Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Scienze e Tecnologie Agrarie, Ambientali e Alimentari

Ciclo XXXI

Settore Concorsuale di afferenza: 07/E1

Settore Scientifico disciplinare: AGR/07

# Durum wheat gene annotation and selection signature detection in the tetraploid wheat germplasm from wild emmer to modern durum wheat

Presentata da: Dott.ssa Danara Ormanbekova

Coordinatore Dottorato :

Chiar.mo Prof. Massimiliano Petracci

Relatore:

Chiar.mo Prof. Roberto Tuberosa

Correlatori:

Dott. Marco Maccaferri

Dott. Sven O. Twardziok

Esame finale anno 2019

# Table of Contents

# 1. Introduction

### 1.1. Taxonomy, origin and evolution of wheat – a brief overview

The species of the genus *Triticum* are classified into three categories based on the level of ploidy with 7 basic sets of chromosomes:

1. Diploid series: 2n=2x=14, genome $A^m$ or A (*T. monococcum, T. urartu, T. speltoides, T. tauschii*);
2. Allotetraploid series: 2n = 4x = 28, AB (*T. turgidum*) or AG (*T. timopheevii*);
3. Allohexaploid series: 2n = 6x = 42, ABD (*T. aestivum*) or $AGA^m$ (*T. zhukovskyi*) (Bennici, 1986; Bálint et al., 2000).

The genus *Triticum* L. belongs to the (*Poaceae*) *Gramineae* family and to the *Triticeae* tribe and *Triticinae* subtribe. *Aegilops, Secale, Agropyron* and *Haynaldia* are other genera belonging to the *Triticinae* subtribe (Gupta, 1972). Although, the *Aegilops* and *Triticum* genera are closely related, they are treated as separate genera as suggested in the Linnaeus's pioneer classification of plants described in his book entitled *Species plantarum* (Linné et al., 1753). The earliest classification of *Triticum* was based on the morphological differences. In 1913 Schulz (Schulz, 1913) classified the *Triticum* species in three main taxonomic groups called Einkorn, Emmer and Dinkel. Subsequently, based on the result of the cytological study conducted by Sakamura (Sakamura, 1918) it was discovered that the three groups differed also in chromosome number: Einkorn is diploid (2n), Emmer is tetraploid (4n) and Dinkel is hexaploid (6n), all with 7 basic sets of chromosomes (Sax and Sax, 1924; Kihara, 1937; McFadden and Sears, 1944). Soon after, Sakamura's student Kihara in his dissertation designated the genomic formulae for diploid, tetraploid and hexaploid wheats as AA, AABB and AABBDD, respectively (Kihara, 1924, 1930). Since these discoveries, there were continuous modifications in the classification of wheat species. In 1979 Dorofeev et al. published one of the earliest nomenclature classifications of *Triticum* (Dorofeev et al., 1979). Later, Mac Key (MacKey, 1966, 1988) and van Slageren (van Slageren, 1994) introduced a new classification system with slight modifications. Today, these classifications are still followed by the scientific community (for a comparative classification table see Kilian et al., 2010).

The origin of the polyploid *Triticum* species is mostly based on allopolyploidization; various interspecific crosses occurred between the genus *Triticum* and *Aegilops* (Kerby and Kuspira, 1987). Initially it was thought that the donor of the A genome was *T. monococcum* ssp*. aegilopoides*. Later several studies confirmed that the A genome was contributed to both *Triticum durum* and *Triticum aestivum* species by the wild wheat *Triticum urartu* (Belea, 1971; Chapman et al., 1976; Nishikawa, 1983; Sallares and Brown, 1999; Dvořák, 2001). In addition to these studies, it is known that *T. monococcum* is more resistant to stem and leaf rust, while *T. urartu* like other wheat species is

susceptible to these diseases (Belea and Fejér, 1980). Subsequently, the formation of the tetraploid species *T. turgidum* ssp*. dicoccoides* (AB) may have occurred due to the hybridisation between *Triticum urartu* (A) and a species related to the *Sitopsis* section of the *Aegilops* genus (S) (van Slageren, 1994; Bálint et al., 2000; Dvořák, 2001). In 1999 Nancy Blake et al. concluded that none of the species belonging to the *Sitopsis* section of *Aegilops* is a donor, but are actually the descendants of the progenitor (Blake et al., 1999). However, soon later, RFLP analysis showed that out of the five species belonging to the *Sitopsis* section of *Aegilops*, the cytoplasm of *T. turgidum* and *T. aestivum* resulted to be closely related to *Ae. speltoids* (Wang et al., 2000).

Wild emmer wheat *T. turgidum* ssp*. dicoccoides* is believed to be the ancestor of the cultivated *turgidum* forms as it is the most ancient species belonging to the *turgidum* group. The *T. turgidum* species are easy to cross and are able to produce fertile progeny (Bálint et al., 2000).

The hexaploid species *T. aestivum* (ABD) known as common or bread wheat evolved over the last 10,000 - 8,500 years and may have occurred due to the second hybridization that involved cultivated emmer *T. turgidum* ssp. *dicoccon* (AB genome), a descendant of *T. turgidum* ssp*. dicoccoides*, and *Ae. tauschii* (D genome). *T. aestivum* ssp. *spelta* is believed to be the ancient form of hexaploid wheat and was initially found in Europe (McFadden and Sears, 1944) and later the cultivated forms were found in Asia (Kuckuck and Schiemann, 1957). *T. aestivum* ssp. *spelta* is a hulled subspecies of hexaploid wheat. This ancient form may have given rise to the naked types of wheat, including *T. aestivum* ssp*. aestivum*, mostly known today as a common wheat.

*T. urartu* is the donor of the A genome for all the polyploid wheat species, except for *T. zhukovskyi* ($AGA^m$). The last may be considered a spontaneous amphiploid of *T. monococcum* ($A^m$) and *T. timopheevii (*AG) (Belea and Fejér, 1980). The ancestors of the tetraploid *T. timopheevii* (AG) may have been the A genome of *T. urartu* and S genome *Ae. speltoides*. Although *Ae. speltoids* is the probable donor of B and G genomes of *timopheevii* and *turgidum* groups, the G genome results to be almost identical to the S genome of *Ae. speltoids*, while the B genome probably was subjected to evolutionary divergence, and shows only slight similarity with the S genome of *Ae. speltoids* (Dvorak and Zhang, 1990; Mori et al., 1997; Wang et al., 2000). Thus, tetraploid species might have undergone two different evolutionary pathways suggesting that the *turgidum* group is relatively ancient compared to the later developed *timopheevii* group (Dvorak and Zhang, 1990).

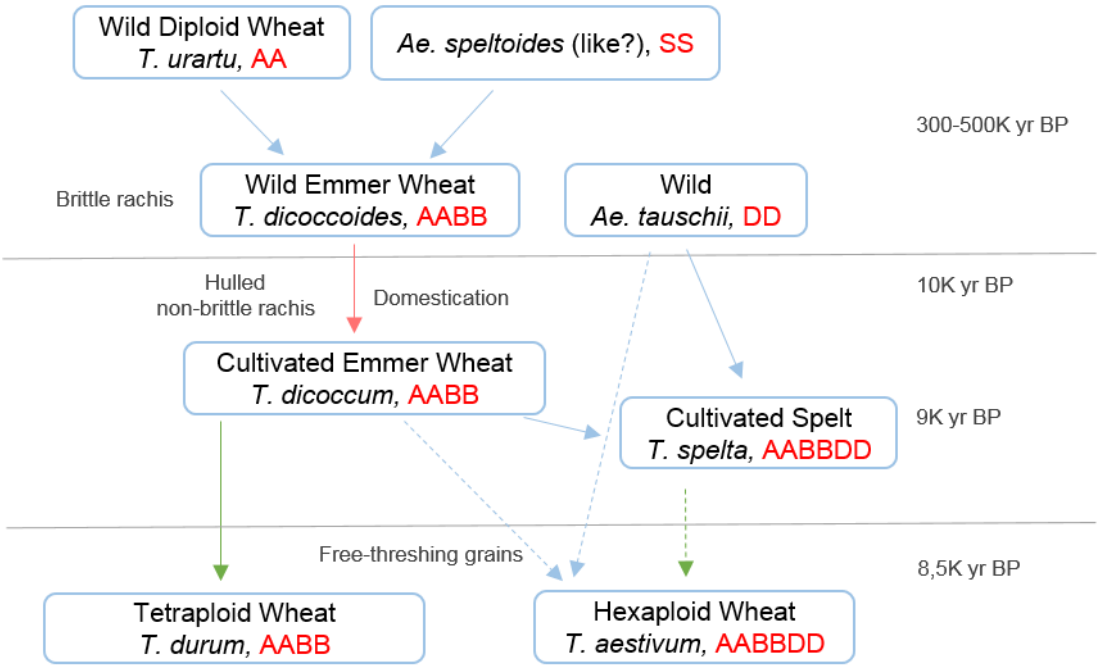



Figure 1. Overview of wheat origin and evolution. Blue arrows indicate hybridization events, red arrows indicate domestication events and green arrows indicate selection events (adapted from Salamini et al., 2002; Kilian et al., 2010; Peng et al., 2011).

It is believed that the diploid progenitors of wheat diverged from a common ancestor about 2.5-4.5 million years ago (Huang et al., 2002). The current model of wheat evolution suggests that the tetraploid *Triticum* (AB) formed about 0.5 million years ago (Salse et al., 2008). Subsequently, around 10,000 years ago, tetraploid wheat hybridized with diploid goat grass (D genome, *Aegilops tauschii*) to form the hexaploid bread wheat (Nesbitt and Samuel, 1996).

It is essential for scientific community and crop breeders to reach a pronounced knowledge on the origin of wheat and its taxonomy in order to understand better the morphological and genetic diversity (Goncharov, 2011), and foresee the success of new accessions and cultivars.

### 1.2. Economic importance and use of durum wheat

*Triticum aestivum* and *Triticum durum* are the most important among the cultivated wheats. Durum wheat (*Triticum durum* Desf.) is a minor naked crop belonging to the genus *Triticum*. It covers only 8-10% of the land dedicated to wheat cultivation and harvesting. The remaining land is dedicated to the hexaploid bread wheat cultivation. Both are economically important crops that are mainly cultivated for the market production and human consumption purposes.

Unique quality characteristics of durum wheat render it different from other existing classes of wheat. The major preeminent use of durum wheat grain is the production of semolina that is used in pasta products. However, in some countries, particularly in Morocco, traditional breads are also made using durum wheat flour. Moreover, in North Africa, the most frequently consumed types of food like couscous and bulgur are produced using durum wheat. In Latin the word "durum" stands for "hard" due to the hardest kernel among all types of wheat. Durum wheat possesses such features like high protein content and good gluten strength. These unique features make it an ideal choice for producing pasta products. The kernels of durum wheat have an amber color and usually larger than those of any other wheat classes. In addition, another superior characteristic of durum wheat is its yellow endosperm; this gives pasta its saturated golden color. Strong gluten characteristics of durum wheat allow it to form non-sticky dough ideal for pasta production purposes. Semolina that possesses strong gluten properties also allows producing pasta products with exceptional cooking characteristics. Generally there are two known sub-classes of durum wheat, which are conventional varieties with moderate gluten strength and extra-strong varieties with extra-strong gluten properties (Clarke et al., 2005).

Durum wheat is native mainly in the Mediterranean regions, Near East and Southwest Asia (Maccaferri et al., 2014). Durum wheat is a cereal crop best suited to a relatively dry climate, with warm to hot days and cool nights during the growing seasons, which is typical of the Mediterranean region. Seed germination takes place at temperatures as low as 2°C, but normally the optimal temperature is 15°C (Bozzini, 1988). Majority of durum wheat cultivars produced in the world grow in spring; however, some varieties of durum wheat lines grow in winter (Donmez et al., 2000; Schilling et al., 2003). The worldwide planted and harvested area occupied by durum wheat in 2016-17 equaled approximately to 16,1 million hectares, with an average production being about 39,9 million metric tons (International Grain Council, 2017). In 2017 according to International Grains Council European Union (25% particularly, Italy, Spain and Greece), Canada (15%), Turkey (10%), Mexico and USA (6% each), Kazakhstan, Algeria and Morocco (each 5%) were the largest durum wheat producers in the world. Leading durum wheat exporters were Canada, Mexico, EU, USA, Kazakhstan, Australia and Turkey (International Grain Council, 2017). In 2017-18, a reduction in grain production is predicted from 39.9 to 39.4 million tons for durum wheat. Moreover, there is a 20% and 17% reduction in the durum wheat planted area in Canada and USA, respectively, the two major wheat producing and exporting countries (International Grain Council, 2017). The population is expected to grow by nearly 9 billion individuals by 2050 and there is a need to meet the predicted cereal production thresholds in order to feed the demands of the increasing population.

### 1.3. Domestication of wheat

Nearly 12,000-10,000 years ago the first foundation of civilization appeared when mankind began to cultivate wheat and other staple crops. They began to shift from hunting to sedentary lifestyle leading to the formation of the agriculture-based society. It is important to understand the difference between the terms cultivation and domestication. Cultivation involves planting, growing and harvesting of either wild or domestic forms of wheat. Whereas, domestication aims to tame wild forms of wheat by means of genetic selection, i.e. modifying particular traits (Salamini et al., 2002).

Early farmers domesticated wheat from naturally occurring hybrids and, in course of time, turned them into high yielding prosperous crops that are easy to harvest. There are three common morphological differences, so called "domestication syndrome" traits (Hammer, 1984), which emerged during the transition from wild to tamed forms, all favorable to harvesting:

- Seed size - the domesticated forms of wheat have larger seeds compared to small seeds of wild forms;
- Brittle rachis – due to a tough rachis in domesticated forms the seeds are held together in a harvestable ear, whereas, the spikelet of the wild ears fall apart at maturity through fragmentation of the rachis by shattering or disarticulation (Salamini et al., 2002; Peng et al., 2011);
- Free-threshing – domesticated forms of wheat have thinner glumes that allow an easier release of the naked kernel, while in wild forms the glumes are tough tightly attached to the seed, and requires drying prior to release. (Salamini et al., 2002). Figure 2 illustrates wheat spikes showing brittle rachis and non-brittle rachis, non-free threshing and free-threshing grains from wild to domesticated forms.


Figure 2. Transition from wild emmer wheat to modern wheat species. Wheat spikes showing brittle rachis, Br (a) and non-brittle rachis (b, c, d), hulled (a, b) and naked, free-threshing (c, d) grains from tetraploid wild emmer, domesticated emmer, durum wheat and hexaploid common wheat. Material provided by Simona Corneti.

According to evolutionary history of wheat only wild einkorn and wild emmer wheats were subjected to domestication (Peng et al., 2011). The diploid einkorn wheat *T. monococcum* was one of the pioneering crops domesticated from its wild progenitor *T. boeticum* (AA). The domestication of the einkorn wheat occurred in the Fertile Crescent near the Karacadağ mountain region in Turkey (Braidwood et al., 1969; Nesbitt and Samuel, 1996; Salamini et al., 2002; Kilian, 2007; Kilian et al., 2010; Peng et al., 2011). This was identified using AFLP (Amplified Fragment Length Polymorphism) technology (Heun et al., 1997). The Fertile Crescent region (indicated by red line in Figure 3) possessed all the favorable elements for human diet such as abundant number of animals and different kinds of plants growing. Moreover, a core area (indicated by blue line in Figure 3) where the growth of several crops and legumes were intersected was found near the Karacadağ area in south-western Turkey (Lev-Yadun et al., 2000).

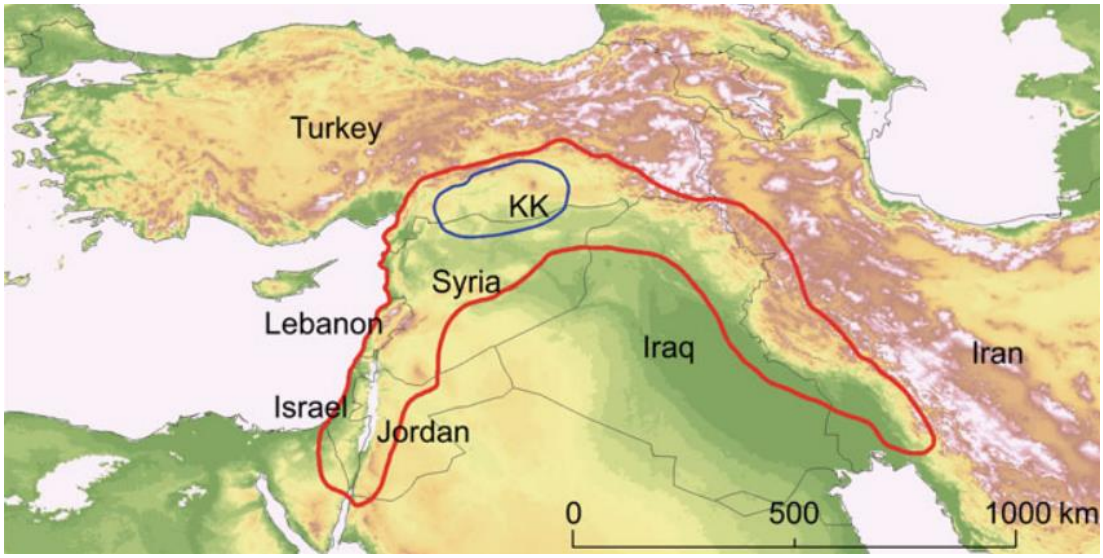



Figure 3. The Fertile Crescent (red line) and "core area" (blue line) near the Karacadağ (KK) mountain region in southwest Turkey (Kilian et al., 2010).

Although in Neolithic agriculture *T. monococcum* was one of the most important species, nowadays it is rarely cultivated and is more considered for a relict plant specific for high-value market-niches (Salamini et al., 2002). Its cultivation spread to Balkans, Cyprus, Greece, Hungary and Bulgaria before it started to drop during the Bronze Age (Salamini et al., 2002). *T. urartu* is the second wild diploid *Triticum* species that occurs partially in the Fertile Crescent but has never been domesticated, though it has a great contribution in wheat evolution as an A genome donor. The two wild species have crossing barrier and are morphologically distinguishable (Peng et al., 2011; Zohary et al., 2012).

Domestication of tetraploid emmer wheat (AABB) is a further important step in the evolution of modern polyploid wheat varieties. The tetraploid emmer wheat was domesticated from its wild progenitor *T. diccocoides*. The latter has brittle rachis; the spikelets shatter and fall apart at the maturity. Whereas, the domesticated tetraploid wheat species have non-brittle rachis, and the spikelets do not fall apart, making the harvesting process less difficult compared to its wild progenitors. Domesticated emmer wheat, *T. dicoccum* has hulled seeds. Its remains are present at Abu Hureyra from 10,400 years BP (Salamini et al., 2002). It is still cultivated in some countries like Ethiopia, Russia, Central Italy and Spain. The suggested geographical distribution of domesticated emmer includes the western parts of the Fertile Crescent, southeastern Turkey and eastern Iran and Iraq (Salamini et al., 2002; Zohary et al., 2012). Recently, Özkan *et al.* (2010) suggested a double independent cultivation of wild emmer wheat in the southern Levant and in the northern Levant

(Özkan et al., 2010). However, today among the tetraploid wheat species, *T. durum* with free-threshing seeds is the widely cultivated one.

