1S01BWXLW, HJKAX1S01BYL7P_1, HJKAX1S01CG6JT_1, HJKAX1S01CQ90A, HJKAX1S01CZBT4_2, HJKAX1S01D0963, HJKAX1S01D3Z3G_1, HJKAX1S01DDGKK, HJKAX1S01DEAKG, HJKAX1S01DNGW4, HJKAX1S01DPOPU, HJKAX1S01DT6F5, HJKAX1S01EJ63H, HJKAX1S01EK99Y, HJKAX1S01EMILZ, HJKAX1S01EPOQF, HJKAX1S01ERUB5	CTG117
0.63-0.67		B5PZZ, HJKAX1S01A1UYT, HJKAX1S01A8F6O, HJKAX1S01AHVHY, HJKAX1S01AJGR2, HJKAX1S01ALBHW_1, HJKAX1S01ALZE3_1, HJKAX1S01ATHT9, HJKAX1S01AVYJ4, HJKAX1S01B888M, HJKAX1S01BQJ9E_1, HJKAX1S01BTBB2, HJKAX1S01CNGXM, HJKAX1S01CUOZ8_1, HJKAX1S01CZY4M_2, HJKAX1S01D0TCI, HJKAX1S01D3179, HJKAX1S01D4VZK, HJKAX1S01D8ZWT, HJKAX1S01DGA0V, HJKAX1S01DJER1, HJKAX1S01EC7QM, HJKAX1S01EE12I, HJKAX1S01EL42U	CTG14
0.63-0.67		BQDJS, BF483719, BF484212, HJKAX1S01A18X0, HJKAX1S01A5BNJ_1, HJKAX1S01A8XU6, HJKAX1S01A9U9C, HJKAX1S01AD4FT_1, HJKAX1S01AD803, HJKAX1S01AF6Y7, HJKAX1S01AK035, HJKAX1S01AK4K0_1, HJKAX1S01AMNLR_1, HJKAX1S01AMP3Y, HJKAX1S01AUQQR, HJKAX1S01AZ757, HJKAX1S01B2TI9, HJKAX1S01B60A8_1, HJKAX1S01B77YA, HJKAX1S01B7JL7, HJKAX1S01B8QWS, HJKAX1S01BD7KT_2, HJKAX1S01BFM22_1, HJKAX1S01BIHWS, HJKAX1S01BJKKZ_1, HJKAX1S01BOPNV, HJKAX1S01BP7HO, HJKAX1S01BPBCP, HJKAX1S01C0XYL_1, HJKAX1S01C438M_1, HJKAX1S01C4TO7_1, HJKAX1S01C7O25, HJKAX1S01C87UH_1, HJKAX1S01CA2KW_1, HJKAX1S01CIT6V, HJKAX1S01CMUQ9, HJKAX1S01CZMSZ, HJKAX1S01D0J8E, HJKAX1S01D35V3_2, HJKAX1S01D72PG, HJKAX1S01D86UD, HJKAX1S01D8NSB, HJKAX1S01DLUKQ, HJKAX1S01DN5P2, HJKAX1S01DRFO0, HJKAX1S01DRFUU, HJKAX1S01DSW0E, HJKAX1S01DWI64, HJKAX1S01DY77U, HJKAX1S01DZ7XI, HJKAX1S01EAL70, HJKAX1S01ED1BY, HJKAX1S01EM0M8, HJKAX1S01EO56A, HJKAX1S01EO8II, HJKAX1S01EVGYU, HJKAX1S01EY011	CTG115
0.63-0.67		HJKAX1S01AU0PR_1, HJKAX1S01BTQ49, HJKAX1S01BV159, HJKAX1S01DOSQS, HJKAX1S01DYLQY	CTG81
0.63-0.67		BUH5R, D8U1H, GPI:C:758334, BE444720, HJKAX1S01A146U, HJKAX1S01A4T0V_1, HJKAX1S01A5I2W, HJKAX1S01A8VE4, HJKAX1S01A9D7T, HJKAX1S01A9QAF_1, HJKAX1S01AG8SM, HJKAX1S01AI7SW, HJKAX1S01AOUM4_1, HJKAX1S01APRNZ, HJKAX1S01AR1KC, HJKAX1S01AU5EA_1, HJKAX1S01AX2GY_1, HJKAX1S01B0LHG, HJKAX1S01B3DV7, HJKAX1S01B3IHJ_1, HJKAX1S01B5D48_1, HJKAX1S01B8BXX, HJKAX1S01B9F53_1, HJKAX1S01BEUXL, HJKAX1S01BLYQ6_1, HJKAX1S01BM7M4, HJKAX1S01BSMGH, HJKAX1S01BUEXD, HJKAX1S01BUH5R_1, HJKAX1S01BZ75R, HJKAX1S01C3EM0, HJKAX1S01C3WNF, HJKAX1S01C7QV4, HJKAX1S01C8UJZ, HJKAX1S01CB8Q9, HJKAX1S01CFT55, HJKAX1S01CH51F, HJKAX1S01CIKO2, HJKAX1S01CNCNC_2, HJKAX1S01CO6XK, HJKAX1S01CUQB7, HJKAX1S01CV3BC, HJKAX1S01CV9WG, HJKAX1S01CVF9K, HJKAX1S01CZ3ON, HJKAX1S01CZAUJ, HJKAX1S01D0QH7, HJKAX1S01D0RJK_1, HJKAX1S01D3P97, HJKAX1S01DALDV, HJKAX1S01DBYEP, HJKAX1S01DJP4A, HJKAX1S01DO9LI, HJKAX1S01DRN46, HJKAX1S01DTIA7, HJKAX1S01DUZ0Q, HJKAX1S01DY81J, HJKAX1S01E2UZ7, HJKAX1S01E41SX, HJKAX1S01EA00F, HJKAX1S01EB384, HJKAX1S01EESIY, HJKAX1S01EGKWX, JKAX1S01EKPZ0, HJKAX1S01EL17F, HJKAX1S01EP0WT, HJKAX1S01EVDIP, HJKAX1S01EVXOU	CTG102
0.63-0.67 junction		C7J3U, BE470750, HJKAX1S01A7M7H_1, HJKAX1S01A8G7B, HJKAX1S01A9K39, HJKAX1S01AG6N7_2, HJKAX1S01AGBML, HJKAX1S01AKHI3, HJKAX1S01AL5J6, HJKAX1S01AM6QR_2, HJKAX1S01AO06D_1, HJKAX1S01APOQR, HJKAX1S01AQC78_1, HJKAX1S01ATC5F_2, HJKAX1S01AU5YH, HJKAX1S01AVFQ3, HJKAX1S01AZCW3_1, HJKAX1S01B07OU, HJKAX1S01B0UPA, HJKAX1S01B3V8V, HJKAX1S01B6J87_1, HJKAX1S01BANIX_1, HJKAX1S01BAOBK, HJKAX1S01BDUC1_1, HJKAX1S01BDY2R, HJKAX1S01BIU13, HJKAX1S01BM4MV, HJKAX1S01BM7ST, HJKAX1S01BO3FA, HJKAX1S01BOI0V, HJKAX1S01BRN76, HJKAX1S01BWWV8, HJKAX1S01BXTLX_1, HJKAX1S01BY2X8, HJKAX1S01BYC6R_1, HJKAX1S01BZ8RN, HJKAX1S01BZJ3R_1, HJKAX1S01C0X0F_1, HJKAX1S01C1PTR, HJKAX1S01C21X5, HJKAX1S01C5B07, HJKAX1S01CCR23_1, HJKAX1S01CD5Z0_1, HJKAX1S01CFIOF, HJKAX1S01CGP6U, HJKAX1S01CH1BV, HJKAX1S01CL7Y8, HJKAX1S01CO5MD, HJKAX1S01COAY5, HJKAX1S01CPHNH_1, HJKAX1S01CVEWZ, HJKAX1S01CWU40, HJKAX1S01CWY9F, HJKAX1S01CY6Q4, HJKAX1S01D0AYT, HJKAX1S01D31GD, HJKAX1S01D4C9M, HJKAX1S01DAMIK, HJKAX1S01DSF7K, HJKAX1S01DTOJH, HJKAX1S01DU2Z8, HJKAX1S01DX2HJ, HJKAX1S01DZCSQ, HJKAX1S01E6I84, HJKAX1S01EO9P9, HJKAX1S01EX9U6	CTG78
0.63-0.67 junction		BA0XF, CAJFV, GPI:C:728956, HJKAX1S01DT5BD, BF292081, HJKAX1S01A1URF_1, HJKAX1S01A6VLE, HJKAX1S01AFX6F, HJKAX1S01AGCXF, HJKAX1S01AP8OI, HJKAX1S01AS0EZ, HJKAX1S01ATG0H_1, HJKAX1S01AW18C, HJKAX1S01B6W94, HJKAX1S01B6W94_1, HJKAX1S01B8GT3_1, HJKAX1S01BADPF, HJKAX1S01BDDYD, HJKAX1S01BR0N0_1, HJKAX1S01BVREN, HJKAX1S01C0R2D, HJKAX1S01C9CYQ_1, HJKAX1S01CDOE1, HJKAX1S01CEMIB, HJKAX1S01CENIX, HJKAX1S01CGKB8_1, HJKAX1S01CJ5KW_1, HJKAX1S01CMXJS, HJKAX1S01CNIFD, HJKAX1S01CU0X5, HJKAX1S01CWPXK_1, HJKAX1S01D0Z6S, HJKAX1S01D8AYP, HJKAX1S01D8XPL, HJKAX1S01DJUT1, HJKAX1S01DMJG4, HJKAX1S01DUCQC, HJKAX1S01EK889	[CTG98-CTG54-CTG68]
0.63-0.67	mapped to	A78ID, GPI:C:758029, HJKAX1S01ANCRF, HJKAX1S01AO5ZI, HJKAX1S01B2617_1, HJKAX1S01B5FIS, HJKAX1S01B8JO3, HJKAX1S01B98BO, HJKAX1S01BC3MN, HJKAX1S01BN6FG, HJKAX1S01C37RG, HJKAX1S01CX857, HJKAX1S01DU33U, HJKAX1S01ED9TT	CTG89