Tetraploid *T. durum* and hexaploid *T. vulgare* are free-threshing naked wheats representing the final domestication step of the *Triticum* species. Hexaploid wheats (AABBDD) were generated as a result of hybridization between the domesticated tetraploid wheat (AABB) and a diploid *Ae. tauschii* (McFadden and Sears, 1944). However, the distribution area of *T. diccocoides* does not overlap with the distribution area of *Ae. tauschii*. Therefore, *T. turgidum dicoccum* was suspected to be involved in this hybridization that probably occurred 9,000 years ago at the southern Caspian basin, where *Ae. tauschii*, the D-genome donor grows (Salamini et al., 2002).

Killian *et al*. suggested a five step strategy approach to understand better the domestication of cereals (Kilian et al., 2010).

1. The use of an extensive and complete collection of germplasm, which covers the different distributions areas of the species, and the use of wild progenitor accessions collected from their natural habitats.
2. The comparison of a great number of wild and domesticated accessions and cultivars.
3. The recognition of the wild ancestor in the wild gene pool by genetic similarity comparison with its domesticated descendants.
4. The comparison of wild and domesticated crops by means of various molecular techniques.
5. The cloning of genes associated with domestication.

### 1.4. Wheat genome sequencing – state of the art

Unlike many other staple crops, wheat species have extensively huge and complicated genome. Therefore, sequencing of the wheat genome was one of the main challenges for scientific community. A rapid increase in the sequence output of next generation sequencing revolutionized the research community. Wheat species tend to have large genome (5-17 Gb) and a high level of repetitive content. For example, >90% of wheat genome is made by repetitive elements (Gill et al., 2004). In addition, the most recently cultivated wheat species are allopolyploids formed through inter-specific hybridization. Durum wheat is an allotetraploid grain; it contains two different genomes (AABB) and 28 chromosomes, which contain the full diploid complement of chromosomes from each of its progenitor species. This means that each chromosome pair in the A genome has a homoeologous chromosome pair in the B genome and they are closely related to each other.

The diploid progenitor of polyploid wheat varieties diverged from a common ancestor 2.5-4.5 million years ago (Huang et al., 2002). This relatively recent divergence might explain the high similarity among coding regions of wheat homoeologs (Choulet et al., 2010). The dynamic nature of wheat species should be taken into consideration as well. High level of repetitive elements in the intergenic regions causes increased rate of methylation, insertions and deletions. Subsequently, these regions in wheat tend to diversify even faster than rapidly evolving gene families (Cantu et al., 2010) and may affect the neighboring genes causing alteration in regulation, gene deletions and transpositions. Gene deletions are associated with a potential negative effect buffered by polyploidy. Gene transposition leads to higher generation of pseudogenes. Further, high divergence of alternative splicing causes even further diversification in the structure and hence structure of homoelogous genes (Paux et al., 2008; Akhunova et al., 2010; Brenchley et al., 2012). Due to a collection of all these elements, wheat is represented as one of the most challenging genomes to analyze.

The development of Next Generation Sequencing (NGS) technologies have revolutionized the understanding of crop genomes, thus expanding the opportunities for crop genetics and breeding (Edwards et al., 2013). Sequencing of wheat genome is not the central issue anymore. The main challenge has shifted towards the accurate assembly and detailed annotation of genome (Uauy, 2017). In the last 10-15 years, wheat genome sequencing has made a great progress. As the result of international collaboration of several research units, it was possible to create draft sequences of diploid, tetraploid and hexaploid wheats of unprecedented accuracy.

In 2005, The International Wheat Genome Sequencing Consortium (IWGSC) has launched a project aimed to establish a reference genome for wheat species, in particular *T. aestivum* cv. Chinese Spring. 21 chromosomes or chromosome arms of bread wheat cv. Chinese Spring were separated by flow cytometric sorting, in order to construct a high quality wheat genome assembly for each of these chromosomes using Bacterial Artificial Chromosome (BAC) libraries and physical maps (Shi and Ling, 2018). Chromosome sorting, DNA isolation and construction of BAC libraries were carried out by the Prof. J. Dolezel's group at the Institute of Experimental Botany in the Czech Republic, while the physical map construction and BAC sequencing were assigned to various groups of various institutions worldwide (Feuillet et al., 2011; Eversole et al., 2014; Shi and Ling, 2018).

In 2008 scientists from INRA (Institut National de la Recherche Agronomique, France) successfully sorted chromosome 3B and generated a physical map using BAC clones (Paux et al., 2008). A minimal tilling approach was used to select and sequence the BAC clones. The pseudomolecule of 3B had the length of 774 Mb (Paux et al., 2008). Nowadays, all 21 chromosomes of cv. Chinese Spring have been sorted and their physical maps are constructed. The physical maps

and the sequences of most of them can be consulted on IWGSC website (http://www.wheatgenome.org/Projects/IWGSC-Bread-Wheat-Projects). In 2012, using the Whole Genome Shotgun Sequencing with 454 pyrosequencing, Illumina and SOLiD methods, *T. aestivum* cv. Chinese Spring, *T. monococcum* accession 4342-96 and *Ae. tauschii* ssp *strangulata* accession AL8/78 were sequenced. The Chinese Spring assembly was 5.42 Gb in length with 96,000 predicted genes. According to authors, they observed a reduced number of gene families present in bread wheat in comparison to their diploid progenitors (Brenchley et al., 2012). Two years later, the first chromosome based draft sequence of bread wheat cv. Chinese Spring was published; 10.2 Gb of genome was sequenced using flow cytometry and Illumina sequencing technology. The work highlighted the high conservation of gene copies present in chromosomes, dynamic nature of common wheat with frequent gene losses and insertion duplication events (International Wheat Genome Sequencing Consortium, 2014). An improved genome of Chinese Spring was published in 2017 by Earlham Institute (UK), often known to wheat community as TGAC annotation (Clavijo et al., 2017). In this annotation, 104,091 High Confidence (HC) genes were identified by combining mate-pair Illumina RNA-seq and PacBio (Pacific Biosciences) full-length cDNA sequences. Several genome rearrangements were identified and confirmed. This improved assembly represented >78% of the whole genome, much higher compared to 49% assembled by IWGSC in 2014 (Clavijo et al., 2017; Shi and Ling, 2018). Zimin et al. (Zimin *et al.*, 2017) reported a more complete genome assembly of the hexaploid wheat. A combination of Illumina reads and long PacBio reads were used to produce the assembly. The authors used MaSuRCA (Zimin et al., 2013) and FALCON (Chin et al., 2016) algorithms in order to merge the sequences. Apart from Chinese Spring, four additional varieties of bread wheat Robigus, Paragon, Claire, Cadenza and one durum wheat variety Kronos have been released (http://www.earlham.ac.uk/grassroots-genomics) (Uauy, 2017).

A major breakthrough was made in sequencing of crops with the advent of a new software package called DenovoMAGIC2 (NRGene, NesZiona, Israel) (Avni et al., 2017; Mascher et al., 2017; International Wheat Genome Sequencing Consortium et al., 2018, Maccaferri et al., under revision). It aims to perform the scaffold assembly from short Illumina sequencing reads. The method relies on a novel 3D chromosome-conformation capture coupled with high-throughput sequencing (Hi-C) data. The software reached a success in crop sequencing at an unprecedented level. Recently, IWGSC announced the completion of a high quality sequence of hexaploid wheat cv. Chinese Spring (International Wheat Genome Sequencing Consortium, 2018) and released the genomic data as well as annotation for public access (http://www.wheatgenome.org/News/Latest-news/RefSeq-v1.0-URGI). In addition to Chinese Spring several other crops were sequenced utilizing DenovoMAGIC2 package. The approach was already adopted to generate high-quality reference genome sequences of the barley cultivar Morex (Beier et al., 2017; Mascher et al., 2017), wild emmer wheat *T. turgidum*

ssp. *dicoccoides* accession Zavitan (Avni et al., 2017) and *Ae. tauschii* ssp. *strangulata* accession AL8/78 (Luo et al., 2017; Zhao et al., 2017). Additionally, it was used to assembly the tetraploid *T. durum* cultivar Svevo (Maccaferri et al., unpublished). Table 1 shows a detailed summary progress of wheat genome sequencing from 2008 to 2018.

It is worth mentioning one of the pioneering reference transcriptome sequences of durum wheat. The assembly of tetraploid wheat cultivar Kronos was constructed *de novo* in 2013 (Krasileva et al., 2013) and since then was widely used by the scientists. Since wheat gene coding regions correspond only to 1-2% of the total genome, a *de novo* transcriptome assembly demonstrated that the expressed region of a genome could be effective and sufficient for some research purposes, while escaping the technical problems associated to the assembly of highly repetitive intergenic regions. Nevertheless, knowledge of the whole genome sequence is highly desirable.

Table 1. Progress in wheat genome sequencing.

| **Genome** | **Sequencing Strategy/Method** | **Reference** |
|---|---|---|
| *T. aestivum* cv. Chinese Spring 3B chromosome | BAC-by-BAC | (Paux et al., 2008) |
| *T. aestivum* cv. Chinese Spring, *T. monococcum* accession 4342-96, *Ae. tauschii* ssp *strangulata* accession AL8/78 | Roche 454 pyrosequencing on the GS FLX Titanium and GS FLX+ platforms, Illumina, SOLiD | (Brenchley et al., 2012) |
| *T. aestivum* cv. Chinese Spring | Flow sorted chromosome arms using Illumina HiSeq 2000 or Genome Analyser Iix. | (Eversole et al., 2014; International Wheat Genome Sequencing Consortium (IWGSC), 2014) |
| *T. aestivum* cv. Chinese Spring 3B chromosome | BAC, Roche 454 pyrosequencing on the GS FLX Titanium | (Choulet et al., 2014) |
| *T. aestivum* Synthetic W7984 and Opata M85, 90 doubled haploid (DH) lines derived from W7984/Opata $F_1$ hybrids; the 'SynOpDH' | Illumina HiSeq 2000 | (Chapman et al., 2015) |
| *T. aestivum* cv. Chinese Spring | Illumina HiSeq 2500 | (Clavijo et al., 2017) |
| *T. aestivum* cv. Chinese Spring | Illumina + Pacific Biosciences | (Zimin et al., 2017a) |
| *Ae.tauschii* ssp. *strangulata* accession AL8/78 | -WGS Pacific Biosciences mega-reads + optical BioNano<br>-Illumina and PacBio sequences | (Luo et al., 2017)<br>(Zimin et al., 2017b) |

| | | |
|---|---|---|
| Wild Emmer Wheat *T. turgidum* ssp. *dicoccoides*, accession Zavitan | Illumina HiSeq2500 | (Avni et al., 2017) |
| *T. urartu* (Tu) accession G1812 (PI428198) | BAC-by-BAC sequencing, single molecule real-time whole-genome shotgun sequencing, linked reads and optical mapping | (Ling et al., 2018) |
| *T. aestivum* cv. Chinese Spring | Physical maps for all chromosomes, BAC libraries, BioNano optical maps, RH maps, GBS maps | (International Wheat Genome Sequencing Consortium (IWGSC) et al., 2018) |
| *T. durum* ssp. *durum* cv. Svevo | Illumina HiSeq2500 | (Maccaferri *et al.,* unpublished) |

### 1.5. Durum wheat variety Svevo

"... a unique grain in the world" – this is how in 2005 the durum wheat variety Svevo has been described in one of the advertisements for a well-known italian brand of pasta Barilla. The variety Svevo is the result of collaboration between Bologna Seed Production Company and Barilla. It was selected for excellent qualitative properties such as protein accumulation and the high index of yellow semolina. Svevo has a good production potential determined by the three components of production: the fertility of the spike, weigth of 1000 seeds and number of ears per square meter. The grain is distinguished by it's exceptional quality features, excellent semolina color index and the extraordinary aptitude to industrial processing (Società Produttori Sementi Spa, 2012). Moreover, Svevo is being widely used in breeding programs. Table 2 below summarizes the main characteristics of durum wheat variety Svevo.

Table 2. Characteristics of *T. durum* variety Svevo.

| **Variety characteristics** | |
|---|---|
| Pedigree | Line CIMMYT/Zenit |
| Release date | 1996 |
| **Plant characteristics** | |
| Seasonal type | spring |
| Heading time | very early |
| Height | medium high |
| Awns color | brown |
| Potential yield | medium high |
| **Grain quality** | |
| Test weight | good |
| Yellow index (“b” Minolta) | high (24-26) |
| Protein content | very high |
| Gluten quality (scale 1-10) | good (5) |
| **Resistance to:** | |
| Powdery mildew | good |
| Leaf rust | medium |
| Septoria | medium |
| Cold | medium |

## 2. Objectives

Durum wheat cultivar Svevo has been a quality and productivity durum variety in Italy for more than a decade. Genome assembly, gene prediction and annotation of durum wheat is a valuable resource for researchers and breeders. Therefore, the International Durum Wheat Genome Sequencing Consortium assembled a high-quality draft genome sequence of the durum wheat cultivar Svevo. The assembly resulted in a set of 14 pseudomolecules of 9.96 Gb size. The availability of wild emmer and durum wheat genomes allow us to understand better the evolution and domestication of tetraploid wheat. We used a Global Tetraploid wheat germplasm Collection (GTC), composed of 1,854 accessions of up to ten different species and subspecies from a wide range of areas. These accessions represent the four principal germplasm groups that are involved in the history of tetraploid domestication and selection processes: Wild Emmer Wheat - WEW, Domesticated Emmer Wheat - DEW, Durum Wheat Landraces - DWL and Durum Wheat Cultivars - DWC. The work herein described had the following objectives:

- Predict the gene models of the durum wheat cultivar Svevo assembly
- Assess the evidence and the pattern of gene expression in high-confidence genes
- Predict and compare the NB-LRR-encoding loci in durum wheat and wild emmer wheat
- Identify the population structure of the Global Tetraploid wheat Collection
- Detect the selection signatures and diversity reduction in tetraploid wheat germplasm from wild emmer to modern durum wheat

# 3. Background

With the effort of The International Durum Wheat Genome Sequencing Consortium the genome of durum wheat cultivar Svevo (release 1996 CIMMYT line/Zenit) was sequenced and *de novo* assembled using the protocols described for wild emmer wheat (Avni et al., 2017) and the major findings are illustrated in Figure 4. The consortium involved diverse research groups, such as CREA (Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria, Italy), CNR (Consiglio Nazionale delle Ricerche, Italy), University of Bologna (Italy), IPK Gatersleben (Germany), University of Saskatchewan (Canada), Helmholtz Center Munich (Germany), University of Tel-Aviv (Israel), Montana State University (USA), AgriBio Centre for AgriBioscience (Australia), USDA U.S Department of Agriculture (USA). The scaffold assembling was performed using DenovoMAGIC2 (NRGene, NesZiona, Israel) using a novel 3D chromosome-conformation capture coupled with high-throughput sequencing (Hi-C) data. A Svevo × Zavitan genetic map was used to order and orient the scaffolds (Avni et al., 2014). The assembly resulted in a set of 14 pseudomolecules of 9.96 Gb size.

In this study, the gene models of Wild Emmer Wheat (WEW) accession Zavitan were annotated as well using the same pipeline and annotation data sources, with the purpose to compare the divergence between the wild and domesticated wheat genomes. As a result, WEW had 67,182 High Confidence (HC) genes and 271,179 Low Confidence (LC) genes.

In addition to this comparison, a Global Tetraploid Collection (GTC) of wheat consisting of 1,856 accessions that represent the four major domestication, breeding related and diverse geographic region germplasm groups (Wild Emmer Wheat-WEW, Domesticated Emmer Wheat-DEW, Durum Wheat Landraces-DWL, Durum Wheat Cultivars-DWC) were genotyped using the wheat iSelect 90K SNP Infinium assay (Wang et al., 2014). A set of 17,340 informative SNPs and 5,774 SNPs with $r^2$=0.5 were used for genetic diversity, population structure analysis and for selection signature identification.

Therefore, the material presented in this thesis makes part of a joint project produced and coordinated by several international units, which aimed to decipher the genome of durum wheat.

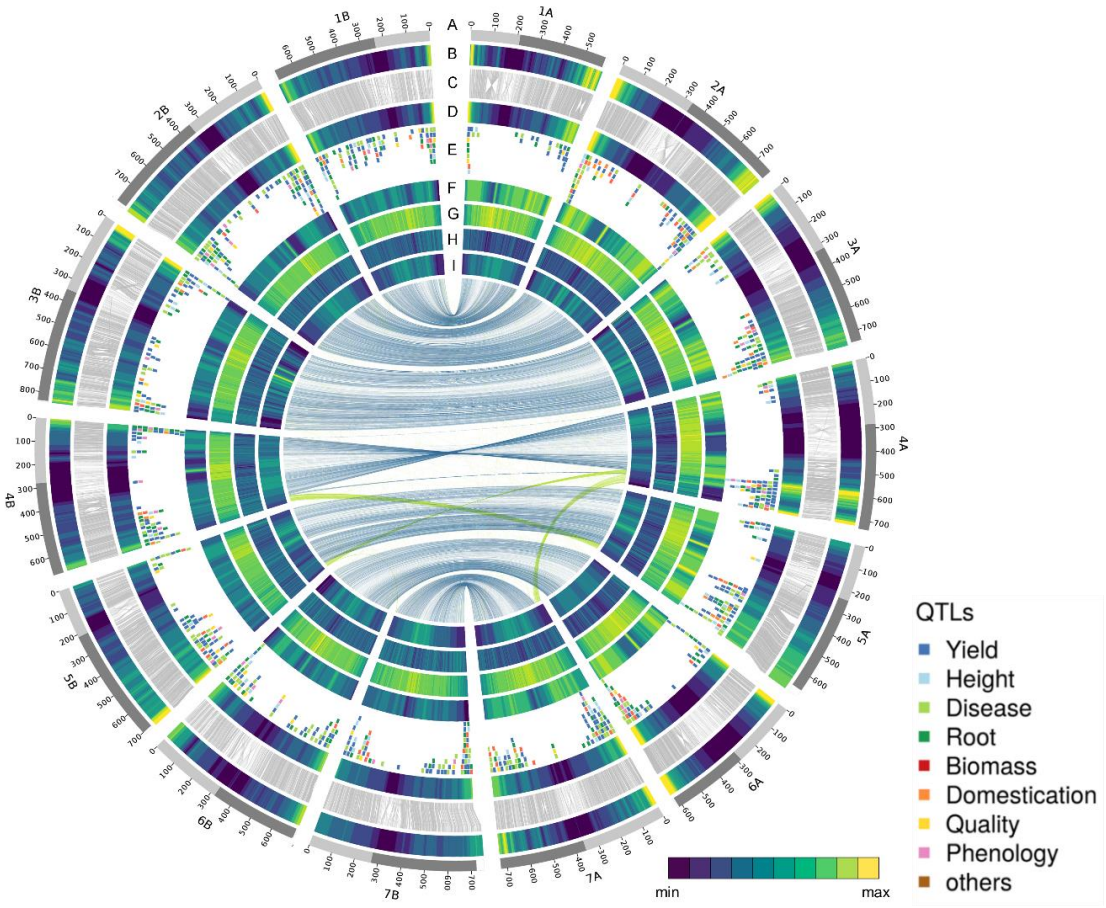



Figure 4. Structural, functional, and conserved synteny landscape of the durum wheat genome. Tracks from outside to inside: **(A)** Chromosome number and size (100 Mbp tick size); **(B)** Density of WEW high confidence gene models (HC; 0 to 25 genes per Mb); **(C)** Links connecting homologous genes between WEW and DW; **(D**) Density of DW high confidence gene models (HC; 0 to 22 genes per Mb); **(E)** Locations of published QTLs (QTL peak positions reported); **(F)** *K-mer* frequencies (2,400-4,700 per Mb); **(G)** LTR-retrotransposons density (0-95 per Mb); **(H)** DNA transposons frequency (0-35 per Mb) ; **(I)** Mean expression of HC genes ([log(FPKM+1)], mean expression value at all conditions, ranges from 1.6 to 8.2); FPKM, Fragments per kilobase-Million. Chromosomal cross-links in center connect DW homoeologous genes between subgenomes; blue links, connections between homoeologous chromosomes; green links, large translocated regions.