0.63-0.67	0.63-0.67 but order is uncertain	BF474953, HJKAX1S01A7UQV_1, HJKAX1S01A8TZB_2, HJKAX1S01ANOY5, HJKAX1S01AVZGE, HJKAX1S01B5VHT, HJKAX1S01BSGSF_2, HJKAX1S01C00XN_1, HJKAX1S01C0IAL, HJKAX1S01C60HW, HJKAX1S01CPOQ8, HJKAX1S01CSAS3, HJKAX1S01CWIDP, HJKAX1S01CYW8J_1, HJKAX1S01DELG0, HJKAX1S01DFVTG, HJKAX1S01DH3HO, HJKAX1S01DHOT1, HJKAX1S01DQJV3, HJKAX1S01EBBPF, HJKAX1S01EKSIV, HJKAX1S01EMS5L, HJKAX1S01ESWLW, synopGBS119454_5_6_B, synopGBS126656	CTG97
0.63-0.67		BG607697, HJKAX1S01B42H9_2	CTG46
0-0.63		BE499257, HJKAX1S01A34JE_1, HJKAX1S01A3ZKL, HJKAX1S01ADS67, HJKAX1S01BJFRB, HJKAX1S01BKVL3, HJKAX1S01BM5S6, HJKAX1S01COGRB, HJKAX1S01D4X3K, HJKAX1S01DR6B1, HJKAX1S01DSKKG	[CTG143-CTG120-CTG82]
0-0.63		GPI:C:728036, BF202632, HJKAX1S01A4YRJ_1, HJKAX1S01A9K8O, HJKAX1S01AECOA_1, HJKAX1S01AFZC8, HJKAX1S01AHSIJ_1, HJKAX1S01AMYSS, HJKAX1S01ATVBU, HJKAX1S01AUKGD, HJKAX1S01AVSPC, HJKAX1S01B3APK, HJKAX1S01BERCW, HJKAX1S01BGKQ5, HJKAX1S01BJD33_1, HJKAX1S01BJSNF, HJKAX1S01BMBPQ_1, HJKAX1S01BMJ5R, HJKAX1S01C02HV, HJKAX1S01C0XVX, HJKAX1S01C2GNE_1, HJKAX1S01C5LY4, HJKAX1S01C6F4R, HJKAX1S01CQ23G, HJKAX1S01D0404, HJKAX1S01DAV60, HJKAX1S01DC73P, HJKAX1S01DIBLR, HJKAX1S01DN1TQ, HJKAX1S01DOH9J, HJKAX1S01DRJUA, HJKAX1S01DUK7C, HJKAX1S01E565L, HJKAX1S01EFQMG, HJKAX1S01ENW61, HJKAX1S01EQ3M7, HJKAX1S01EXT6F	[CTG149-CTG140]
0-0.63		HJKAX1S01A0IWX, HJKAX1S01AR1WB, HJKAX1S01AXG90, HJKAX1S01B015D, HJKAX1S01BFOKH, HJKAX1S01BPOV5, HJKAX1S01DJSTL, HJKAX1S01EY4R1	CTG25
0-0.63		BE497093, GPI:C:741009, GPI:C:744654, HJKAX1S01A180B_2, HJKAX1S01A6TPM, HJKAX1S01AGYZ1, HJKAX1S01AH8NN, HJKAX1S01AHZBG, HJKAX1S01AIDXK_1, HJKAX1S01ALFCG, HJKAX1S01AUG8G, HJKAX1S01B1W96, HJKAX1S01B5DOH, HJKAX1S01BBM1D, HJKAX1S01BG0MO, HJKAX1S01BIXSB, HJKAX1S01BP5F2_1, HJKAX1S01BUS5I, HJKAX1S01BUY5G, HJKAX1S01C2BHH, HJKAX1S01C2QMJ_1, HJKAX1S01C4F9I, HJKAX1S01CQB5D_1, HJKAX1S01CQR4D, HJKAX1S01CWDDE, HJKAX1S01CYVKC, HJKAX1S01CZEPV_1, HJKAX1S01D4LGC, HJKAX1S01D67I2, HJKAX1S01D6PG8, HJKAX1S01D9DXQ_2, HJKAX1S01DAR2Q, HJKAX1S01DARR7, HJKAX1S01DAWY4, HJKAX1S01DD40V, HJKAX1S01DDM74, HJKAX1S01DLZKE, HJKAX1S01DMKUI, HJKAX1S01DOYS9, HJKAX1S01DS8ZI, HJKAX1S01DYU1G, HJKAX1S01DZTRS, HJKAX1S01E2JSA, HJKAX1S01EA0JP, HJKAX1S01ETGUO, HJKAX1S01EZ5AJ	CTG124
0-0.63		A1Z95, CFD74, BE636795, HJKAX1S01A1BOC, HJKAX1S01AFQCP, HJKAX1S01AOWPC_1, HJKAX1S01AQQVV, HJKAX1S01ASP54, HJKAX1S01B1I4F_1, HJKAX1S01B1LFP, HJKAX1S01BBDK2_1, HJKAX1S01BI22I_2, HJKAX1S01BPNQZ, HJKAX1S01BRGQX, HJKAX1S01CFFRJ, HJKAX1S01CHQN7_1, HJKAX1S01D312L, HJKAX1S01D4904, HJKAX1S01D83H6_1, HJKAX1S01DGWXW, HJKAX1S01DPOGU, HJKAX1S01DV5W4, HJKAX1S01ECZ5S, HJKAX1S01EEZWJ, HJKAX1S01EWB7B, HJKAX1S01EXYES	[CTG144-CTG121]
0-0.63		BXQWE, BF485261, GPI:C:746971, BE443751, HJKAX1S01A62C5_1, HJKAX1S01A7A8Q_1, HJKAX1S01A7J6H_1, HJKAX1S01A8VUC_1, HJKAX1S01AI0RH_2, HJKAX1S01ASXPP, HJKAX1S01AYD4U, HJKAX1S01B5OO8, HJKAX1S01BBVN8_1, HJKAX1S01BGFWN_1, HJKAX1S01BRADD, HJKAX1S01C9P4Z, HJKAX1S01CFQ7M, HJKAX1S01CLRU8, HJKAX1S01CUN29, HJKAX1S01DU8K9, HJKAX1S01ET2LL, HJKAX1S01EVK0B, HJKAX1S01EWPHW	[CTG100-CTG71]
0-0.63		HJKAX1S01AU2R5, HJKAX1S01B0Y2T, HJKAX1S01B3WFU_1, HJKAX1S01B8UOO, HJKAX1S01BHDZ0, HJKAX1S01BQ6DT, HJKAX1S01BX79U, HJKAX1S01BZHBP, HJKAX1S01C3XOH, HJKAX1S01C60VA, HJKAX1S01D5VR9, HJKAX1S01DBDY5, HJKAX1S01DHU8D, HJKAX1S01ENBNW	CTG101
0-0.63		BIC1N, AX462334, CFD78, HJKAX1S01AHTYI, HJKAX1S01B21Q2, HJKAX1S01B72B6_1, HJKAX1S01BAHDB, HJKAX1S01BBYNW, HJKAX1S01BDBB3, HJKAX1S01BDXF0, HJKAX1S01BGRQQ, HJKAX1S01BZMZ3_1, HJKAX1S01C0EH6_2, HJKAX1S01C94G9_3, HJKAX1S01CWJ6K_1, HJKAX1S01D02T9, HJKAX1S01DFYJO, HJKAX1S01DXC8A, HJKAX1S01EALL9, HJKAX1S01EIMS3	[CTG159
0-0.63		BE490408, HJKAX1S01A7OJD, HJKAX1S01AXE2D, HJKAX1S01BEME5, HJKAX1S01BHONX, HJKAX1S01BHTUD, HJKAX1S01BYJSE, HJKAX1S01BZR5I_1, HJKAX1S01C3QQB, HJKAX1S01C665H, HJKAX1S01CCDHQ, HJKAX1S01CP8L8, HJKAX1S01CRWKS, HJKAX1S01CVI8T, HJKAX1S01D3WSX, HJKAX1S01DKA6E, HJKAX1S01E18B6, HJKAX1S01E4H0T, HJKAX1S01ECB9T	CTG150]