# 4. Chapter 1. Gene content and pattern of gene expression in durum wheat

## 4.1. Materials and Methods

### 4.1.1. Annotation of protein-coding genes

Annotation pipeline data sources

The gene annotation pipeline combined evidences from protein reference sequences and gene expression data to predict transcript sequences on the genome assembly. A wide range of RNA-seq libraries, full-length transcript sequences as well as protein sequences and several publicly available datasets of wheat were included in the pipeline. Subsequently, Open Reading Frames (ORFs) were predicted based of the transcript structures. Finally, a confidence classification of the predicted genes was performed in order to distinguish sets of high confidence genes (described below).

Plant material and RNA extraction: RNA-seq of nine pools from cultivar Svevo

Fifty-seven different tissue and treatment combinations (abiotic and biotic stresses, nutrients, hormones, heavy metals and various organs) of cultivar Svevo were collected and stored at -80°C. RNAs for most of the tissues were extracted using Direct-zol TM RNA miniprep kit (Zymo Research), according to manufacturer's protocol. A modification was made for the samples of grain and anthers and ovaries, where a pre-process of Trizol treatment was applied to reduce the amount of carbohydrates in the sample. RNA integrity was assessed using capillary electrophoresis on a Fragment Analyzer™ system (Advanced Analytical Technologies, Inc.) and analysed using the PROSize™ data analysis software. Samples with RIN number below six were discarded and agarose gel electrophoresis was used to confirm RNA size and integrity. Sequencing (two lanes, paired-end, 2 x 126 cycles) was done at the Crown Institute for Genomics at the Weizmann institute, Rehovot, using an Illumina HiSeq2500 instrument, according to manufacturer's instructions. The RNA-seq samples from these different tissues and treatments were then gathered in nine mega-pools.

RNA-seq samples from cultivar Svevo

RNA-seq samples of cv. Svevo were extracted from four different combinations of tissues: (i) coleoptile and leaves at the seedling stage, (ii) apex of seminal roots at the seedling stage, (iii) ovaries and anthers at the beginning of anthesis (Zadoks 60) and (iv) developing grains at the growth stage Z70 (ZADOKS et al., 1974). Total RNA was isolated from 100 mg using Total Spectrum Plant RNA (Sigma-Aldrich) according to the manufacturer's instructions. RNA was quality checked using the RNA 6000 Nano kit on a 2100 Bioanalyzer (Agilent) and quantified on Nanodrop spectrophotometer (Thermo Scientific). The library preparation and Illumina sequencing is as follows: for each tissue

derived from Svevo, two cDNA libraries (eight in total) were constructed from 4 μg total RNA using the Illumina TruSeq RNA Sample preparation kit, according to the manufacturer's protocol. After PCR enrichment, cDNA fragments were separated on a 2% low agarose 1X TAE gel; two size fractions, one with an average fragment size of 280 bp and another with a fragment size ranging from 380 to 480 bp were extracted and then purified using the Gel- Extraction kit (Qiagen). The two libraries from each tissue were pooled together and quantified with Bioanalyzer using High Sensitivity kit (Agilent). The libraries were loaded as a pool on a Cluster Generation machine in a single Illumina flow cell lane following the standard Illumina protocol. A paired-end sequencing protocol was conducted with the Illumina GAIIx generating 150 bp reads in pairs.

Thirteen durum wheat varieties

Thirteen durum wheat varieties representing the worldwide durum wheat elite germplasm (*Triticum durum*) including cv. Svevo (*Triticum turgidum* subsp*. durum* Desf.) were grown in growth chamber under the optimal conditions for wheat, long days 16/8 hours day/night photoperiod regime at 20/16 °C day/night temperature regime. The 13 varieties (described in Table 4) that span the breeding era from 1940 to 2005 and include the diverse germplasm, were chosen to produce RNA-seq from leaf, root and grain tissues. The libraries were prepared as previously described for Svevo, with gel size selection of 500-600 bp fragments. The purified libraries were used to produce two pools, 6-plex and 7-plex, and after quantification on the Bioanalyzer were loaded on two lanes of a HiSeq2000 sequencer run. Sequence reads were produced with the standard Illumina pipeline.

RNA-seq for grains at six levels of development cv. Svevo and Senatore Cappelli

Caryopses from two durum wheat cultivars Svevo and Senatore Cappelli were collected at six different developmental stages (3, 5, 11, 16, 21 and 30 days after anthesis) and used for RNA extraction. Libraries were prepared with the Illumina TruSeq Stranded Total RNA with Ribo-Zero Plant sample prep kit (Illumina Inc., San Diego, CA) following manufacture's protocol. The prepared indexed libraries were evaluated with the High sensitivity D1000 screen Tape (Agilent Tape Station 2200), then quantified with ABI9700 qPCR instrument using the KAPA Library Quantification Kit (Kapa Biosystems, Woburn, MA, USA) and sequenced on the Hiseq2000 with a 100 cycles of paired-end sequencing module using the 164 Truseq SBS kit v3.

The list of the datasets used for gene prediction is summarised in Table 3.

RNA-seq from public literature

In addition, we used several publicly available RNA-seq data sets.

- Wild Emmer Wheat (*Triticum diccocoides*), 20 RNA-seq samples from different tissues and developmental stages (Avni et al., 2017);
- Bread wheat cv. Chinese Spring (*Triticum aestivum*), 30 RNA-seq samples from five tissues (grain, leaf, root, spike, stem) and three developmental stages (Pingault et al., 2015);
- Durum wheat cv. Kronos (*Triticum turgidum*), RNA-seq samples from young roots, young shoots, spike (Krasileva et al., 2013);
- RNA-seq extracted from glumes of two wild emmer wheats, two landraces and two durum wheat cultivars (Zou et al., 2015);
- Bread wheat cv. Chinese Spring, Illumina RNA-seq reads from leaves, roots, seedling, seed, stem and spike (Clavijo et al., 2017).

Full-length transcript sequences

Moreover, several full-length cDNA sequences were used in the annotation pipeline. We used Pacific Biosciences (PacBio) full-length cDNA sequences of bread wheat cv. Chinese Spring from five different bread wheat tissues such as leaves, roots, seedling, seed, stem and spike (Clavijo et al., 2017). Also full-length coding sequences (CDSs) of the *Triticeae* crops from the *Triticeae* full length cDNA sequences TriFLDB database were included in the gene annotation pipeline (Mochida et al., 2009).

Gene annotation pipeline

We used the gene annotation pipeline developed by the Plant Genome and Systems Biology (PGSB) group at the Helmholtz Center Munich, Germany (Avni et al., 2017; Mascher et al., 2017; International Wheat Genome Sequencing Consortium et al., 2018). Figure 5 illustrates the details of the pipeline.

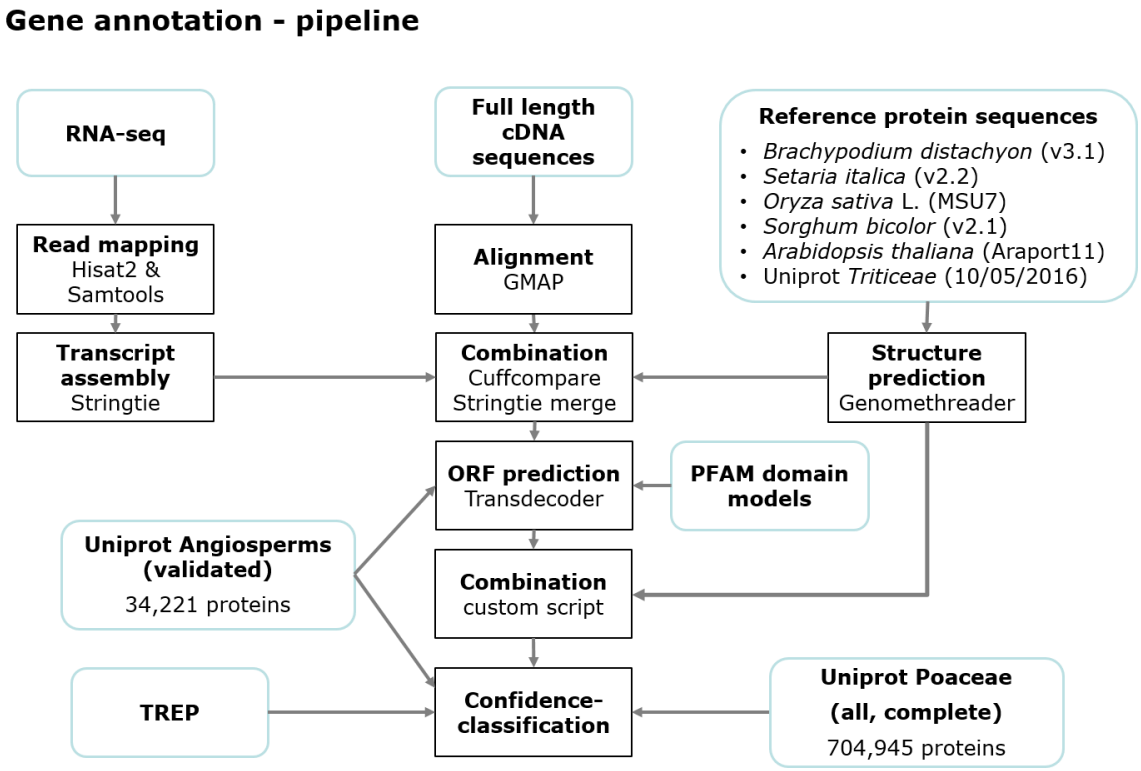



Figure 5. Gene structure prediction pipeline developed by PGSB group in Germany.

Alignment of reference protein sequences

The spliced alignment tool Genomethreader version 1.6.6 (Gremme et al., 2005) was used to align protein sequences from the following related grass species to the Svevo genome assembly:

- *Brachypodium distachyon* (The International Brachypodium Initiative, 2010)
- *Setaria italica* (Bennetzen et al., 2012)
- *Oryza sativa* L. (Ouyang et al., 2007)
- *Sorghum bicolor* (Paterson et al., 2009)
- *Arabidopsis thaliana* (Krishnakumar et al., 2015)
- All annotated protein sequences from the *Triticeae* tribe.

The *Triticeae* protein sequences were downloaded from UniProt database released on 10/05/2016 (The UniProt Consortium, 2015). Then, from these sequences, we filtered the ones that have been marked as complete protein sequences. Subsequently, we clustered these sequences by 100% identity. This set included validated protein sequences from SwissProt as well as predicted protein sequences from species that included *Triticum aestivum*, *Aegilops tauschii* and *Hordeum vulgare*.

Genomethreader is a software used to predict the gene structure based on similarities between the protein sequences via spliced alignments. We applied the Genomethreader on each pseudomolecule sequence separately in order to reduce the memory requirement per application. The parameters used were as follows: -startcodon –stopcodon -species rice –gcmincoverage 70 –prseedlength 7 –prhdist 4 –gff3out. Where,

-startcodon, require than an ORF must begin with a start codon

-stopcodon, require that the final ORF must end with a stop codon

-species, specify species to select splice site model

-gcmincoverage, the minimum coverage of global chains

-prseedlength, the length m of the exact seeds used for protein matching

-prhdist, the maximum Hamming distance h a protein match is allowed to have

- gff3out, show output in GFF3 format.

Overall, alignment of protein sequences on the durum wheat assembly predicted 266,429 potential gene loci.

Transcriptomic evidences

We used HISAT2 version 2.0.4 (Pertea et al., 2016) to align all sets of RNA-seq libraries to the Svevo genome assembly (parameter: --dta). HISAT2 (hierarchical indexing for spliced alignment of transcripts) is a freely available software that aligns reads to a genome and detects the transcript splice sites. Moreover, in comparison to other existing RNA-seq alignment methods, HISAT2 shows accurate results, has a faster and better performance compared to other alignment tools and uses less memory (Pertea et al., 2016).

Then, the mapped reads were assembled into transcript sequences separately using StringTie (Pertea et al., 2016). StringTie aims to assemble the alignments into transcripts, by creating isoforms and estimating the abundance of these isoforms. We configured the StringTie with the following parameters: -m 150 –t –f 0.3, which means that we set a minimum length of 150 bp for the transcript sequences and a minimum fraction of 0.3 for predicted isoform abundance of the transcript over the most abundant transcript.

Alignment of full-length transcript sequences

The full length cDNA sequences from public databases as well as publicly available bread wheat IsoSeq sequences from six different tissues (leaf, root, seedling, seed, spike and stem) (Clavijo et al., 2017) were aligned to the reference genome using GMAP version 06/30/2016 (Wu and Watanabe, 2005; Wu et al., 2016). GMAP aims to map and align the large number of cDNA sequences to a genome with minimal memory requirements. The program does not use probabilistic splice site models and neglects the sequencing errors and various polymorphism sites (Wu and Watanabe, 2005). We applied GMAP using the following parameter - K 50000, in order to align all sequences to the assembly with the restricted maximum intron size of 50,000 bp.

Combination of evidences and prediction of open reading frames

All transcript predictions from the above described evidences were combined using Cuffcompare from Cufflinks software suite (Trapnell et al., 2010, 2012). We used *StringTie* with the *merge* parameter (--merge -m 150) in order to combine the overlapping transcript sequences and hence, remove the redundant transcripts and fragments. Then, from the resulting GTF output files (genome-based transcript structure files), we extracted the transcript sequences using *cufflinks_gtf_to_cdns_fasta.pl* script from the TransdeCoder package version 3.0.0 (https://github.com/TransDecoder/TransDecoder). TransdeCoder is used to predict and find the coding regions within the transcripts (Haas et al., 2013). We then used TransDecoder.LongOrfs (parameter: -p 0) to extract the longest open reading frames for each transcript sequences and to translate them into predicted protein sequences. We then compared these potential protein sequences to the reference protein database using BLASTP (version NCBI blast 2.3.0+, parameter: -max_target_seqs 1, -evalue 1e-05) and checked for the abundance of the known protein domains using Hmmscan version 3.1b2 (http://hmmer.org/, Durbin, 1998). The resulting two tables of output were used as queries into TransDecoder.Predict in order to select a single best open reading frame for each transcript structure. The final gene predictions were combined with the protein structure predictions from Genomethreader to compensate for potentially differentiating open reading frame predictions by the two tools.

Confidence classification

We applied a confidence classification to all predicted protein/transcript sequences in order to differentiate the sequences into (i) canonical proteins, (ii) non-coding transcripts, (iii) incomplete genes and (iv) transposable elements. Therefore, we aligned all potential protein sequences using BLAST against two reference protein databases. The first database consisted of all validated *Magnoliophyta* protein sequences from Uniprot (UniMag) and the second database contained all

annotated *Poaceae* protein sequences from Uniprot (UniPoa) (downloaded on 03/08/2016). Non-complete protein sequences were filtered from the second database. Furthermore, in order to detect and filter out the transposons, we aligned all potential protein sequences using BLAST against the translated TREP database (release 16, http://botserv2.uzh.ch/kelldata/trep-db/index.html, Sabot et al., 2005).

Based on the E-value distribution of best hits, those with an E-value below $10^{-10}$ were considered as significant hits. In order to avoid the fragmented alignments, which might be present due to fragmented protein annotations or local alignments of domains, we applied the thresholds for the significant alignments based on query and subject coverages. For the comparison with the protein databases, only alignments with query and subject coverage of at least 90% were considered as representative hits, and for the comparison with the TREP database, the alignments with a query coverage of at least 75% were considered as representative hits. Based on the representative BLAST hits and the completeness of protein sequences (with annotated start and stop codons), all potential transcript sequences were then classified into two confidence classes and five subclasses:

- High confidence (HC) transcripts: Coding sequence with annotated start and stop codon and representative hit to reference protein sequence (query coverage >90% and subject coverage >90% and E-value $<10^{-10}$);
- HC1: hit to validated protein sequence (*Magnoliophyta*);
- HC2: hit to predicted protein sequence (*Poaceae*);
- Low confidence (LC) transcripts: Coding sequences that were annotated as incomplete or that showed only insufficient homology to reference proteins or were possible candidates for transposons;
- LC1: incomplete coding sequence, but significant match to reference protein sequence;
- LC2: complete coding sequence, but no significant match to reference protein;
- REP: match to transposon elements database.

The loci that contained at least one high confidence transcript was considered as a high confidence gene. All low confidence transcripts that were overlapping with the high confidence transcripts were removed in order to prevent merging of neighboring loci by low confidence transcripts.

Validation of the genome assembly and annotation

In order to evaluate the completeness of the genome assembly and to determine the quality of the annotation, we performed a double validation step procedure. Firstly, we used the BUSCO tool

(Benchmarking Universal Single-Copy Orthologs, version 2, Embyophyta odb9) to determine the abundance of strongly conserved genes in all annotated gene sets and HC gene sets (Simão et al., 2015). In addition, the predicted gene models were verified using 216 experimentally validated complete gene sequences kindly provided by Jorge Dubcovsky (University of California, Davis, CA). We used these sequences as queries in a BLASTX (version ncbi-blast-2.2.26+) search against the whole set of proteins (LC and HC).