0-0.63		BE498768, BF473658, C0VNS, HJKAX1S01A1RCC, HJKAX1S01A40SK_1, HJKAX1S01A6QJ9_1, HJKAX1S01A9H5F_1, HJKAX1S01A9N2X, HJKAX1S01AEHM7, HJKAX1S01AJAQV, HJKAX1S01ALH1O, HJKAX1S01ALWPD, HJKAX1S01ASL57, HJKAX1S01AXO03, HJKAX1S01B0ABN, HJKAX1S01B14OH, HJKAX1S01B3H5S_3, HJKAX1S01B45BL_1, HJKAX1S01BAWMR, HJKAX1S01BE4KV, HJKAX1S01BEY0P, HJKAX1S01BGVUW, HJKAX1S01BHSRP, HJKAX1S01BJFL1, HJKAX1S01BLWOD, HJKAX1S01BOEIK, HJKAX1S01BTOCJ, HJKAX1S01BV9HJ_1, HJKAX1S01BWULC_1, HJKAX1S01C09WH_1, HJKAX1S01C0VNS_1, HJKAX1S01C1B75, HJKAX1S01C252M, HJKAX1S01C2OOT, HJKAX1S01C4R2N, HJKAX1S01C7HMY, HJKAX1S01CBBE9, HJKAX1S01CBUUH, HJKAX1S01CEFTK_1, HJKAX1S01CG092_1, HJKAX1S01CG5AG, HJKAX1S01CIY61, HJKAX1S01CJPN3, HJKAX1S01CKCMM, HJKAX1S01CPT8Q, HJKAX1S01CTEZK_1, HJKAX1S01D0CJT, HJKAX1S01D2ZTU, HJKAX1S01D3APN, HJKAX1S01D3KTC, HJKAX1S01D6CRO_1, HJKAX1S01D6MZB, HJKAX1S01D9S47, HJKAX1S01DAIY6, HJKAX1S01DBFFW, HJKAX1S01DHE09, HJKAX1S01DNPOR, HJKAX1S01DP9VT, HJKAX1S01DQFRA, HJKAX1S01DQQ02, HJKAX1S01DQWL7, HJKAX1S01DRUIX, HJKAX1S01DUVID, HJKAX1S01E2WE1, HJKAX1S01EDL81, HJKAX1S01EEFIM, HJKAX1S01EKROW, HJKAX1S01EKVCT, HJKAX1S01EZRIJ	CTG125
0-0.63		BE606945, HJKAX1S01A1ZCL_2, HJKAX1S01A4PVO, HJKAX1S01A8D02, HJKAX1S01AY7IZ_1, HJKAX1S01B2EEN, HJKAX1S01B8B7Z, HJKAX1S01BCSK4, HJKAX1S01BGM23, HJKAX1S01BVG5I, HJKAX1S01BVMEB, HJKAX1S01C21ZF, HJKAX1S01C3BWZ, HJKAX1S01C5NQY, HJKAX1S01CEZ4D, HJKAX1S01D0Q65, HJKAX1S01DA0E1, HJKAX1S01DS69M, HJKAX1S01DTMWB, HJKAX1S01EA4QY, HJKAX1S01EIP6C	CTG142
0-0.63		BG604740, HJKAX1S01A9E0N, HJKAX1S01AGXDH_1,HJKAX1S01AJVK3, HJKAX1S01BADKM, HJKAX1S01BI1FF, HJKAX1S01BLOPY, HJKAX1S01C80NZ, HJKAX1S01CT63B, HJKAX1S01D0A1J, HJKAX1S01DA19P, HJKAX1S01DAKW0, HJKAX1S01DIIGQ, HJKAX1S01DQHDD, HJKAX1S01EJSMO, HJKAX1S01EJWQN, HJKAX1S01EK8AT	CTG5
0-0.63		BYIFL, EXX6A, GPI:C:729592, BE591974, BE490781, BE591461, HJKAX1S01A0GAM, HJKAX1S01A3JCJ, HJKAX1S01A3JPQ, HJKAX1S01A4V4G, HJKAX1S01A7D6L, HJKAX1S01A7MCG, HJKAX1S01A7VJ0, HJKAX1S01A9IU7, HJKAX1S01A9L82_1, HJKAX1S01AD2LT, HJKAX1S01AERNY, HJKAX1S01AF62V_1, HJKAX1S01AF8Z4_1, HJKAX1S01AFLP1, HJKAX1S01AGQW8, HJKAX1S01AJVB0_1, HJKAX1S01ALICS_1, HJKAX1S01AOSZE_1, HJKAX1S01AQFDW, HJKAX1S01AQXAZ, HJKAX1S01ASV0A_1, HJKAX1S01ATKU3, HJKAX1S01B0L3S, HJKAX1S01B2H41, HJKAX1S01B2PR8, HJKAX1S01BES6E, HJKAX1S01BH05G, HJKAX1S01BJTGJ, HJKAX1S01BL5SE, HJKAX1S01BL7FA, HJKAX1S01BLL22, HJKAX1S01BMV52, HJKAX1S01BQFRY_2, HJKAX1S01BTY1E, HJKAX1S01BU9MJ, HJKAX1S01BW595, HJKAX1S01C1DLU, HJKAX1S01C5B1Q, HJKAX1S01C5CLD, HJKAX1S01C6F1Z_1, HJKAX1S01C77SJ, HJKAX1S01C9CYC, HJKAX1S01C9LL1, HJKAX1S01CA1CU, HJKAX1S01CF43N, HJKAX1S01CFZ2P, HJKAX1S01CGSHQ, HJKAX1S01CHYBI, HJKAX1S01CIEB1, HJKAX1S01CMOXI, HJKAX1S01CP1C1_1, HJKAX1S01CPU3S, HJKAX1S01CQSVL, HJKAX1S01CUHC7, HJKAX1S01CWVR7_1, HJKAX1S01CY5QG_1, HJKAX1S01CYR9E, HJKAX1S01CYZJP, HJKAX1S01D4AVF, HJKAX1S01DDJ4H, HJKAX1S01DDL6H, HJKAX1S01DEV48, HJKAX1S01DFD81, HJKAX1S01DG3DQ, HJKAX1S01DGSFU, HJKAX1S01DGUFJ, HJKAX1S01DHS1R, HJKAX1S01DJF7V, HJKAX1S01DMRCR, HJKAX1S01DVMQX, HJKAX1S01DWBDC, HJKAX1S01E0O6B, HJKAX1S01E1ENY, HJKAX1S01E2N80, HJKAX1S01E4L46, HJKAX1S01E4RF7, HJKAX1S01EACSM, HJKAX1S01EI1GL, HJKAX1S01EKHG2, HJKAX1S01ENC26, HJKAX1S01EO4N7, HJKAX1S01ERU6A, HJKAX1S01ERW22, HJKAX1S01EXDLS, HJKAX1S01EYHHB, HJKAX1S01EYPXD	[CTG156
0-0.63		WMS358, HJKAX1S01ARCBJ, HJKAX1S01ASPAQ, HJKAX1S01ATZWT, HJKAX1S01B04RQ, HJKAX1S01B64O0_1, HJKAX1S01B8ZKY, HJKAX1S01B8ZKY_1, HJKAX1S01BKP8T, HJKAX1S01BKW53_1, HJKAX1S01BNSOS_1, HJKAX1S01BTIZZ, HJKAX1S01BYFXT, HJKAX1S01CJAA3, HJKAX1S01COOBU_1, HJKAX1S01CPPKC, HJKAX1S01CQR6R, HJKAX1S01CXT3Z, HJKAX1S01CZFHJ, HJKAX1S01D3CPD, HJKAX1S01D3M6P, HJKAX1S01D7PIM_1, HJKAX1S01D8IEF, HJKAX1S01DJAZH, HJKAX1S01DY8OA, HJKAX1S01DYD1U, HJKAX1S01DZGY4, HJKAX1S01EEUPZ, HJKAX1S01EHUJ2, HJKAX1S01EL9F4, HJKAX1S01EMJ57, HJKAX1S01EOI0R	CTG157]

0-0.63		B7QNM, HJKAX1S01B7QNM_1, HJKAX1S01B9MNZ_1, HJKAX1S01BAV97, HJKAX1S01BRWJZ, HJKAX1S01DCJ4B, HJKAX1S01E2WL4	CTG95
0-0.63		AFQ6M, BE404490, HJKAX1S01BOBJ9_1	CTG2
0-0.63		BZB42, CAGIE, BE444644, BE500291, HJKAX1S01A14N8_1, HJKAX1S01A6Y6R, HJKAX1S01A75OR, HJKAX1S01A7YQ8, HJKAX1S01A91XM, HJKAX1S01ADNOS_1, HJKAX1S01AEF3A_1, HJKAX1S01AFGD9, HJKAX1S01AFZAB_1, HJKAX1S01AHD5D_1, HJKAX1S01AI7CB_1, HJKAX1S01AIHGE, HJKAX1S01AOK3I_1, HJKAX1S01AQCXN, HJKAX1S01ARIOW, HJKAX1S01ART2Q_1, HJKAX1S01ARVXO, HJKAX1S01AT38K_1, HJKAX1S01AVXCJ_1, HJKAX1S01B1NJG_2, HJKAX1S01B2LU7_1, HJKAX1S01B3R65_1, HJKAX1S01B3S3B, HJKAX1S01B43O1, HJKAX1S01B8G76, HJKAX1S01BCUJ6, HJKAX1S01BDTVT_1, HJKAX1S01BIUZS, HJKAX1S01BJRZV_1, HJKAX1S01BLR42, HJKAX1S01BLY7X_1, HJKAX1S01BMDH6, HJKAX1S01BN2NQ, HJKAX1S01BPSGZ, HJKAX1S01BSDRV, HJKAX1S01BUKIE_1, HJKAX1S01BVL6K_2, HJKAX1S01BVVMJ_1, HJKAX1S01BYJV6_1, HJKAX1S01BZB42_1, HJKAX1S01C0U0X_1, HJKAX1S01C11DG, HJKAX1S01C3D3M, HJKAX1S01C3DSZ, HJKAX1S01C4PJ6_1, HJKAX1S01C5Z0L_1, HJKAX1S01C871O, HJKAX1S01C9I9E_1, HJKAX1S01C9IE1, HJKAX1S01CAC7F_1, HJKAX1S01CAGIE_1, HJKAX1S01CC4IW, HJKAX1S01CC657, HJKAX1S01CFZNN, HJKAX1S01COA45, HJKAX1S01CPXMH, HJKAX1S01CQA3C, HJKAX1S01CREL0, HJKAX1S01CTCV0, HJKAX1S01CVAAW, HJKAX1S01CXOHS, HJKAX1S01CY9YZ, HJKAX1S01D13LQ, HJKAX1S01D3TDB, HJKAX1S01D4H6V, HJKAX1S01D4MO7, HJKAX1S01DDP19, HJKAX1S01DEZ4S, HJKAX1S01DFEFA, HJKAX1S01DKB2P, HJKAX1S01DO8VD, HJKAX1S01DOB2V, HJKAX1S01DPX42, HJKAX1S01DRJAS, HJKAX1S01DSI8Z, HJKAX1S01DU9TI, HJKAX1S01DX8NU, HJKAX1S01DXVH5, HJKAX1S01DYLHS, HJKAX1S01DYSJD, HJKAX1S01DZX98, HJKAX1S01EA4PP, HJKAX1S01EC836, HJKAX1S01EFB0R, HJKAX1S01EGSL7, HJKAX1S01EGWOI, HJKAX1S01EH27D, HJKAX1S01EKRDC, HJKAX1S01ELHNL, HJKAX1S01ELQ2S, HJKAX1S01ETU2G, HJKAX1S01ETZEL, HJKAX1S01EVTVY, HJKAX1S01EY5TK	CTG138
0-0.63		BE604729, HJKAX1S01A06QF, HJKAX1S01A2GYR, HJKAX1S01A3R4H_1, HJKAX1S01A4MV4, HJKAX1S01A7MQ2, HJKAX1S01ARCHY, HJKAX1S01ARDFO, HJKAX1S01AS7FU_2, HJKAX1S01AW9CU, HJKAX1S01AXSE2_1, HJKAX1S01AZ7G1, HJKAX1S01B07MH_2, HJKAX1S01B0UGI, HJKAX1S01B2JKX, HJKAX1S01B3X79, HJKAX1S01BB4MQ, HJKAX1S01BC2G7, HJKAX1S01BFA3W_1, HJKAX1S01BO3IL_1, HJKAX1S01BPTSV, HJKAX1S01BPYUO, HJKAX1S01BZMB6, HJKAX1S01C3D4K, HJKAX1S01C92SG, HJKAX1S01CA292, HJKAX1S01CCP4R, HJKAX1S01CDBM0_1, HJKAX1S01CFAYF, HJKAX1S01D2OLY, HJKAX1S01D53SX_1, HJKAX1S01D64T0, HJKAX1S01DDAJX, HJKAX1S01DFRTL, HJKAX1S01DGSKS, HJKAX1S01DH5EM, HJKAX1S01DII0K, HJKAX1S01E5953, HJKAX1S01ECHJE, HJKAX1S01EED8I, HJKAX1S01EHWNI, HJKAX1S01ETXRX	CTG135
0-0.63		BTCOT, GH722882, HJKAX1S01A3M9G, HJKAX1S01A7SYZ, HJKAX1S01A99ZE, HJKAX1S01AZ5XR, HJKAX1S01B6W77, HJKAX1S01B8KRV, HJKAX1S01BLNL2, HJKAX1S01BUGTT, HJKAX1S01C0AZM, HJKAX1S01C0SVI, HJKAX1S01C2QQP_1, HJKAX1S01C8Q9W, HJKAX1S01C9SJ1, HJKAX1S01CDGMS, HJKAX1S01CIFWA, HJKAX1S01CQ0M0_1, HJKAX1S01CSNIP, HJKAX1S01CUEGO, HJKAX1S01DCZ9G, HJKAX1S01DER4J, HJKAX1S01DHA7J, HJKAX1S01E23NS, HJKAX1S01EDT5X, HJKAX1S01EFQDK, HJKAX1S01EVKXG, HJKAX1S01EVYMG, HJKAX1S01EWBBH	CTG96