Furthermore, to validate the predicted protein sequences, we downloaded all available *Triticeae* protein sequences from the Uniprot database (downloaded on 27/04/2017), filtered for sequences that were marked as complete and clustered sequences by 100% sequence identity. This procedure has identified a set of 204,773 unique reference protein sequences.

### 4.1.2. Pattern of gene expression

We were interested to investigate the expression pattern of the predicted and validated high confidence genes. For these purposes, we mapped RNA-seq libraries to the Svevo genome assembly. We used different RNA-seq libraries for different gene expression analyses purposes:

- Sixteen RNA-seq libraries of cv. Svevo only (9 library pools, grain, leaf, root, anther_ovaries, seed_milk and grain at 6 developmental stages) were used to investigate the number of HC genes expressed;
- 57 RNA-seq libraries (13 varieties and 3 tissues, grain at 6 developmental stages for two varieties Svevo and Senatore Cappelli, and additional tissues for Svevo; leafa, roota, anther_ovaries a-b, seed_milk a-b) were mapped to Svevo assembly, in order to study the expression of genes in specific tissues, and the mean expression (or expression density) variation along the chromosomes;
- RNA-seq libraries for 13 durum wheat varieties were used in order to study the expression profiles of durum wheat elite varieties and understand the variation between both the varieties and tissues.

We used HISAT2 version 2.0.4 (Pertea et al., 2016) to map the RNA-seq libraries to the assembly. The resulting output files in SAM format were sorted and converted to BAM files using samtools (Li et al., 2009; Li, 2011). Then, the BAM files were filtered for reads that aligned concordantly exactly 1 times based on the mapping quality > 40. The replicates, where applicable, were merged prior to read count. Subsequently, the transcript abundance was estimated by running the StringTie with –eB options (Pertea et al., 2016) and read count matrix was generated using a

provided python script *prepDe.py* (http://ccb.jhu.edu/software/stringtie/dl/prepDE.py). The gene expression was quantified by FPKM (fragments per kilobase of exon per million fragments mapped). The genes that were not expressed at all were consequently removed. Finally, the expression matrix was log normalized in R (R Core Team, 2013) using a Bioconductor package *DESeq2* (Love et al., 2014) and quantified as log(FPKM+1).

The normalized expression matrix was further used for clustering and variance analysis. We performed various clustering analyses:

- hierarchical cluster analysis based on dissimilarities using R function *hclust* (method = "complete") (Murtagh, 1985; R Core Team, 2013);
- heatmap clustering based on sample to sample distances using *pheatmap* function (Kolde, 2018);
- two-dimensional and three-dimensional Principal Component Analysis (PCA) using *prcomp* function in R *stats* package (R Core Team, 2013).

We investigated the number of genes expressed in each subgenome (A and B) and the difference in number of genes expressed in different libraries. Additionally, we investigated the mean expression level per gene (mean expression value across all 57 samples) along each chromosome, mean expression density of genes (number of libraries in which gene was expressed; from 1 to 57), and number of genes that are expressed using a library of 57 samples.

We attempted to analyze the major variation pattern within the tissues and varieties. Based on the last we performed hierarchical clustering analysis on the strongest PC scores to identify genes that are highly or lowly expressed in particular tissues for different varieties. In order to do this, we used sparse PCA analysis (Zou et al., 2006) using the *SPC* function of R package *PMA* (Witten et al., 2009). Sparse PCA zeroes out irrelevant features from PC loadings. The advantage is that we can find important features that contribute to major variation patterns. Additionally, using the R *stats* function *var* (R Core Team, 2013) we performed a variance expression analysis between the varieties for each tissue projected on the Svevo assembly, which allowed us to identify the chromosome regions that drove the major expression variation patterns.

### 4.1.3. NB-LRR gene family organization in durum and wild emmer wheat

NB-LRRs are plant disease resistance genes containing nucleotide-binding leucine-rich repeat domains. They form one of the biggest gene families in plants and have an important role in

plant resistance mechanisms and plant innate immune systems (Jupe et al., 2012; Marone et al., 2013; Bouktila et al., 2015; Lee and Yeom, 2015).

In this study, we used an NLR-annotator version 0.7 pipeline kindly provided by B. Steuernagel (John Innes Centre, UK) (https://github.com/steuernb/NLR-Annotator) to annotate the loci associated with NLRs both in durum wheat (DW) cv. Svevo and wild emmer wheat (WEW) accession Zavitan. NLR-annotator predicts the NLR loci by searching for amino acid motifs associated with the NLRs within six open reading frames. NLR-loci is defined by the first and last motifs associated to NLR and does not predict the NLR genes. In order to identify the loci potentially encoding NLR genes the pseudomolecules were,

- First fragmented in 20 kb segments overlapping by 5 kb;
- Next, the NLR-associated amino acid motifs were searched within all six frame translated amino acid sequences using the NLR-parser (Steuernagel et al., 2015);
- Finally, the NLR-annotator generates information on predicted NLR loci, aligned motifs, domains, whether these loci are potentially complete, partial or pseudogenes.

Further, the NLRs were compared to their corresponding RNA-seq based gene models using Cuffcompare (Trapnell et al., 2010, 2012) in order to identify possible novel loci not present in the transcriptome-based annotations.

Table 3**.** RNA-seq libraries used in the gene prediction pipeline.

| **Origin** | **File in public repository** | **Tissue** | **Reads** | **Reference** |
|---|---|---|---|---|
| *T. durum* cv. Svevo | This study | 57 different treatments at seedling and adult plants organized in 9 Pool of RNA samples | 2.8 billion reads | This study |
| *T. durum* cv. Svevo | This study | Grain at 6 developmental stages | 1.7 billion reads | This study |
| *T. durum* cv. Senatore Cappelli | This study | Grain at 6 developmental stages | 1.4 billion reads | This study |
| *T. aestivum* cv. Chinese Spring | The Illumina and PacBio reads are available at study accession PRJEB15048 at EMBL-EBI European | Leaf; Root; Seedling; Seed; Stem; Spike | 3 billion reads | (Pingault et al., 2015) |

| | | | | |
|---|---|---|---|---|
| | Nucleotide Archive | | | |
| *T. durum* cv. Kronos | Bioproject PRJNA191054 for T. turgidum. Raw data is available at the Short Read Archive (accession numbers: SRR769749, SRR769750, SRR863375, SRR863376, SRR863394 SRR863377,SRR863384, SRR863385, SRR863386, SRR863387, SRR863389, SRR863390, SRR863391) | Young roots; young shoots; spike; grain | 0.5 billion reads | (Krasileva et al., 2013) |
| *T. aestivum* cv. Chinese Spring | RNA-Seq data have been deposited under accession number ERP004714 | Grain; leaf; root; spike; stem | 2 billion reads | (Clavijo et al., 2017) |
| *T. durum* cv. Altar84 | This study | grain; root; leaf | 117 million reads | |
| *T. durum* cv. Capeiti8 | This study | grain; root; leaf | 190 million reads | |
| *T. durum* cv. Claudio | This study | grain; root; leaf | 156 million reads | |
| *T. durum* cv. Creso | This study | grain; root; leaf | 174 million reads | |
| *T. durum* cv. Edmore | This study | grain; root; leaf | 195 million reads | |
| *T. durum.* Cv. Kofa | This study | grain; root; leaf | 160 million reads | |
| *T. durum* cv. Meridiano | This study | grain; root; leaf | 250 million reads | |
| *T. durum* cv. Neodur | This study | grain; root; leaf | 211 million reads | |
| *T. durum* cv. Saragolla | This study | grain; root; leaf | 178 million reads | |
| *T. durum* cv. | This study | grain; root; leaf | 161 million | |

| | | | | |
|---|---|---|---|---|
| Strongfield | | | reads | |
| *T. durum* cv. Valnova | This study | grain; root; leaf | 200 million reads | |
| *T. durum* cv. Yavaros79 | This study | grain; root; leaf | 133 million reads | |
| *T. durum* cv. Svevo | This study | grain; root; leaf; seed_anthesis; seed_milk | 181 million reads | |
| *T. turgidum* *dicoccoides* Zavitan accession | WEW: GeneBank LSYQ00000000 BioProject PRJNA310175 | Leaf; root; flag leaf; developing spikes; glumes; flowers; grain; | 0.5 billion reads | (Avni et al., 2017) |
| Two wild emmer, two landraces, two durum cultivars | NCBI Short Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra/) under the accession numbers: SRR2084071, SRR2084163, SRR2084091, SRR2084165, SRR2084092, SRR2084160. | Glumes | 0.15 billion reads | (Zou et al., 2015) |

Table 4. Thirteen durum wheat accessions description.

| Accession | Year | Germplasm | Pedigree | Genotype feature | Breeder |
|---|---|---|---|---|---|
| Capeiti 8 | 1940 | Italian | Cappelli/Eiti | Founder | Istituto Sperimentale per la Cerealicoltura |
| Creso | 1974 | Italian/CIMMYT | CpB 144//Yt54-N10-B/ Cp2 63 Tc | Founder and parent of mapping population | ISEA |
| Valnova | 1975 | Italian/North America | Giorgio-324//Senatore Cappelli/Yuma | Founder | Istituto Sperimentale per la Cerealicoltura |
| Edmore | 1978 | North Am. | D6530//Leeds / Calvin | Founder | Western Plant Breeders |
| Yavaros 79 | 1979 | CIMMYT-'70 | Jori /Anhinga //Flamingo | Founder | CIMMYT |
| Altar 84 | 1984 | CIMMYT-'80 | RUFF"S"/FG"S"//MEXI75/3/SHWA"S" | Founder | CIMMYT |
| Neodur | 1987 | French/North America | 184-7/Valdur//Edmore | Parent of mapping population | Florisem |
| Kofa | 1996 | Desert Durum | dicoccum alpha pop-85 S-1 | Parent of mapping population | Westbred |
| Svevo | 1996 | Italian/CIMMYT | Cimmyt line/zenit sib | Parent of mapping population | Produttori Sementi S.p.A. |
| Meridiano | 1998 | Italian | Simeto/WB881//Duilio/F21 | Parent of mapping population | Produttori Sementi S.p.A. |
| Saragolla | 2002 | Italian/CIMMYT | Iride/O114 | Elite genotype | Produttori Sementi S.p.A. |
| Claudio | 2004 | Italian/CIMMYT | Sel.Cimmyt35/Durango//ISEA1938xGrazia | Parent of mapping population | Società Italiana Sementi, SIS |
| Strongfield | 2004 | North America | AC Avonlea'/DT665 | Elite genotype | Agriculture and Agri-Food Canada |

## 4.2. Results

### 4.2.1. Gene annotation and genome analysis

To distinguish genes from transposable elements, fuctional genes from pseudogenes and meaningful coding sequences from random ones, we applied a confidence classification for all predicted transcript protein sequences using BLASTP search against three databases: (i) TREP, (ii) Annotated *Poaceae* proteins and (iii) Validated *Magnoliophyta* proteins. Based on the E-value distribution for the best hits to each predicted protein, we set an E-value threshold to $10^{-10}$ (Figure 6).

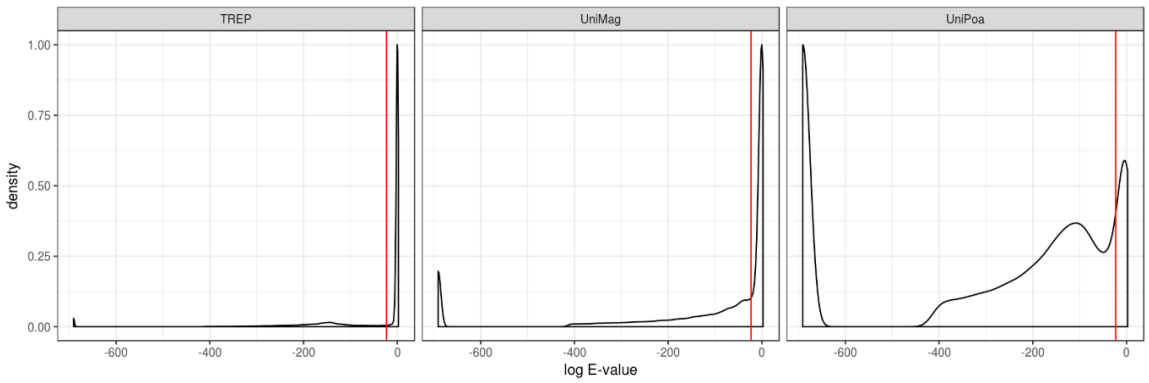
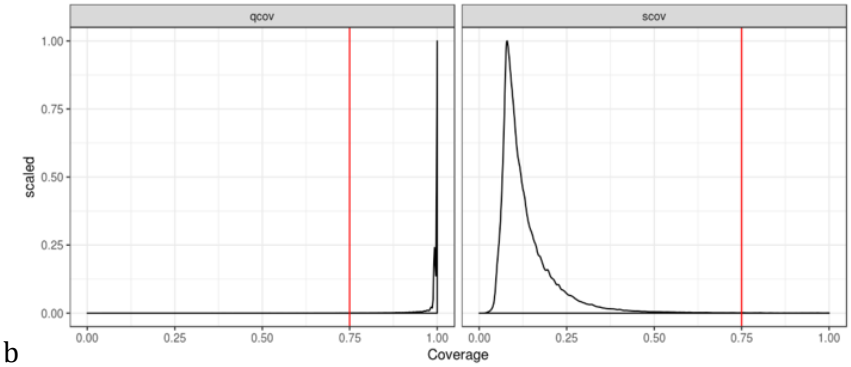


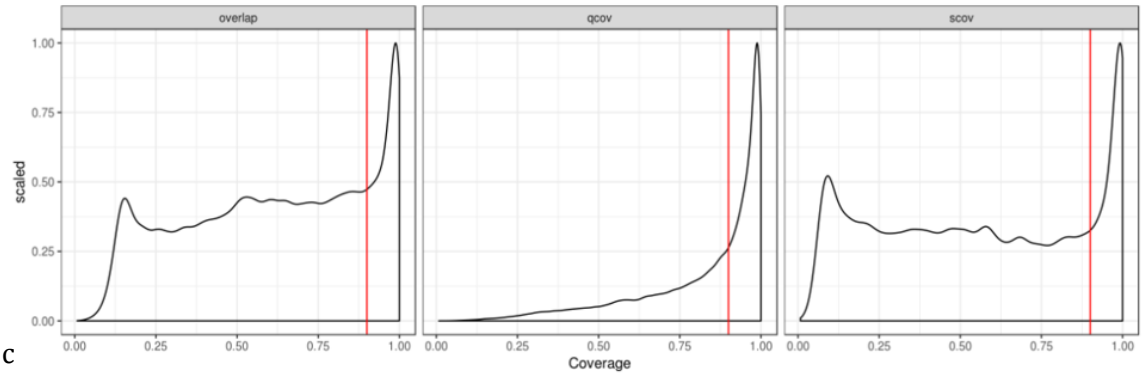

Figure 6. E-value distribution for BLAST result of predicted proteins against the three reference proteins databases. Red line indicates E-value=$10^{-10}$.

The best reference alignment was then selected for each query sequence and database based on alignment significance with maximal overlap between query and subject sequence for the two protein databases or as an alignment with maximal query coverage for TREP database. Multiple alignments for a protein with same coverage or query coverage were possible. To further filter reference alignments, we chose coverage thresholds for best alignments between predicted proteins and reference proteins. For the TREP database, most significant best alignments covered completely the predicted proteins while only parts of the reference proteins were covered. We chose a query coverage threshold of 75% to filter the best alignments further. For the UniMag database, a high amount the best alignments had a high query coverage as well as a high subject coverage. There was also a high amount of low subject coverage alignments, which indicates a significant number of fragmented protein sequences in the prediction set. Based on coverage distribution, we set the threshold for UniMag database to 90%. For the UniPoa database, most of the best alignments had a high query coverage as well as a high subject coverage. Based on coverage distribution, we set threshold to 90% as well (Figure 7).

a




b




c

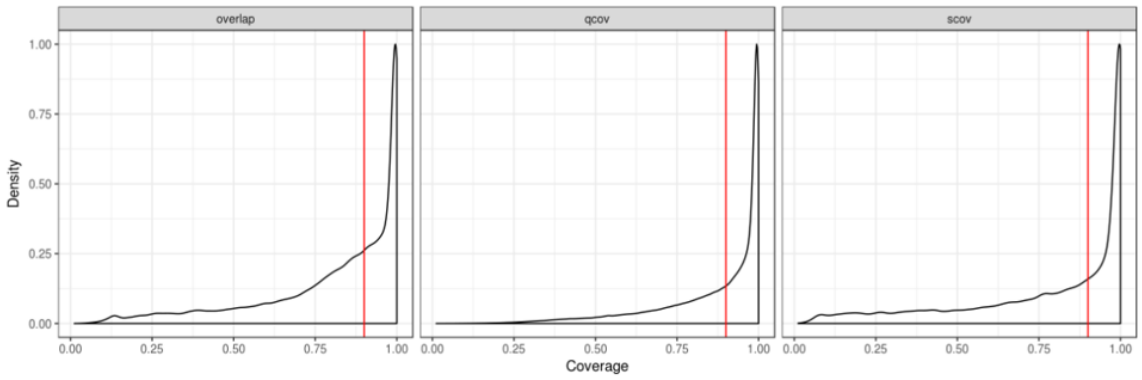



Figure 7. Coverage of significant best alignment to a) TREP b) *Poaceae* c) *Magnoliophyta* databases.

Using the gene annotation pipeline described above we predicted a total number of 369,963 genes: **66,559 high confidence (HC) and 303,404 low confidence (LC) genes**.

The table 5 summarizes the results of the gene annotation.

Table 5. Annotation statistics of durum wheat cv. Svevo high confidence (HC) and low confidence (LC) genes.

| Feature | HC | HC A | HC B | HC Un | LC |
|---|---|---|---|---|---|
| Number of genes | 66,559 | 31,718 | 32,275 | 2,566 | 303,404 |
| Mean loci size (bp) | 6,681 | 6,246 | 7,268 | 4,662 | 1,089 |
| Median loci size (bp) | 2,091 | 2,174 | 2,092 | 1,304 | 428 |
| Number of single transcript genes | 31,283 | 14,307 | 15,359 | 1,617 | 282,546 |
| Number of multi transcript genes | 35,276 | 17,411 | 16,916 | 949 | 20,858 |
| Number of transcripts | 196,153 | 96,213 | 94,259 | 5,681 | 341,975 |
| Mean transcripts per gene | 2.95 | 3.03 | 2.92 | 2.21 | 1.13 |
| Mean CDS size (bp) | 1,241 | | | | 520 |
| Median CDS size (bp) | 1,056 | | | | 414 |
| Mean exons per transcript | 4.6 | | | | 1.2 |
| Median exons per transcript | 3 | | | | 1 |
| Number of single exon transcripts | 17,250 | | | | 273,063 |
| Number of multi exon transcripts | 49,309 | | | | 30,341 |

The number of predicted HC transcripts was slightly higher on subgenome B compared to subgenome A, except for chromosomes 4 and 7 as expected due to ancient translocation between 7B and 4A. We found 2,566 HC genes on chromosome unknown (chrUn). There were 196,153 transcripts with an average of 2-3 transcripts per gene. Out of 66,559 genes 31,283 had only one transcript, the rest 35,276 had multiple number of transcripts. The mean length of coding sequences was 1,241 bp. Most identified coding sequences (64.6 %) translated to complete protein sequences with start and stop codon, and their mean length was 273.8 amino acids (Figure 8 and 9).

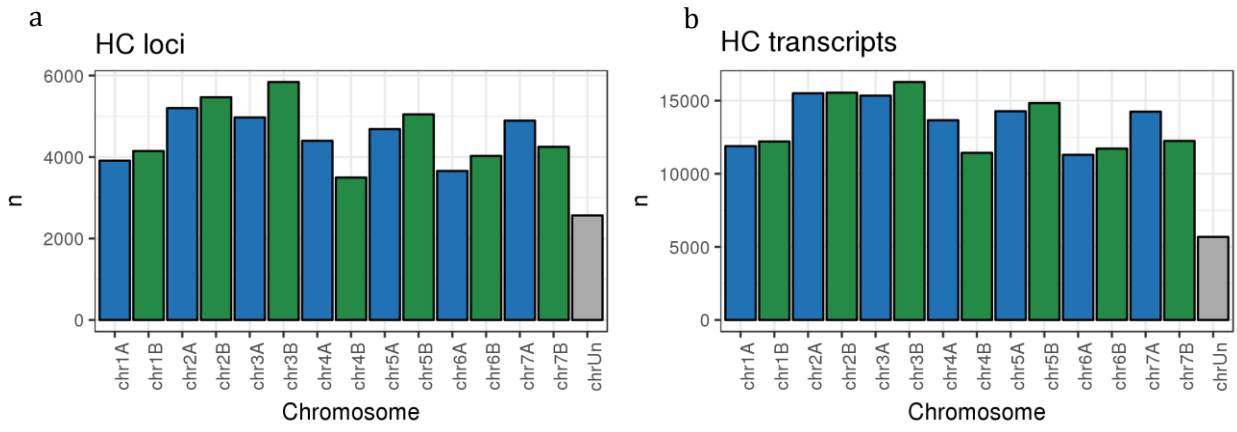



Figure 8. Number of predicted a) HC loci and b) HC transcripts per chromosome.

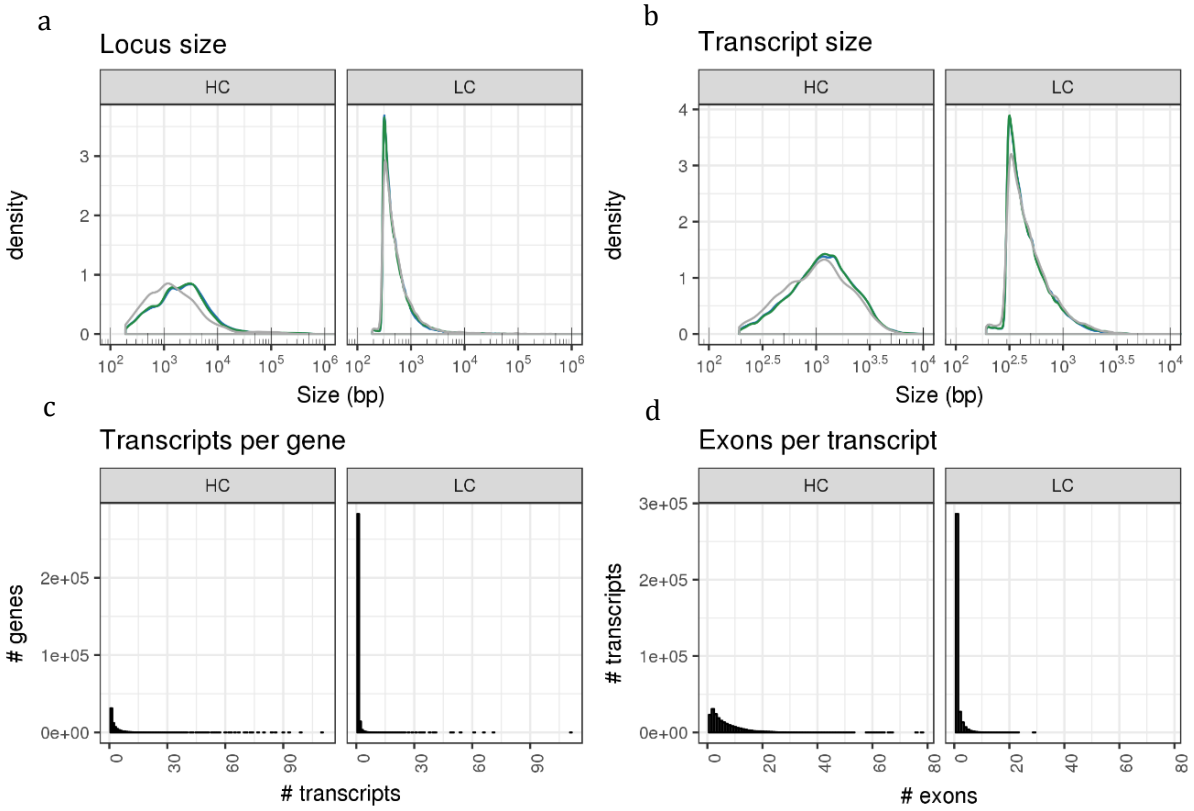



Figure 9. a) Size of predicted loci. b) Size of predicted transcript d) Number of transcripts per gene d) Number of exons per transcript.

Validation of the genome assembly and annotation

The 98.1% (n = 1,413) of BUSCO genes were found in the predicted gene set. This high value indicates that the assembly represents an almost complete fraction of the gene space. Furthermore, 96.1% of the BUSCO genes were fully represented by the HC gene sets (Figure 10).

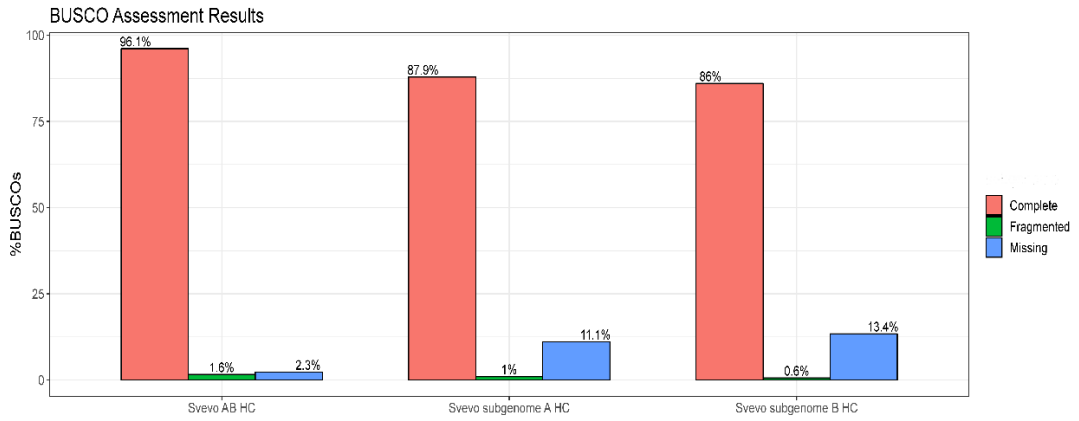



Figure 10. BUSCO validation results.

In addition to the BUSCO analysis, the predicted proteins were validated using 216 experimentally determined genes. The 97.7% (n = 211) of these genes were represented by at least one annotated gene with at least 75% protein coverage and E-value < $10^{-05}$. Most of them were represented in the HC gene sets (95.4%). Then, the set of 204,773 unique reference protein sequences was searched in BLAST against the HC gene set and 194,131 proteins had a significant hit to predicted genes (E-value < $10^{-05}$). From these genes, an annotated gene with at least 75% query coverage represented 92.3%. These results indicate that high confidence gene sets represent a large amount of already known protein sequences. Missing *Triticeae* genes that are not represented by the annotations may also include transposons and species-specific genes, especially genes that belong to *Aegilops tauschii* or wheat D subgenome.

### 4.2.2. Gene expression pattern

We investigated the gene expression pattern of HC gene sets. 2,566 genes on chrUn were excluded for this analysis. Therefore, the analysis revealed that out of 63,993 HC genes 61,269 (95.8%) genes were expressed at least in one of the 57 samples and 2,724 (4.3%) of genes were not expressed at all. The 21,878 genes (34.2%) were expressed in all 57 samples. We found that the mean expression level per gene (mean expression value of all 57 samples) across all the 57 samples along each chromosome, and mean expression density of genes (number of libraries in which gene was expressed; from 0 to 57) was higher in the centromere distal regions of the chromosomes rather than centromere proximal regions (Figure 11). While the mean expression and expression breadth were higher in the centromere proximal regions, the average number of genes expressed were higher in the distal regions and lower in the centromere regions (Figure 12). This is in correlation with the number of genes present in the centromere and telomere regions (Figure 4).

We analyzed the expression pattern of these genes along the chromosomes at 20 Mb window. While the number of the expressed genes is higher in the distal regions of the chromosome arms, the number of libraries under which these genes are expressed is lower in the distal regions and higher in the pericentromeric region, suggesting that there are more condition specific genes rather than housekeeping genes.

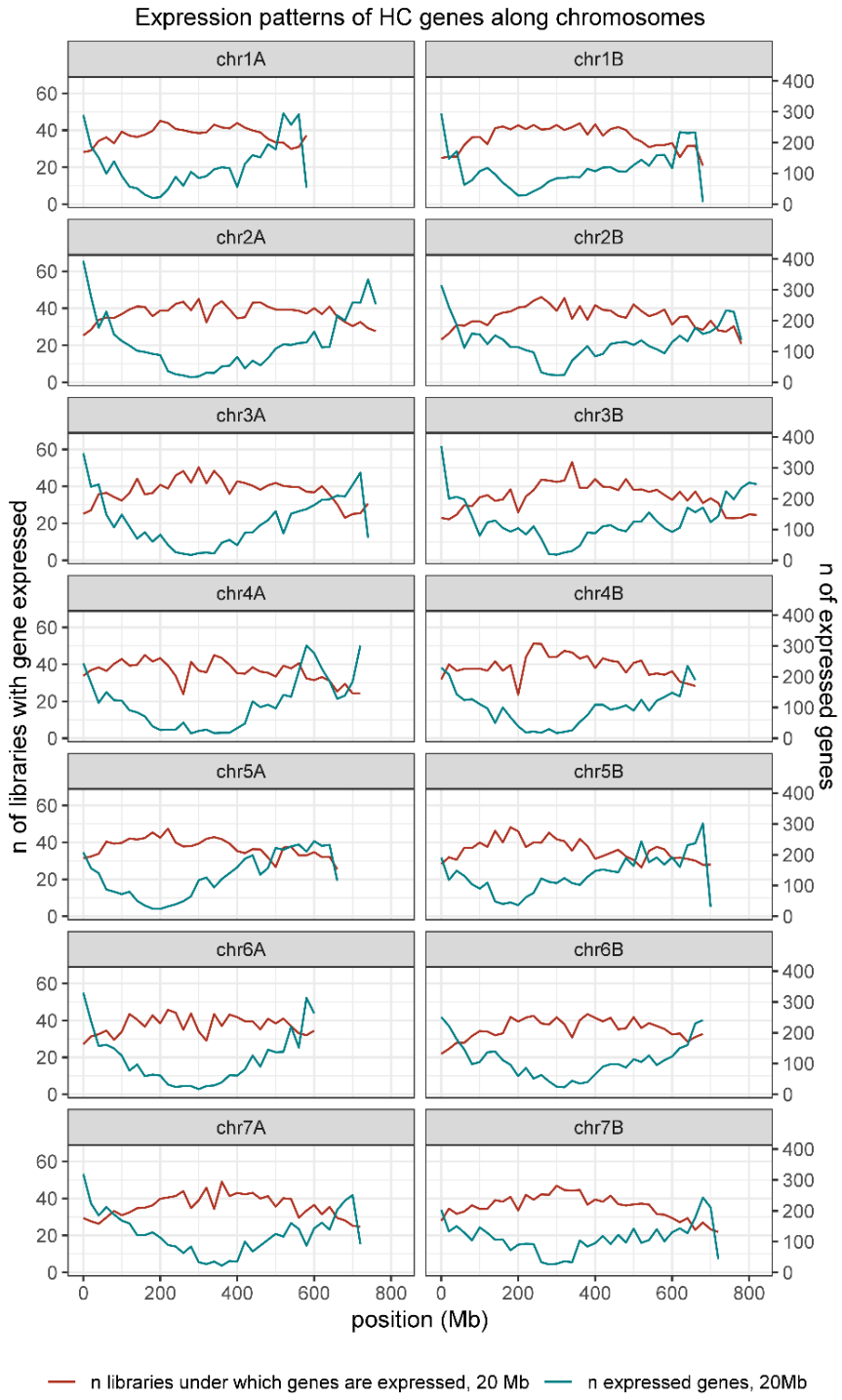



Figure 11. Expression pattern of high confidence genes. Average number of genes expressed at 20 Mb window size – blue line; Expression breadth density (the number of libraries in which the genes were expressed, from 0 to 57) per chromosome at 20 Mb window size - red line.

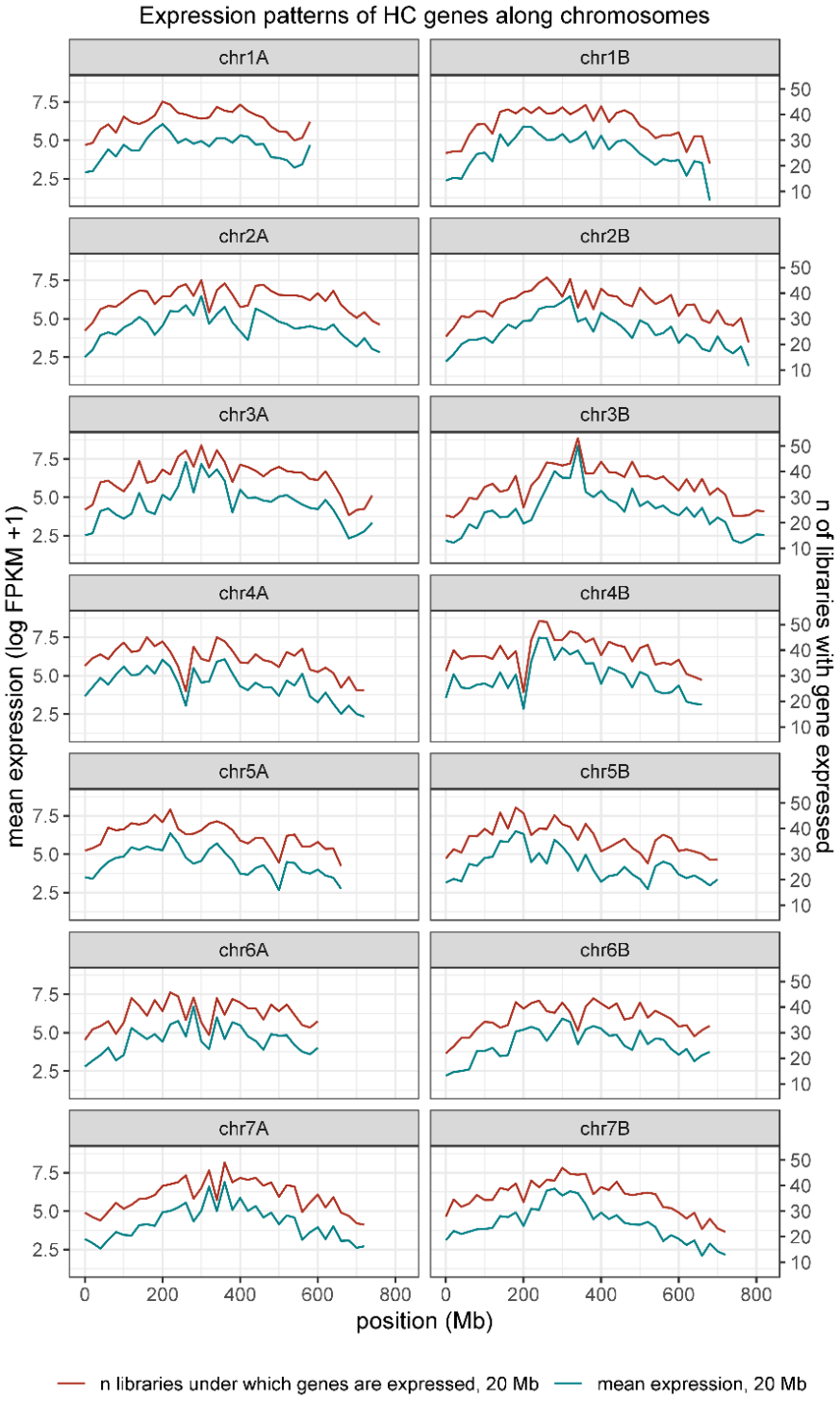



Figure 12. Expression pattern of high confidence genes. Expression breadth density, the number of libraries in which the genes were expressed at 20 Mb window, from 0 to 57 – red line; Expression value across all 57 samples as mean [log(FPKM+1)] over 20 Mb window size – blue line.

There was no significant difference between the number of genes expressed in A and B sub-genomes. In total, there are 31,718 and 32,275 total high confidence gene models in the A and B sub-

genomes, respectively, out of the total 63,993 HC genes on 7 homeologous pairs of chromosomes. On average, there is a 1.7% increased number of expressed genes in subgenome B compared to subgenome A. There are 30,873 HC gene models expressed in the B subgenome and 30,396 HC genes expressed in the A subgenome. Considering the cultivars and the tissue/organ libraries overall, the percentage of genes mapped to the Svevo reference genome varied from 48.0% (Altar84, Capeiti 8, Claudio, Saragolla leaves) to a maximum of 61.0% (Meridiano, Strongfield roots and grains) (Figure 13).