0-0.63		HJKAX1S01AKAVO, HJKAX1S01C2VXY, HJKAX1S01DA8PE, HJKAX1S01DV89Q	CTG61
0-0.63		HJKAX1S01A6ED3_3, HJKAX1S01A6FRH, HJKAX1S01ANBH4, HJKAX1S01B1HV3_1, HJKAX1S01B4XDY, HJKAX1S01B5MQ7, HJKAX1S01BA33Q, HJKAX1S01BE2Z7_2, HJKAX1S01BJQG7, HJKAX1S01BK8WL, HJKAX1S01BNLCQ_1, HJKAX1S01BOZPD, HJKAX1S01BRF2J, HJKAX1S01BS34C_2, HJKAX1S01BWU5Y, HJKAX1S01CEU9S, HJKAX1S01CGN2C, HJKAX1S01CHGID, HJKAX1S01CI6CC_1, HJKAX1S01CIG0P, HJKAX1S01CRTSC, HJKAX1S01D2XQE, HJKAX1S01D3HL3, HJKAX1S01DEUDO, HJKAX1S01DFUEE, HJKAX1S01DGERT, HJKAX1S01DGO4L, HJKAX1S01DIV4Y, HJKAX1S01DJGXM, HJKAX1S01DLWYL, HJKAX1S01DWO9Y, HJKAX1S01EBY4L, HJKAX1S01EH1Y3, HJKAX1S01EJF10, HJKAX1S01ENNHQ, HJKAX1S01ESWWG, HJKAX1S01EWFTN	CTG87
0-0.63		HJKAX1S01BN0H2	CTG55
0-0.63		HJKAX1S01A353Y, HJKAX1S01AJLEK, HJKAX1S01AMRNL, HJKAX1S01B79U7_1, HJKAX1S01BVOG0_3, HJKAX1S01C5P90, HJKAX1S01EFC5Y	[CTG74-CTG109]
0-0.63		HJKAX1S01ANPI7, HJKAX1S01ANPI7_1, HJKAX1S01AVSNZ_1, HJKAX1S01AY9ZK_1, HJKAX1S01B45J4, HJKAX1S01B7TQH_1, HJKAX1S01C4BTV, HJKAX1S01CJN0Q_2, HJKAX1S01CKPWS, HJKAX1S01CZTRC, HJKAX1S01E3KDN, HJKAX1S01EHR2B	CTG21
0-0.63		BE471016, HJKAX1S01A766A, HJKAX1S01A777G, HJKAX1S01A9714, HJKAX1S01AKMFT_1, HJKAX1S01AQOX2_1, HJKAX1S01AYT8J, HJKAX1S01B1Z4F, HJKAX1S01B3QTI_1, HJKAX1S01B5VCT, HJKAX1S01BR7WR_1, HJKAX1S01BXPEO, HJKAX1S01CKP5W, HJKAX1S01CMJHG, HJKAX1S01CZ1ZH_1, HJKAX1S01D70US, HJKAX1S01DGY1H, HJKAX1S01DQCJ7, HJKAX1S01EABJD, HJKAX1S01EQEFN, HJKAX1S01EVNEI, HJKAX1S01EYWXU, HJKAX1S01EZVKE	CTG83
0-0.63		BE403618, BE606654, BF485220, HJKAX1S01A1JA2, HJKAX1S01A5EE1_2, HJKAX1S01A5IOK_1, HJKAX1S01AQX07, HJKAX1S01BJABO_1, HJKAX1S01BXGZV_1, HJKAX1S01C01HO, HJKAX1S01C6XPU, HJKAX1S01CYQL1_1, HJKAX1S01DBDQI_2	[CTG127-CTG77]
0-0.63		BE405839, HJKAX1S01A06F6_1, HJKAX1S01A51MN, HJKAX1S01AION7_1, HJKAX1S01ARMA5_1, HJKAX1S01ATWON_1, HJKAX1S01AYJWI, HJKAX1S01B47DU, HJKAX1S01B82XB_1, HJKAX1S01BD76B, HJKAX1S01BG6FB, HJKAX1S01BI3PM, HJKAX1S01CC7DY, HJKAX1S01DDPND, HJKAX1S01DLITF, HJKAX1S01EQNFH, HJKAX1S01ES7OU, HJKAX1S01ETB3Y	CTG53
0-0.63		HJKAX1S01AEPAH_1, HJKAX1S01ANN34, HJKAX1S01ATX39, HJKAX1S01AW9QN, HJKAX1S01AYBTC, HJKAX1S01AZ089, HJKAX1S01B3UBU, HJKAX1S01BLOQY, HJKAX1S01C6LYA_1, HJKAX1S01CMQVW, HJKAX1S01DPHIW, HJKAX1S01EL74D	[CTG136-CTG148]
0-0.63	mapped to 0-0.63 but order is uncertain	BE424775, HJKAX1S01ANFYW, HJKAX1S01AP0TA_1, HJKAX1S01AQ36F_1, HJKAX1S01ATB7G, HJKAX1S01AUTOL_1, HJKAX1S01BA23Z_1, HJKAX1S01BAS0I, HJKAX1S01BRN80, HJKAX1S01BRUTM, HJKAX1S01BS4DX, HJKAX1S01BYQK1, HJKAX1S01BYSP1_1, HJKAX1S01CCR76_2, HJKAX1S01CL3NH, HJKAX1S01CONOO_1, HJKAX1S01CRZ16, HJKAX1S01CZMAY, HJKAX1S01DHLU4, HJKAX1S01DYLBL, HJKAX1S01EXG64, HJKAX1S01EXRJV, HJKAX1S01EXTWT, synopGBS118852	CTG137
0-0.63		GPI:C:743567, BF291319, BF485220	[CTG88-CTG90]
0-0.63		BT6NA	[CTG58-CTG56]
0-0.63		CFD189, HJKAX1S01BU82A_1	CTG113