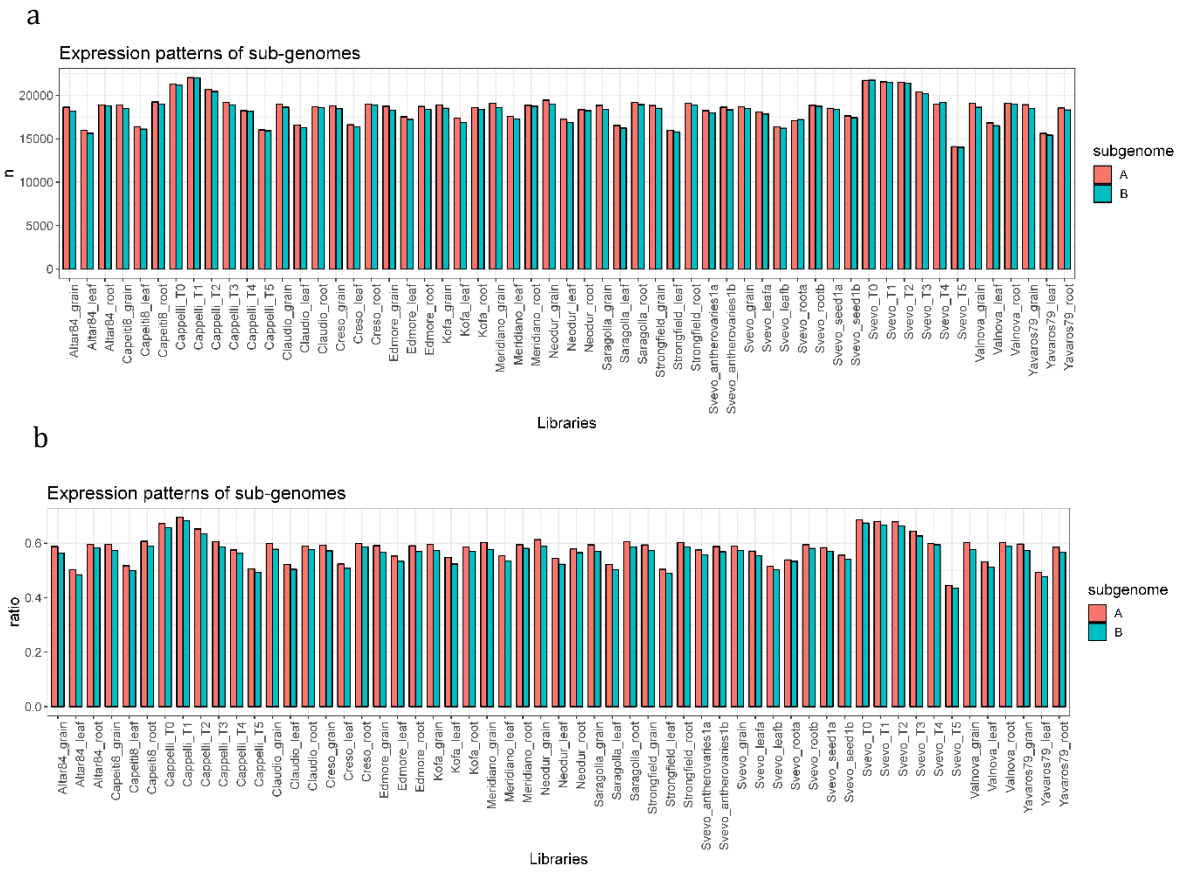



Figure 13. a) Number of genes expressed in A and B subgenomes. 30,396 and 30,873 total expressed HC gene models in A and B subgenomes, respectively. b) Ratio of genes expressed in subgenome A and B over the total number of genes in each subgenome, under all conditions.

Comparison of 13 varieties

We investigated the expression variation of the thirteen worldwide elite durum wheat varieties with the grain, leaf and root tissues. The RNA-seq libraries had no replicates. Out of the 63,993 high confidence gene models, overall,

- 55,428 (86.6%) of genes expressed in all tissues and all varieties;
- 48,007 (75.0%) of genes were expressed in grain;
- 45,142 (70.5%) were expressed in leaf;
- 47,702 (74.5%) of genes were expressed in roots.

There was a stronger variation between tissues than between the varieties. According to the PCA analysis, clustering dendrogram and the heatmap (Figure 14) there are three clear gene expression clustering lead by organs/tissues (leaf, grain and root) for the 13 varieties (leaves, grains and roots accounting for 33.0% variance at two-dimensional clustering). However, at the three-dimensional clustering we observed only 2.0% variance within the tissues.

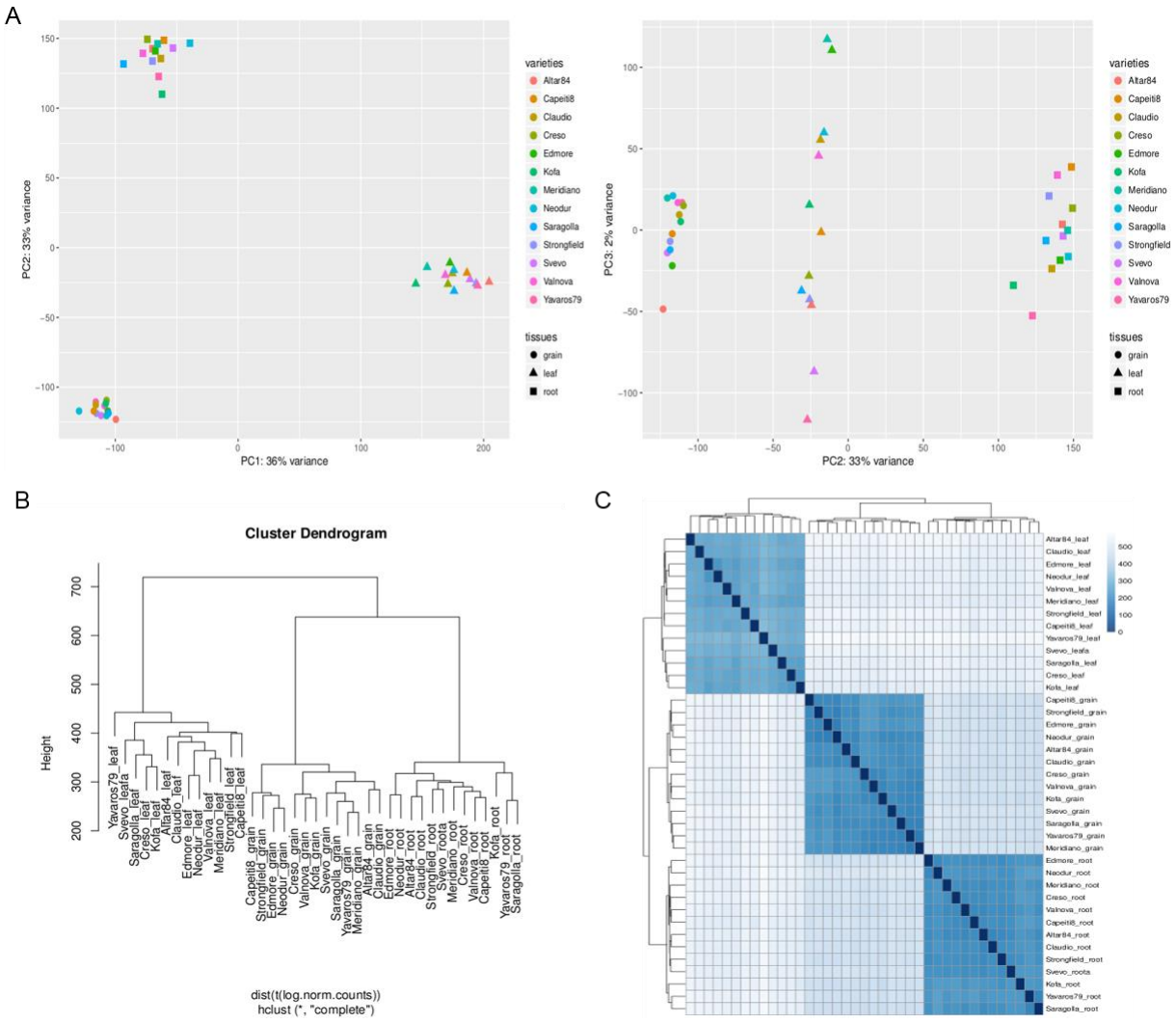



Figure 14. A. Two-dimensional and three-dimentional PCA. B. Hierarchical clustering of normalized gene expression. C. Heatmap clustering of normalized gene expression.

We attempted to analyze the major variation pattern within the tissues and varieties. In order to do this, we used sparse PCA analysis (Zou et al., 2006). Sparse PCA does feature selection and

retains the features that drive the major pattern in the data. Using sparse PCA, we retained 510 genes and plotted their expressions as heatmap (Figure 16, 17). The major variance was explained in PC1 and PC2 (Figure 15A); hence, we used sparse PCA based on the strongest PC1–PC2 scores. We used a k-means estimation and set optimal number of clusters to 4 (Figure 15B).

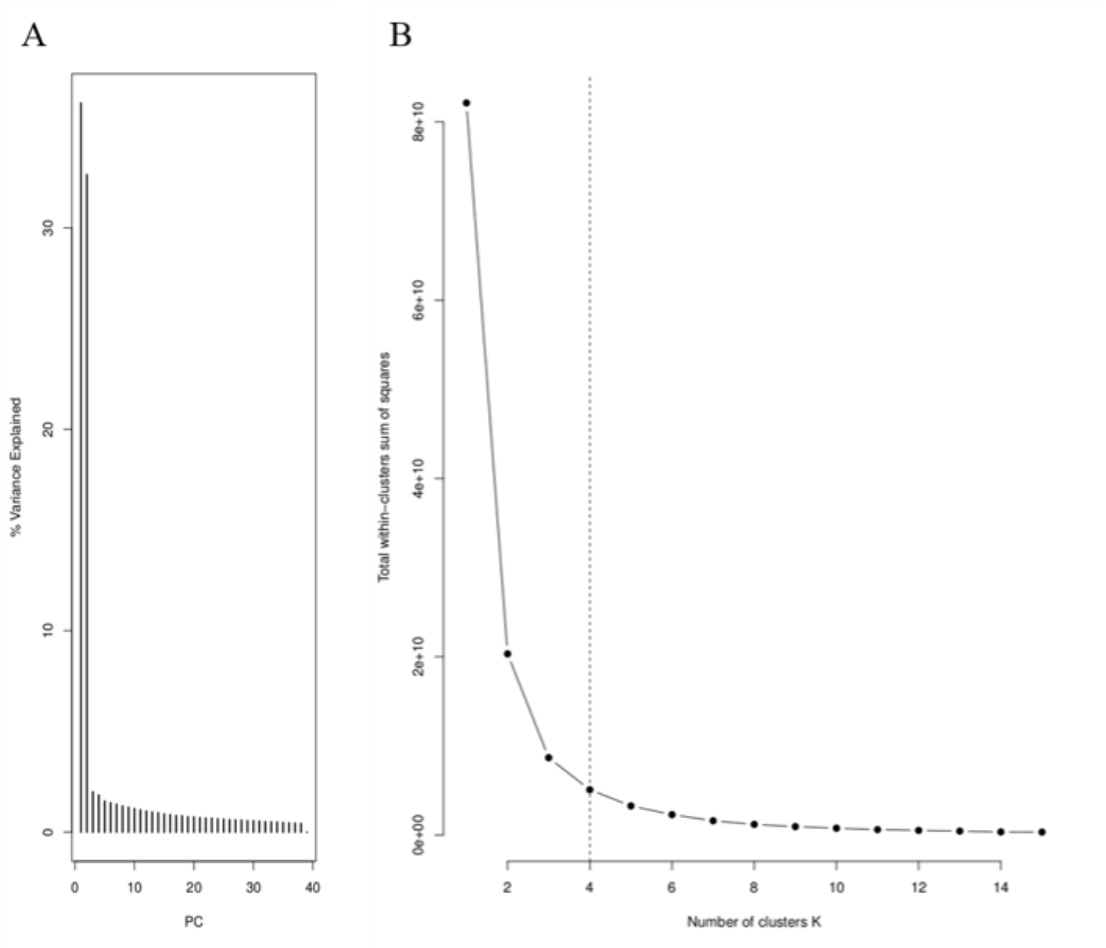



Figure 15. A. Percent of variance explained in principal component analysis. B. K-means clustering aimed to estimate the optimum number of clusters (set to 4).

According to the results of our analyses, by clustering the gene expression profiles and the cultivar's expression profiles several gene expression patterns related to the ancestry relationship among cultivars were evidenced (Figure 17). For example, for grain tissue American and old cultivars formed one cluster, while CIMMYT and Italian germplasm formed a second cluster (Figure 17). Moreover, we calculated the variance along varieties for the three tissues (Figure 18). This analysis allows us to identify the chromosome regions that drive the major expression variation pattern.

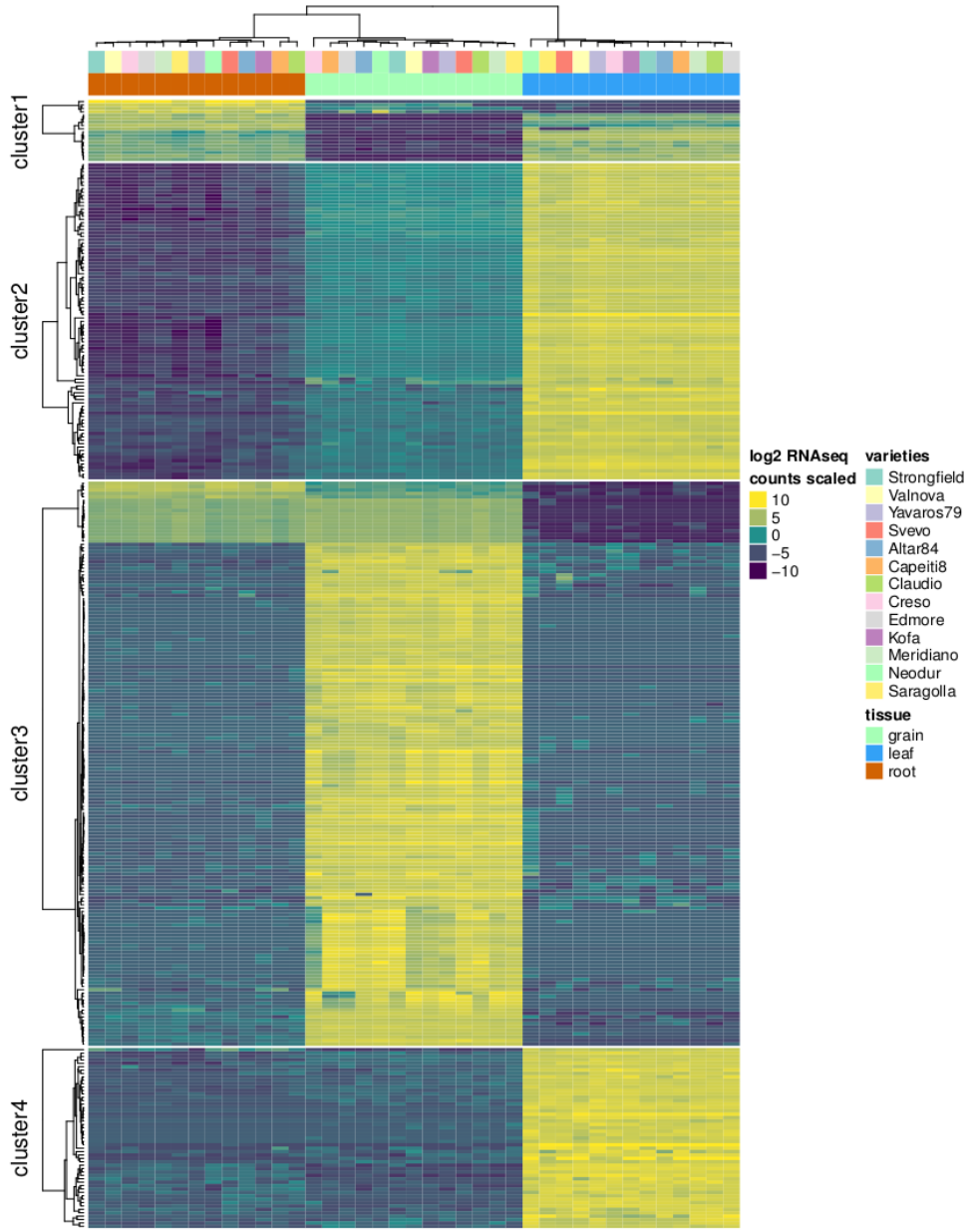



Figure 16. Heatmap of 510 genes that drive the major variation pattern retained after sparse PCA. The heatmap is based on PC1 and PC2 scores.

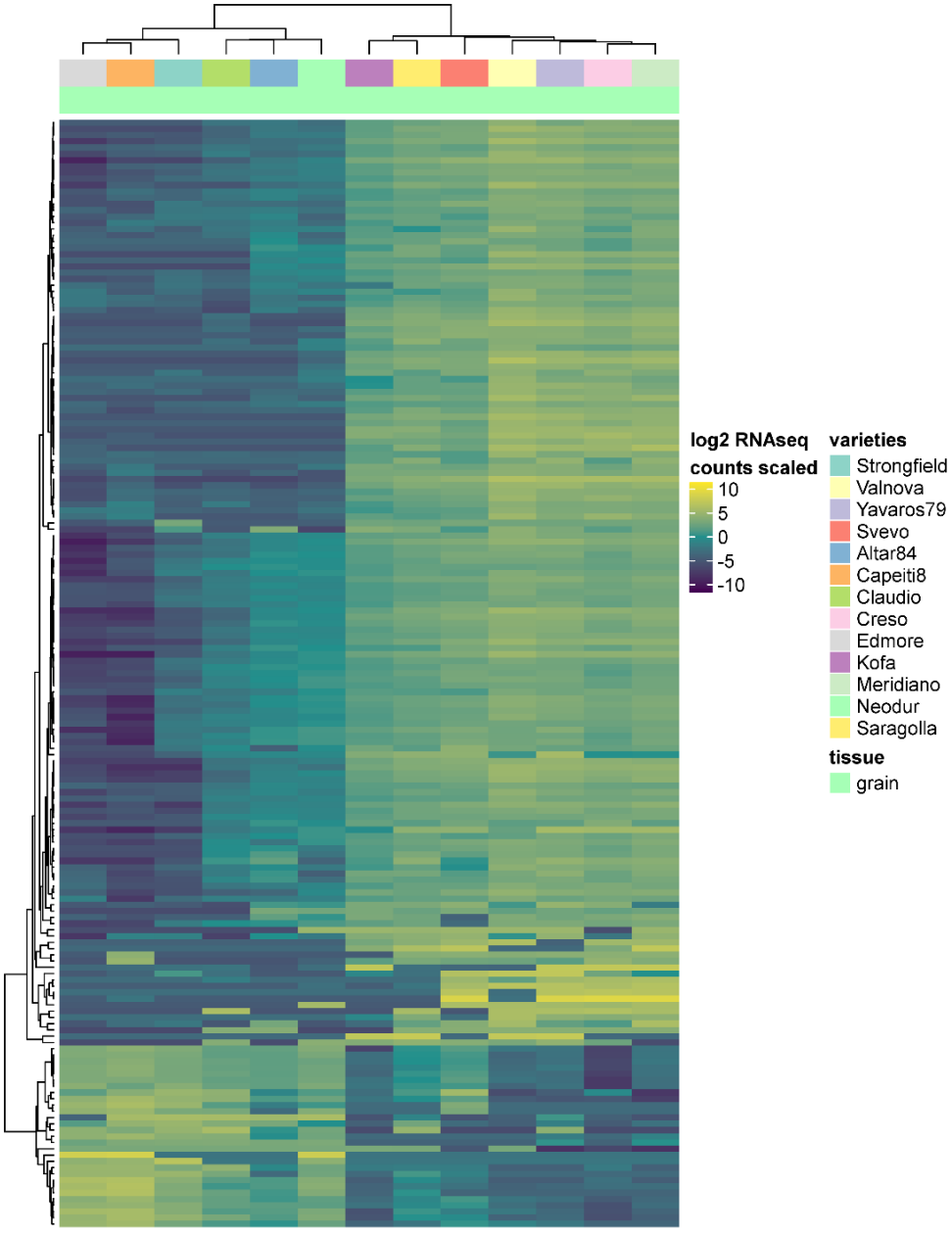



Figure 17. Heatmap of genes expressed in grain tissue that drive the major variation pattern retained after sparse PCA. The heatmap is based on PC1 scores.

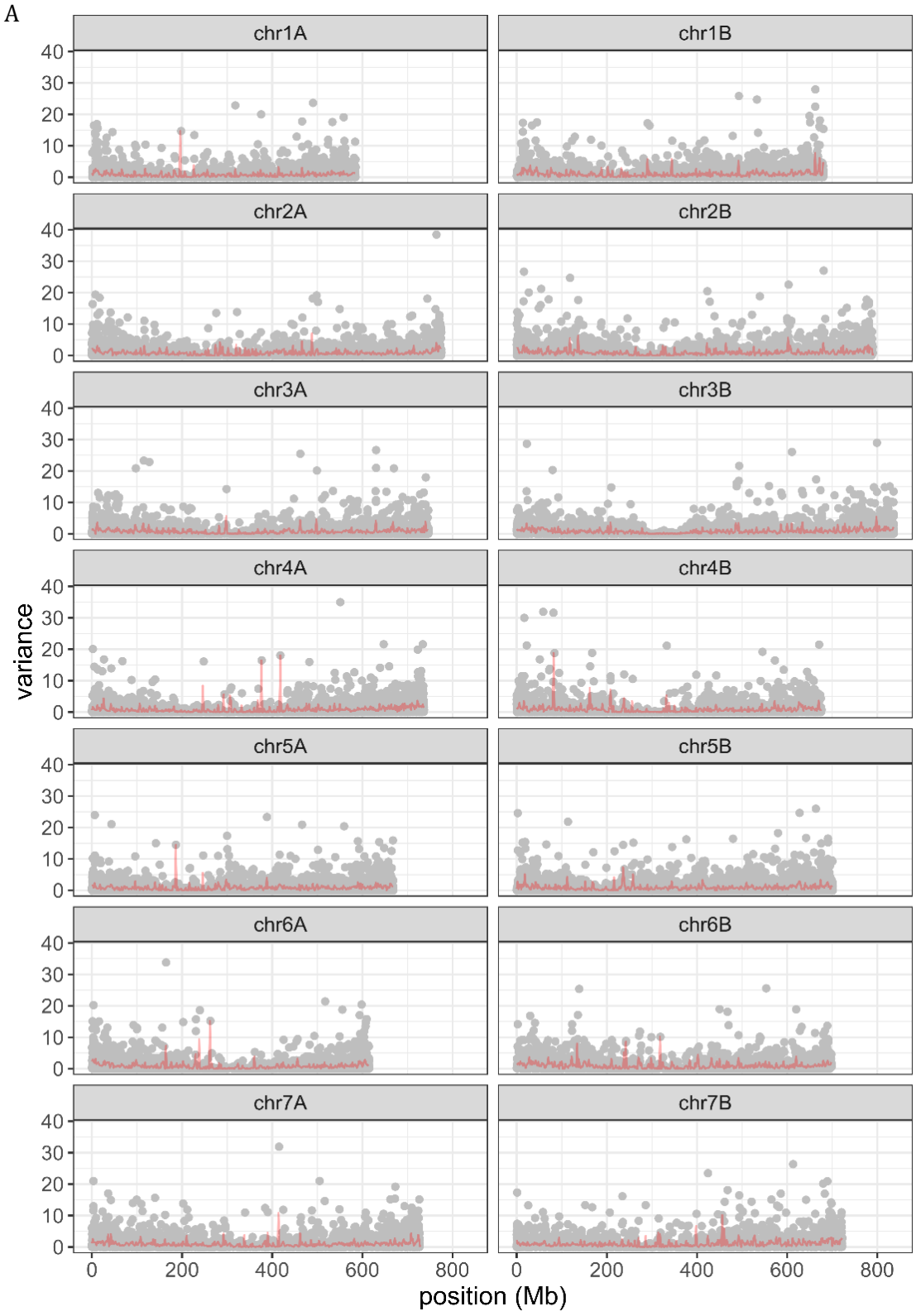



Figure 18. Variance expression analysis of the thirteen varieties for A. grain, B. leaf and C. root tissues. Red line indicated the mean variance at 2 Mb window.

B

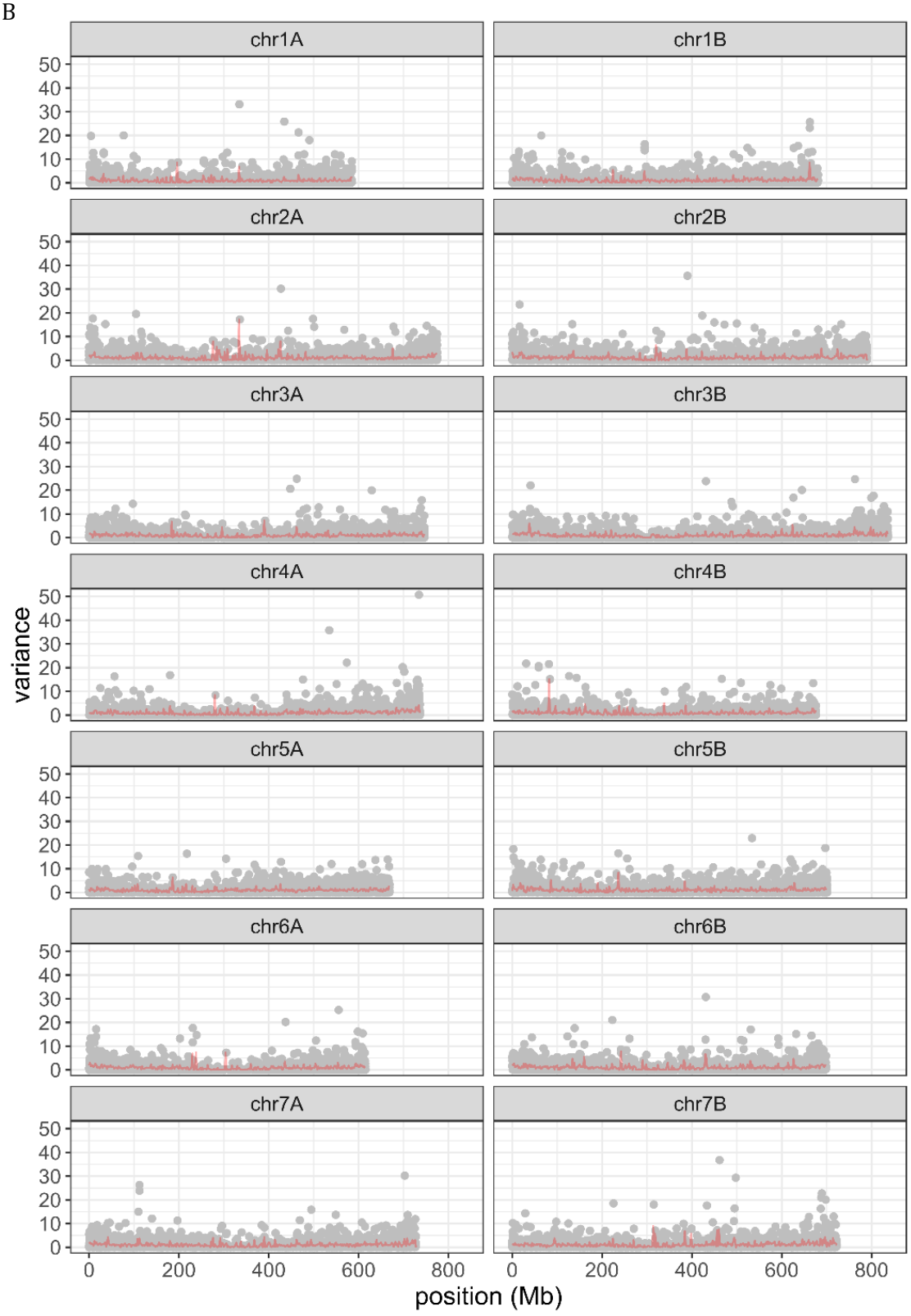



Figure 18. Continued.

C

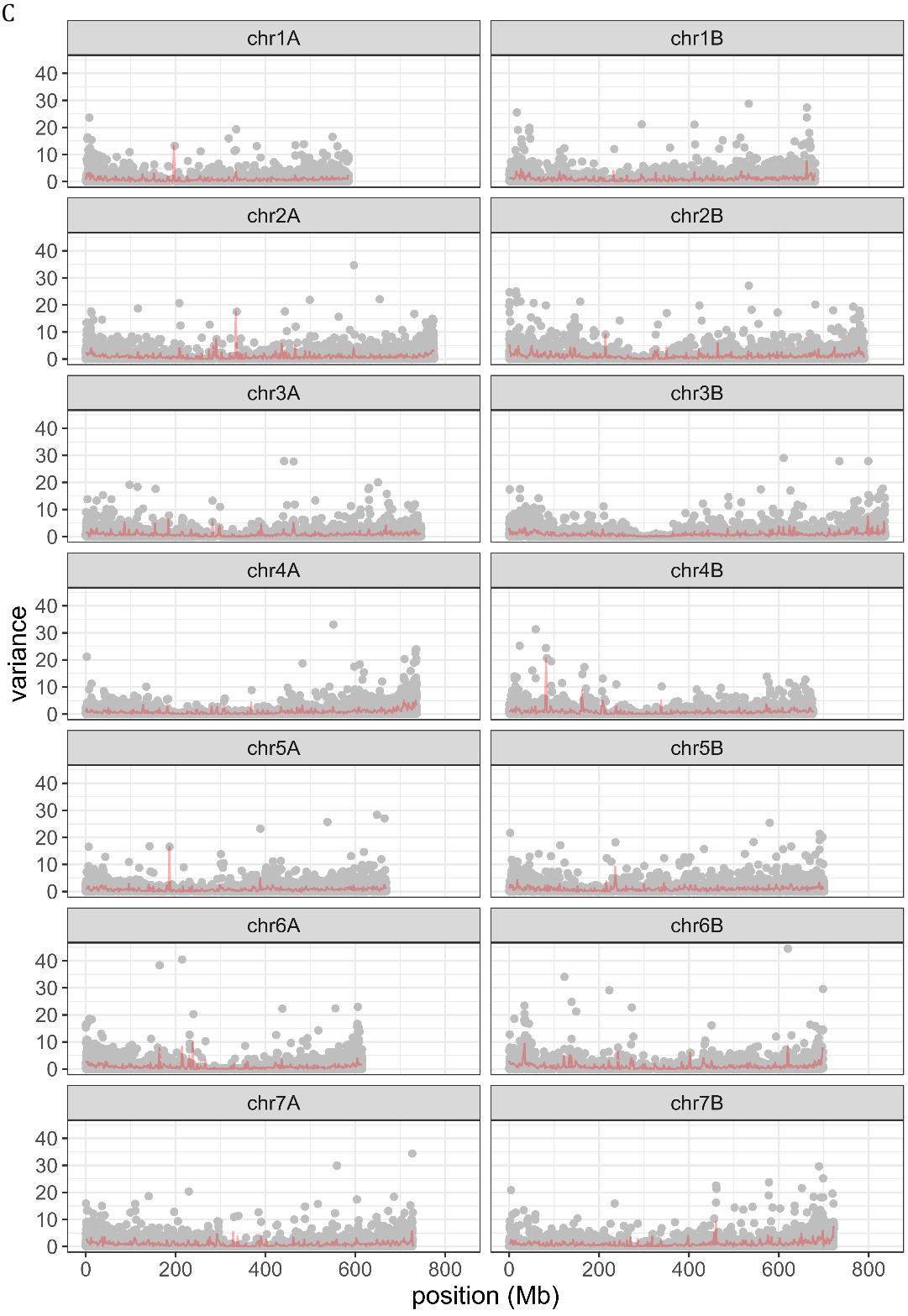



Figure 18. Continued.

### 4.2.3. NB-LRR gene family organization in durum and wild emmer wheat

Using NLR-annotator version 0.7 (https://github.com/steuernb/NLR-Annotator), we predicted 2,442 and 2,420 NB-LRR loci for durum wheat cv. Svevo e wild emmer wheat accession Zavitan, respectively. Compared to RNA-seq gene models, this tool allowed an annotation of an additional 390 loci (16.5%) in Svevo genome and 417 loci (17.2%) in Zavitan (Table 6).

Table 6. NLR-annotator statistic results for durum wheat cv. Svevo and wild emmer wheat accession Zavitan.

| Features | Durum wheat cv. Svevo (%) | Wil emmer wheat, acc. Zavitan (%) |
|---|---|---|
| Missed exons | 99.1 | 99.1 |
| Novel exons | 26.5 | 28.2 |
| Missed introns | 99.0 | 99.0 |
| Novel introns | 61.5 | 58.8 |
| Missed loci | 99.4 | 99.4 |
| **Novel loci** | **16.5** | **17.6** |
| Total union super-loci across all input datasets | 1950 | 1920 |

In durum wheat, out of annotated 2,442 NLR-encoding loci 1,487 were complete genes, 814 were pseudogenes/partial genes and 141 were complete genes on chrUn. In wild emmer wheat, out of predicted 2,420 NB-LRR loci 1,462 were complete genes, 857 were pseudogenes/partial genes and 101 were complete genes located on chrUn (Table 7).

Table 7. The number of total, complete and pseudogene NLR-loci for durum wheat and wild emmer wheat.

| **Features** | **Durum wheat cv. Svevo** | **Wild emmer wheat acc. Zavitan** |
|---|---|---|
| Total | 2,442 | 2,420 |
| Complete NLRs (incl. pseudogenes) | 1,826(1671 without chrUn) | 1,743(1627 without chrUn) |
| pseudogenes | 333 (198 complete) | 336 (180 complete) |
| Partial NLRs | 616 | 677 |
| **Complete NLRs** | **1,487** | **1,462** |

We observed the NLR loci clustering principally at the distal regions of the chromosome arms and overlapping with confidence intervals of disease resistance QTLs known from literature (Figure 19A). The set of complete 1,487 durum wheat NB-LRR-encoding loci were aligned using BLASTP (version blast-2.2.26 , E-value $10^{-10}$) versus the set of 1,467 complete wild emmer wheat NB-LRR loci. The generated output was filtered for identity, query and subject coverage of >70%. As a result, we identified 172 loci specific for Svevo and 136 loci specific for Zavitan. The distribution of these NLR genes on the chromosomes showed that the most dissimilar regions between the durum wheat and emmer wheat were localized on subgenome B (Figure 19B).

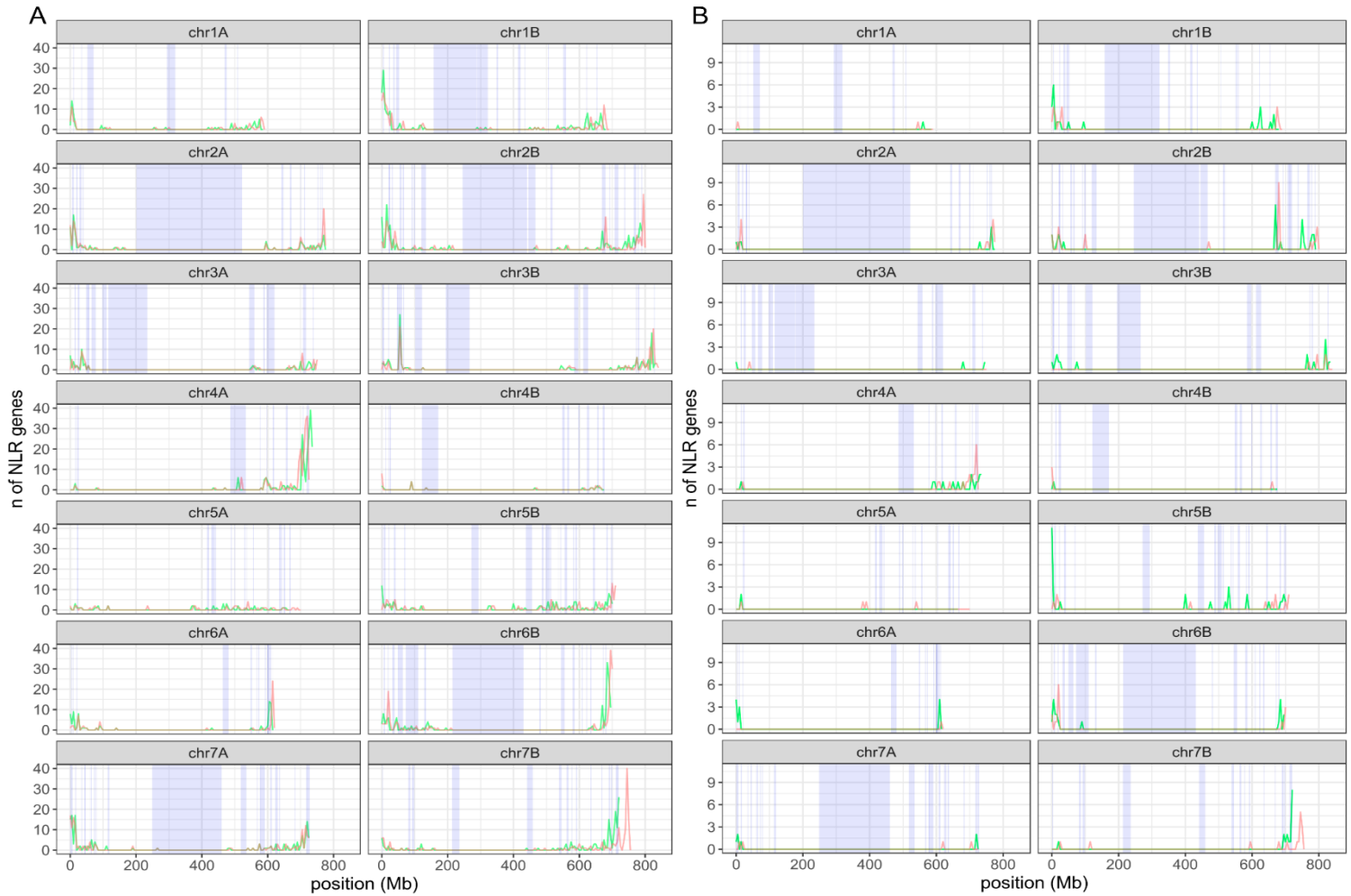



Figure 19. Whole genome NLR gene density graph. Blue transparent area represents confidence intervals of published disease resistance QTLs. (A) Nucleotide-binding leucine-rich repeat (NLR) gene density graph at 5 Mb window. Green color represents Svevo; red, Zavitan. (B) Gene density of 172 NLR loci specific for durum and 136 for wild emmer wheat at 5 Mb window.