APPENDIX J

Reagents and molecular biology kits

6X DNA Loading dye	Thermo Scientific	R0611
Agarose	Sigma	A5093
Boric acid	Sigma	B6768
dNTP Mix	Thermo Scientific	R0193
Ethidium bromide	Applichem	A1151
Ethylenediaminetatraaceticacid (EDTA)	Calbiochem	324503
Ethyl Alcohol Absolut %99.8	Riedel de Haen	32221
GeneRuler 100 bp DNA Ladder	Thermo Scientific	SM0241
GeneRuler DNA Ladder Mix	Thermo Scientific	SM0332
Isopropanol	Merck	1.09634
Nuclease free water	Qiagen	129114
Sodium dodecyl sulfate	Molekula	15171947
Taq DNA polymerase (recombinant)	Thermo Scientific	EP0401
i-Taq^{TM} DNA Polymerase	iNtRON	25022
Trizma(R) base>=99.9%(titration)	Sigma	T1503
Tween$^{®}$ 20	Sigma	P2287
Wizard$^{®}$ Genomic DNA Purification Kit	Promega	A1120

APPENDIX K

Equipment

Autoclave	Hirayama, Hiclave HV-110, JAPAN
	Nüve 0T 032, TURKEY
Balance	Sartorius, BP221S, GERMANY
	Schimadzu, Libror EB-3 200 HU, JAPAN
Centrifuge	Microfuge 18 Centrifuge Beckman Coulter, USA
	Eppendorf, 5415D, GERMANY
	Eppendorf, 5415R, GERMANY
Deepfreeze	-20^{o}C Bosch, TURKEY
	-80^{o}C Thermo electron corporation, USA
Distilled water	Millipore, MilliQ Academic, FRANCE
Electrophoresis	ENDURO™ Gel XL Electrophoresis System, USA
	Labnet, Electrophoresis-Gel System, USA
Gel documentation	Biorad, UV-Transilluminator 2000, USA
Heating block	Thermostat Bio TDB-100, LATVIA
Ice machine	Scotsman Inc., AF20, USA
Incubator	Memmert, Modell 300, GERMANY
Laminar flow	Heraeus, Modell HS 12, GERMANY
Magnetic stirrer	VELP Scientifica, ITALY
Microarray system	Microarray hybridization chamber and assemblies, G2534A, Agilent, USA
	NimbleGen MS 200 Microarray Scanner
Micropipettes	Gilson, Pipetman, FRANCE
	Eppendorf, GERMANY
	Thermo Scientific, USA
Microwave oven	Bosh, TURKEY
Nitrogen tanks	Linde Industrial Gases, TURKEY
pH meter	WTW, pH540, GLP MultiCal, GERMANY
Refrigerator	+4 ^{o}C Bosh, TURKEY
Sequencer	Roche 454 GS FLX Sequencer, Basel, SWITZERLAND

Thermal cycler	Eppendorf, Mastercycler Gradient, GERMANY
	Prime Elite Thermal Cycler, Techne, UK
	PTC-100® Thermal Cycler, Biorad, USA
Tissue lyser	Qiagen Retsch, USA
Vacuum	Heto, MasterJet Sue 300Q, DENMARK
Vortex mixer	Stuart, SA8, UK
Water bath	Memmert, GERMANY

REFERENCES

Agah, A., Aghajan, M., Mashayekhi, F., Amini, S., Davis, R.W., Plummer, J.D., Ronaghi, M. and Griffin, P.B. **(2004)** A multi-enzyme model for Pyrosequencing. *Nucleic Acids Res.*, 32, e166.

Akpinar, B.A., Lucas, S.J. and Budak, H. **(2013)** Genomics approaches for crop improvement against abiotic stress. *Sci. World J.*, 2013.

Akpinar, B.A., Lucas, S.J., Vr, J., Dole, J. and Budak, H. **(2014)** Sequencing chromosome 5D of Aegilops tauschii and comparison with its allopolyploid descendant bread wheat (*Triticum aestivum*). 15:1080, 1–13.

Allen, A.M., Barker, G.L., Wilkinson, P., et al. **(2013)** Discovery and development of exome-based, co-dominant single nucleotide polymorphism markers in hexaploid wheat (*Triticum aestivum* L.). *Plant Biotechnol J*, 11, 279–295.

Berkman, P.J., Skarshewski, A., Lorenc, M.T., et al. **(2011)** Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. *Plant Biotechnol. J.*, 9, 768–75.

Berkman, P.J., Skarshewski, A., Manoli, S., et al. **(2012)** Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theor. Appl. Genet.*, 124, 423–32.

Bolger, M.E., Weisshaar, B., Scholz, U., Stein, N., Usadel, B. and Mayer, K.F.X. **(2014)** Plant genome sequencing - applications for crop improvement. *Curr. Opin. Biotechnol.*, 26, 31–7.

Bolot, S., Abrouk, M., Masood-Quraishi, U., Stein, N., Messing, J., Feuillet, C. and Salse, J. **(2009)** The “inner circle” of the cereal genomes. *Curr. Opin. Plant Biol.*, 12, 119–25.

Breen, J., Wicker, T., Shatalina, M., et al. **(2013)** A physical map of the short arm of wheat chromosome 1A. *PLoS One*, 8, e80272.

Brenchley, R., Spannagl, M., Pfeifer, M., et al. **(2012)** Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, 491, 705–10.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. **(2009)** BLAST+: architecture and applications. *BMC Bioinformatics*, 10, 421.

Choulet, F., Alberti, A., Theil, S., et al. **(2014)** Structural and functional partitioning of bread wheat chromosome 3B. *Science*, 345(SI), 1249721.

Choulet, F., Wicker, T., Rustenholz, C., et al. **(2010)** Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell*, 22, 1686–1701.

Cloutier, S., McCallum, B.D., Loutre, C., Banks, T.W., Wicker, T., Feuillet, C., Keller, B. and Jordan, M.C. **(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.

Conesa, A. and Götz, S. **(2008)** Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics*, 2008, 619832.

Deng, W., Casao, M.C., Wang, P., Sato, K., Hayes, P.M., Finnegan, E.J. and Trevaskis, B. **(2015)** Direct links between the vernalization response and other key traits of cereal crops. *Nat. Commun.*, 6, 5882.

Doležel, J., Vrána, J., Safář, J., Bartoš, J., Kubaláková, M. and Simková, H. **(2012)** Chromosomes in the flow to simplify genome analysis. *Funct. Integr. Genomics*, 12, 397–416.

Edwards, D. and Batley, J. **(2010)** Plant genome sequencing: Applications for crop improvement. *Plant Biotechnol. J.*, 8, 2–9.

Endo, T.R. and Gill, B.S. **(1996)** The Deletion Stocks of Common Wheat. *J. Hered.*, 87, 295–307.

Ergen, N.Z. and Budak, H. **(2009)** Sequencing over 13 000 expressed sequence tags from six subtractive cDNA libraries of wild and modern wheats following slow drought stress. *Plant. Cell Environ.*, 32, 220–36.

FAO **(2013)** PART 3 Feeding the world. In *FAO Statistical Yearbook 2013*. pp. 123–158.

Feuillet, C., Langridge, P. and Waugh, R. **(2007)** Cereal breeding takes a walk on the wild side. *Trends Genet.*, 24, 24–32.

Feuillet, C., Leach, J.E., Rogers, J., Schnable, P.S. and Eversole, K. **(2011)** Crop genome sequencing: lessons and rationales. *Trends Plant Sci.*, 16, 77–88.

Foley, J.A., Ramankutty, N., Brauman, K.A., et al. **(2011)** Solutions for a cultivated planet. *Nature*, 478, 337–42.

Frenkel, Z., Paux, E., Mester, D., Feuillet, C. and Korol, A. **(2010)** LTC: a novel algorithm to improve the efficiency of contig assembly for physical mapping in complex genomes. *BMC Bioinformatics*, 11, 584.

Gill, B.S., Appels, R., Botha-Oberholster, A.M., et al. **(2004)** A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium. *Genetics*, 168, 1087–96.

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

Henry, I.M., Nagalakshmi, U., Lieberman, M.C., et al. **(2014)** Efficient Genome-Wide Detection and Cataloging of EMS-Induced Mutations Using Exome Capture and Next-Generation Sequencing. *Plant Cell*, 26, 1382–1397.

Hernandez, P., Martis, M., Dorado, G., et al. **(2012)** Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *Plant J.*, 69, 377–86.

Hutchison, C.A. **(2007)** DNA sequencing: Bench to bedside and beyond. *Nucleic Acids Res.*, 35, 6227–6237.

Jackson, S.A., Iwata, A., Lee, S.H., Schmutz, J. and Shoemaker, R. **(2011)** Sequencing crop genomes: Approaches and applications. *New Phytol.*, 191, 915–925.

Jia, J., Zhao, S., Kong, X., et al. **(2013)** *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature*, 496, 91–5.

Jurka, J., Kapitonov, V. V, Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. **(2005)** Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, 110, 462–7.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. **(2009)** Circos: an information aesthetic for comparative genomics. *Genome Res.*, 19, 1639–45.

Kubaláková, M., Vrána, J., Cíhalíková, J., Simková, H. and Dolezel, J. **(2002)** Flow karyotyping and chromosome sorting in bread wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.*, 104, 1362–1372.

Kumar, M., Gantasala, N.P., Roychowdhury, T., Thakur, P.K., Banakar, P., Shukla, R.N., Jones, M.G.K. and Rao, U. **(2014)** De novo transcriptome sequencing and analysis of the cereal cyst nematode, *Heterodera avenae*. *PLoS One*, 9, e96311.

Lander, E.S., Linton, L.M., Birren, B., et al. **(2001)** Initial sequencing and analysis of the human genome. *Nature*, 409, 860–921.

Li, W., Zhang, P., Fellers, J.P., Friebe, B. and Gill, B.S. **(2004)** Sequence composition, organization, and evolution of the core Triticeae genome. *Plant J.*, 40, 500–11.

Ling, H.-Q., Zhao, S., Liu, D., et al. **(2013)** Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature*, 496, 87–90.

Lobell, D.B., Schlenker, W. and Costa-Roberts, J. **(2011)** Climate trends and global crop production since 1980. *Science*, 333, 616–621.

Lowe, T.M. and Eddy, S.R. **(1997)** tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res.*, 25, 955–964.

Lucas, S.J., Akpınar, B.A., Kantar, M., et al. **(2013)** Physical mapping integrated with syntenic analysis to characterize the gene space of the long arm of wheat chromosome 1A. *PLoS One*, 8, e59542.

Lucas, S.J., Akpınar, B.A., Simková, H., Kubalakova, M., Dolezel, J. and Budak, H. **(2014)** Next-generation sequencing of flow-sorted wheat chromosome 5D reveals lineage-specific translocations and widespread gene duplications. *BMC Genomics*, 15, 1–18.

Lucas, S.J., Šimková, H., Šafář, J., et al. **(2012)** Functional features of a single chromosome arm in wheat (1AL) determined from its structure. *Funct. Integr. Genomics*, 12, 173–82.

Luo, M., Gu, Y.Q., You, F.M., et al. **(2013)** A 4-gigabase physical map unlocks the structure and evolution of the complex genome of *Aegilops tauschii*, the wheat D-genome progenitor. *PNAS*, 110:19, 7940-45.

Luo, M.C., Thomas, C., You, F.M., et al. **(2003)** High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics*, 82, 378–389.

Marcussen, T., Sandve, S.R., Heier, L., et al. **(2014)** Ancient hybridizations among the ancestral genomes of bread wheat. *Science,* 345(SI), 1250092.

Margulies, M. **(2005)** Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437, 376–380.

Mayer, K.F.X., Martis, M., Hedley, P.E., et al. **(2011)** Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell*, 23, 1249–63.

Mayer, K.F.X., Rogers, J., Dole el, J., et al. **(2014)** A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science,* 345(SI), 1251788.

Mayer, K.F.X., Taudien, S., Martis, M., et al. **(2009)** Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol.*, 151, 496–505.

Mcintosh, R. a, Yamazaki, Y., Dubcovsky, J., Rogers, J., Morris, C. and Somers, D.J. **(2008)** 11 th International Wheat Genetics Symposium Brisbane Qld Australia Catalogue of Gene Symbols for Wheat. *Plant Breed.*

Metzker, M.L. **(2010)** Sequencing technologies - the next generation. *Nat. Rev. Genet.*, 11, 31–46.

Metzker, M.L. **(2005)** Emerging technologies in DNA sequencing. *Genome Res.*, 15, 1767–76.

Middleton, C.P., Stein, N., Keller, B., Kilian, B. and Wicker, T. **(2012)** Comparative analysis of genome composition in Triticeae reveals strong variation in transposable element dynamics and nucleotide diversity. *Plant J.*, 73, 347-56.

Morrell, P.L., Buckler, E.S. and Ross-Ibarra, J. **(2011)** Crop genomics: advances and applications. *Nat. Rev. Genet.*, 13, 85–96.

Morris, C.F. **(2002)** Puroindolines: the molecular genetic basis of wheat grain hardness. *Plant Mol. Biol.*, 48, 633–47.

Murat, F., Zhang, R., Guizard, S., et al. **(2014)** Shared subgenome dominance following polyploidization explains grass genome evolutionary plasticity from a seven protochromosome ancestor with 16K protogenes. *Genome Biol. Evol.*, 6, 12–33.

Nelson, J.C., Sorrells, M.E., Deynze, A.E. Van, Lu, Y.H., Atkinson, M., Bernard, M., Leroy, P., Faris, J.D. and Anderson, J.A. **(1995)** Molecular mapping of wheat: major genes and rearrangements in homoeologous groups 4, 5, and 7. *Genetics*, 141, 721–31.

Nelson, W.M., Bharti, A.K., Butler, E., Wei, F., Fuks, G., Kim, H., Wing, R.A., Messing, J. and Soderlund, C. **(2005)** Whole-genome validation of high-information-content fingerprinting. *Plant Physiol.*, 139, 27–38.

Ouyang, S. and Buell, C.R. **(2004)** The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res.*, 32, D360–3.

Ouyang, S., Zhang, D., Han, J., et al. **(2014)** Fine physical and genetic mapping of powdery mildew resistance gene MlIW172 originating from wild emmer (*Triticum dicoccoides*). *PLoS One*, 9, e100160.

Paterson, A.H., Bowers, J.E., Bruggmann, R., et al. **(2009)** The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, 457, 551–6.

Paux, E., Faure, S., Choulet, F., et al. **(2010)** Insertion site-based polymorphism markers open new perspectives for genome saturation and marker-assisted selection in wheat. *Plant Biotechnol. J.*, 8, 196–210.

Paux, E., Roger, D., Badaeva, E., Gay, G., Bernard, M., Sourdille, P. and Feuillet, C. **(2006)** Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. *Plant J.*, 48, 463–474.

Paux, E., Sourdille, P., Salse, J., et al. **(2008)** A physical map of the 1-gigabase bread wheat chromosome 3B. *Science*, 322, 101–4.

Philippe, R., Paux, E., Bertin, I., et al. **(2013)** A high density physical map of chromosome 1BL supports evolutionary studies, map-based cloning and sequencing in wheat. *Genome Biol.*, 14, R64.

Poland, J.A., Brown, P.J., Sorrells, M.E. and Jannink, J.L. **(2012)** Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One*, 7, e32253.

Pont, C., Murat, F., Guizard, S., et al. **(2013)** Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. *Plant J.*, 76, 1030–44.

Poursarebani, N., Nussbaumer, T., Šimková, H., et al. **(2014)** Whole-genome profiling and shotgun sequencing delivers an anchored, gene-decorated, physical map assembly of bread wheat chromosome 6A. *Plant J.*, 79, 334–347.

Quraishi, U.M., Abrouk, M., Bolot, S., et al. **(2009)** Genomics in cereals: from genome-wide conserved orthologous set (COS) sequences to candidate genes for trait dissection. *Funct. Integr. Genomics*, 9, 473–84.

Raats, D., Frenkel, Z., Krugman, T., et al. **(2013)** The physical map of wheat chromosome 1BS provides insights into its gene space organization and evolution. *Genome Biol.*, 14, R138.

Reynolds, M., Foulkes, M.J., Slafer, G. a, Berry, P., Parry, M. a J., Snape, J.W. and Angus, W.J. **(2009)** Raising yield potential in wheat. *J. Exp. Bot.*, 60, 1899–918.

Rustenholz, C., Hedley, P.E., Morris, J., Choulet, F., Feuillet, C., Waugh, R. and Paux, E. **(2010)** Specific patterns of gene space organisation revealed in wheat by using the combination of barley and wheat genomic resources. *BMC Genomics*, 11, 714.

Safár, J., Simková, H., Kubaláková, M., Cíhalíková, J., Suchánková, P., Bartos, J. and Dolezel, J. **(2010)** Development of chromosome-specific BAC resources for genomics of bread wheat. *Cytogenet. Genome Res.*, 129, 211–23.

Saintenac, C., Jiang, D. and Akhunov, E.D. **(2011)** Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Genome Biol.*, 12, R88.

Salse, J. **(2012)** In silico archeogenomics unveils modern plant genome organisation, regulation and evolution. *Curr. Opin. Plant Biol.*, 15, 122–130.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M. and Smith, M. **(1977a)** Nucleotide sequence of bacteriophage φX174 DNA. *Nature*, 265, 687–695.

Sanger, F., Nicklen, S. and Coulson, A.R. **(1977b)** DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.*, 74, 5463–7.

Scalabrin, S., Morgante, M. and Policriti, A. **(2009)** Automated FingerPrint Background removal: FPB. *BMC Bioinformatics*, 10, 127.

Sears, E. and Sears, L. **(1978)** The telocentric chromosomes of common wheat. *Proc 5th Int Wheat Genet Symp*, 389–407.

Sehgal, S.K., Li, W., Rabinowicz, P.D., Chan, A., Simková, H., Doležel, J. and Gill, B.S. **(2012)** Chromosome arm-specific BAC end sequences permit comparative analysis of homoeologous chromosomes and genomes of polyploid wheat. *BMC Plant Biol.*, 12, 64.

Shemesh, R., Novik, A. and Cohen, Y. **(2010)** Follow the leader: preference for specific amino acids directly following the initial methionine in proteins of different organisms. *Genomics, Proteomics Bioinformatics*, 8, 180–9.

Simková, H., Safář, J., Kubaláková, M., et al. **(2011)** BAC libraries from wheat chromosome 7D: efficient tool for positional cloning of aphid resistance genes. *J. Biomed. Biotechnol.*, 2011, 302543.

Simková, H., Svensson, J.T., Condamine, P., et al. **(2008)** Coupling amplified DNA from flow-sorted chromosomes to high-density SNP mapping in barley. *BMC Genomics*, 9, 294.

Smith, D.B. and Flavell, R.B. **(1975)** Characterisation of the wheat genome by renaturation kinetics. *Chromosoma*, 50.

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

Tanaka, T., Antonio, B.A., Kikuchi, S., et al. **(2008)** The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res.*, 36, D1028–33.

Tanaka, T.S., Kobayashi, F.U., Joshi, G.P., et al. **(2014)** Next-Generation Survey Sequencing and the Molecular Organization of Wheat Chromosome 6B, *DNA Res.*, 21, 103–4.

Thakur, K., Chawla, V., Bhatti, S., Swarnkar, M.K., Kaur, J., Shankar, R. and Jha, G. **(2013)** De novo transcriptome sequencing and analysis for *Venturia inaequalis*, the devastating apple scab pathogen. *PLoS One*, 8, e53937.

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

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

The UniProt Consortium **(2012)** Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, 40, D71–5.

Varshney, R.K., Hoisington, D.A. and Tyagi, A.K. **(2006)** Advances in cereal genomics and applications in crop breeding. *Trends Biotechnol.*, 24, 490–9.

Varshney, R.K., Nayak, S.N., May, G.D. and Jackson, S.A. **(2009)** Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.*, 27, 522–30.

Venter, J.C., Adams, M.D., Myers, E.W., et al. **(2001)** The sequence of the human genome. *Science*, 291, 1304–51.

Vitulo, N., Albiero, A., Forcato, C., et al. **(2011)** First survey of the wheat chromosome 5A composition through a next generation sequencing approach. *PLoS One*, 6, e26421.

Vrana, J., Kubalakova, M., Simkova, H., Cihalikova, J., Lysak, M.A. and Dolezel, J. **(2000)** Flow Sorting of Mitotic Chromosomes in Common Wheat (*Triticum aestivum* L.). *Genetics*, 156, 2033–2041.

Wang, Z., Cui, Y., Chen, Y., et al. **(2014)** Comparative genetic mapping and genomic region collinearity analysis of the powdery mildew resistance gene Pm41. *Theor. Appl. Genet.*, 127, 1741–51.

Wang, Z., Gerstein, M. and Snyder, M. **(2009)** RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10, 57–63.

Wendler, N., Mascher, M., Nöh, C., Himmelbach, A., Scholz, U., Ruge-Wehling, B. and Stein, N. **(2014)** Unlocking the secondary gene-pool of barley with next-generation sequencing. *Plant Biotechnol. J.*, 12, 1122–31.

Wicker, T., Buchmann, J.P. and Keller, B. **(2010)** Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res.*, 20, 1229–1237.

Wicker, T., Mayer, K.F.X., Gundlach, H., et al. **(2011)** Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *Plant Cell*, 23, 1706–18. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3123954&tool=pmcentrez&rendertype=abstract [Accessed January 10, 2015].

Winfield, M.O., Wilkinson, P.A., Allen, A.M., et al. **(2012)** Targeted re-sequencing of the allohexaploid wheat exome. *Plant Biotechnol. J.*, 10, 733–42.

Wu, H., Qin, J., Han, J., et al. **(2013)** Comparative high-resolution mapping of the wax inhibitors Iw1 and Iw2 in hexaploid wheat. *PLoS One*, 8, e84691.

Yoshida, T., Nishida, H., Zhu, J., Nitcher, R., Distelfeld, A., Akashi, Y., Kato, K. and Dubcovsky, J. **(2010)** Vrn-D4 is a vernalization gene located on the centromeric region of chromosome 5D in hexaploid wheat. *Theor. Appl. Genet.*, 120, 543–52.

You, F.M., Huo, N., Deal, K.R., Gu, Y.Q., Luo, M.-C., McGuire, P.E., Dvorak, J. and Anderson, O.D. **(2011)** Annotation-based genome-wide SNP discovery in the large and complex Aegilops tauschii genome using next-generation sequencing without a reference genome sequence. *BMC Genomics*, 12, 59.

You, F.M., Luo, M.-C., Gu, Y.Q., Lazo, G.R., Deal, K., Dvorak, J. and Anderson, O.D. **(2007)** GenoProfiler: batch processing of high-throughput capillary fingerprinting data. *Bioinformatics*, 23, 240–2.

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

Zhang, H., Guan, H., Li, J., et al. **(2010)** Genetic and comparative genomics mapping reveals that a powdery mildew resistance gene Ml3D232 originating from wild emmer co-segregates with an NBS-LRR analog in common wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.*, 121, 1613–21.

Zhang, J., Wang, Y., Wu, S., Yang, J., Liu, H. and Zhou, Y. **(2012)** A single nucleotide polymorphism at the Vrn-D1 promoter region in common wheat is associated with vernalization response. *Theor. Appl. Genet.*, 125, 1697–704.