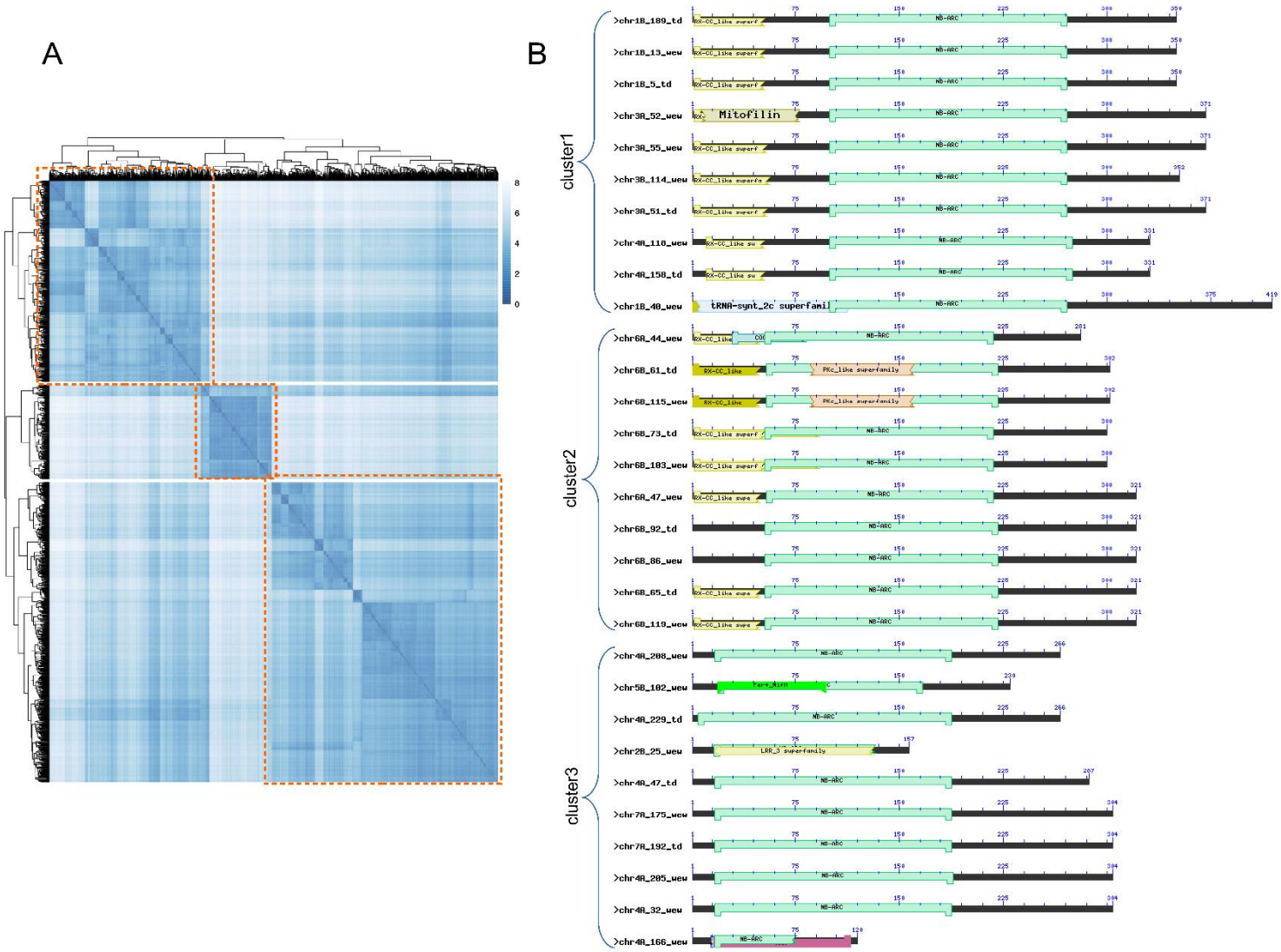



Figure 20. NB-LRR sub-groups. (A) Similarity heatmap of complete sets of durum and wild emmer wheat NB-LRR loci. Orange dashed lines delimit the three NB-LRR loci clusters. (B) Domain composition of 10 most representative NLR loci sequences for the three clusters.

The complete set of NLR genes both for durum and wild emmer wheat were aligned all versus all using clustal-omega-1.2.4 (Sievers et al., 2011). A heatmap of similarity matrix based on alignment distance was generated using the R package *seqinr* and *pheatmap* (Kolde, 2018). Based on this multiple-alignment analysis (the heatmap similarity matrix is reported in Figure 20) we identified three main NB-LRR loci clusters. These main clusters were differentiated at the level of domain composition (Marchler-Bauer et al., 2017), position of domains inside the putative genes as well as the amino acidic difference at the domain level (Figure 20 and 21). While in clusters 1 and 3 the ratio of durum and wild emmer wheat NB-LRRs was similar, cluster 2 was more enriched in durum wheat (8%) NLRs loci compared to wild wheat.

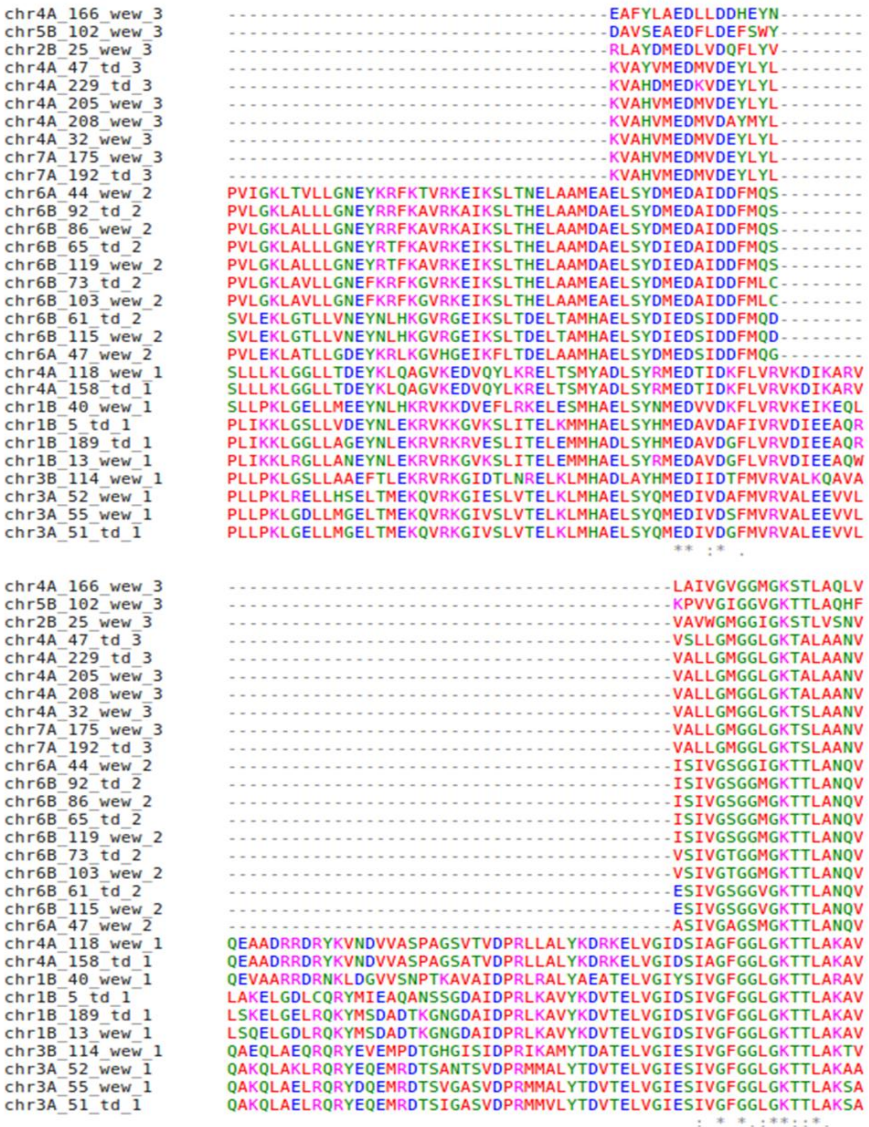



Figure 21. Multiple Sequence Alignment of NLR loci. We report the domain alignment of ten most representative NB-LRR loci sequences for each of the three NLR clusters.

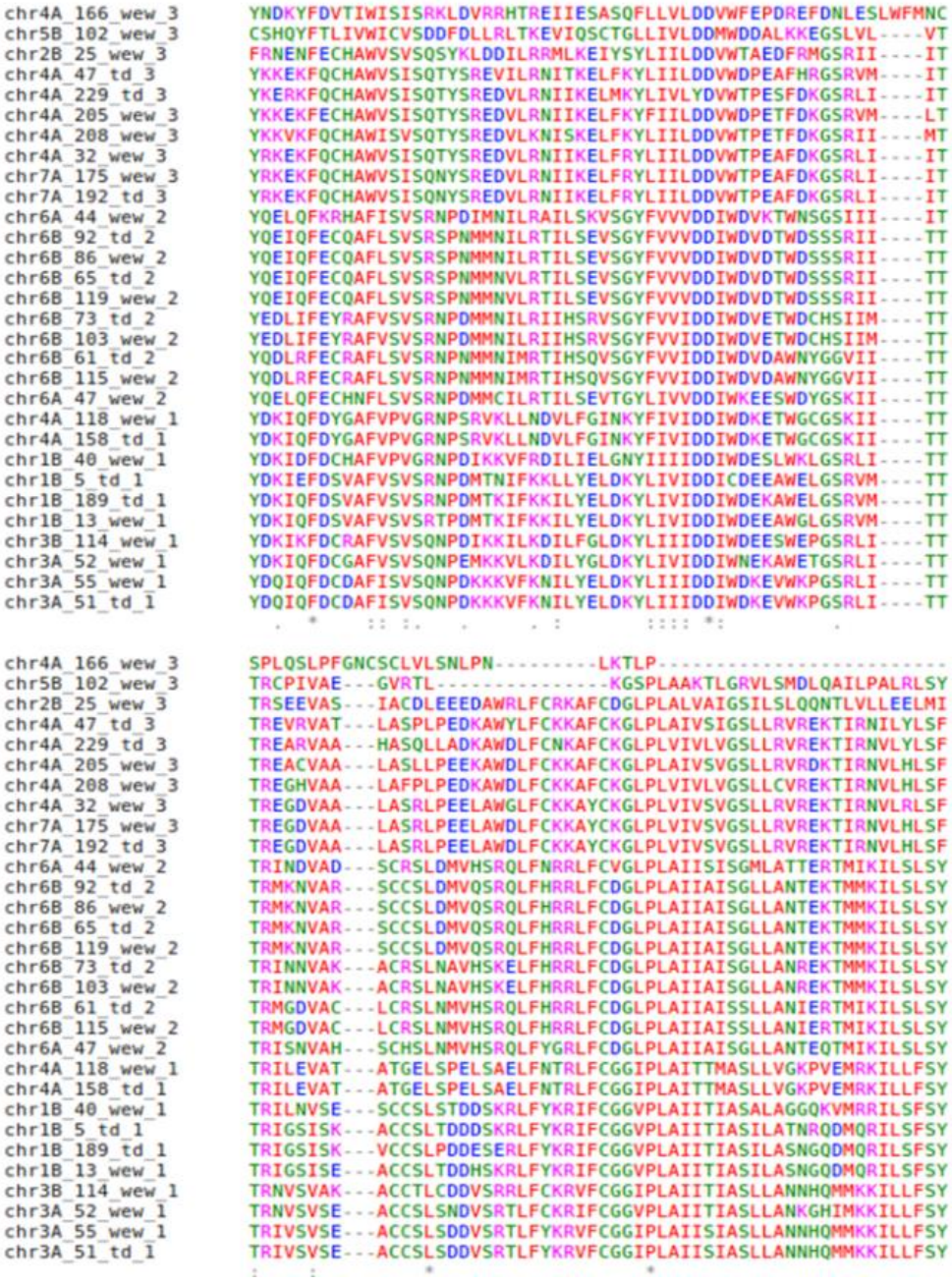



Figure 21. Continued.


Figure 21. Continued.


Figure 21. Continued.

# 5. Chapter 2. Diversity reduction and selection signature in tetraploid wheat germplasm

## 5.1. Materials and Methods

### 5.1.1. Global Tetraploid wheat Collection

The Global Tetraploid wheat Collection (GTC) was composed of up to ten different species and subspecies (Table 8):

- Persian wheat (*Triticum turgidum* L. subsp. *carthlicum (Nevski)* Á. & D. Löve);
- Wild emmer wheat (*Triticum turgidum* L. subsp. *diccocoides* (Körn. ex Asch. & Graebner));
- Durum wheat landraces and cultivars (*Triticum turgidum* L. subsp*. durum* (Desf.) Husn.);
- Emmer wheat (*Triticum turgidum* L. subsp*. dicoccum* (Schrank ex Schübler) Thell*.* and *Triticum ispahanicum* Heslot);
- Polish wheat (*Triticum turgidum* L. subsp*. polonicum* (L.) Thell.);
- Khorasan wheat (*Triticum turgidum* L. subsp*. turanicum* (Jakubz.) Á. & D. Löve);
- Miracle wheat (*Triticum turgidum* L. subsp*. turgidum*);
- Karamyschev's wheat (*Triticum karamyschevii*);
- Ethiopian wheat (*Triticum aethiopicum* Jakubz.).

Table 8. Composition and number of different species and subspecies of 1,856 tetraploid wheat accessions (AABB genome) that composed the Global Tetraploid wheat Collection.

| **Wheat species or subspecies** | **Common name** | **Genome** | **No.** |
|---|---|---|---|
| *Triticum karamyschevii* | Karamyschev's wheat | AABB | 2 |
| *Triticum aethiopicum* Jakubz. | Ethiopian wheat | AABB | 16 |
| *Triticum turgidum* L. subsp. *carthlicum (Nevski)* Á. & D. Löve | Persian wheat | AABB | 20 |
| *Triticum turgidum* L. subsp. *dicoccoides.* (Körn. ex Asch. & Graebner) | Wild emmer wheat | AABB | 115 |
| *Triticum turgidum* L. subsp*. dicoccum* (Schrank ex Schübler) Thell*.* | Domesticated emmer wheat; Emmer | AABB | 364 |
| *Triticum turgidum* L. subsp*. durum* (Desf.) Husn. (landraces) | Durum wheat or pasta wheat | AABB | 806 |

| | | | |
|---|---|---|---|
| *Triticum turgidum* L. subsp. *durum* (Desf.) Husn. (registered cultivars or breeding lines) | Durum wheat or pasta wheat | AABB | 427 |
| *Triticum ispahanicum* Heslot | Domesticated emmer wheat; Emmer | AABB | 2 |
| *Triticum turgidum* L. subsp*. polonicum* (L.) Thell. | Polish wheat | AABB | 22 |
| *Triticum turgidum* L. subsp*. turanicum* (Jakubz.) Á. & D. Löve | Khorasan wheat | AABB | 74 |
| *Triticum turgidum* L. subsp*. turgidum* | Rivet, Cone, English wheat or Miracle wheat | AABB | 8 |

Four main germplasm groups

The GTC collection consisted of 1,856 accessions. These accessions represent the four principal germplasm groups that are involved in the history of tetraploid domestication and selection processes:

- Wild Emmer Wheat, WEW;
- Domesticated Emmer Wheat, DEW;
- Durum Wheat Landraces, DWL;
- Durum Wheat Cultivars, DWC.

The GTC contained germplasm accessions from a wide range of areas including Fertile Crescent, Northern Africa, Europe, India, Ethiopia, Transcaucasia, Central Asia, North and South America (Table 9).

Table 9. Tetraploid diversity panel by country of origin, based on passport data. Accessions were categorized by subspecies and geographical aggregates in accordance with the United Nations M-49 list, except for the Fertile Crescent (Turkey, Southern Levant).

| **Tetraploid wheat subspecies/Geographical area** | **Accessions no.** |
|---|---|
| **Wild Emmer Wheat (*T. turgidum* subsp. *dicoccoides*)** | |

| | |
|---|---|
| Fertile_Crescent_Southern_Levant (Lebanon, Syria, Jordan, Israel) | 67 |
| Fertile_Crescent_North-East (Turkey, Karacadag) | 36 |
| **Domesticated Emmer Wheat (*T. turgidum* subsp. *dicocccum, T. ispahanicum*)** | |
| Fertile Crescent (Turkey) | 17 |
| Fertile Crescent (Southern Levant, Lebanon-Syria-Jordan-Israel-Palestine) | 25 |
| Fertile_Crescent (general, not detailed) | 11 |
| Eastern Africa (Ethiopia-Kenia) | 46 |
| Southern Asia (Iran-Afghanistan) | 39 |
| Southern Asia (India) | 18 |
| Western Asia-Transcaucasia (Armenia-Georgia-Daghestan-Azerbaijan) | 31 |
| Western Asia (Oman-Yemen-Kuwait-Saudi Arabia) | 12 |
| Northern Africa (Morocco-Tunisia) | 8 |
| Southern Europe (Greece-Albania-Serbia-Bosnia-Montenegro-Italy-Spain-Portugal) | 66 |
| Western Europe (Austria-Switzerland-Germany) | 14 |
| Eastern Europe (RussianFederation-Belarus-Poland-Ukraine) | 22 |
| Eastern Europe (Romania-Slovenia-Hungary-Czech Republic-Bulgaria) | 21 |
| Northern Europe (UK) | 15 |
| Central Asia (Kazakhstan-Uzbekistan) | 2 |
| Unknown origin | 7 |
| **Durum wheat landrace** | |
| Fertile Crescent (Turkey) | 93 |
| Fertile Crescent (Southern Levant, Lebanon-Syria-Jordan-Israel-Palestine-Iraq) | 83 |
| Fertile Crescent (Cyprus) | 17 |
| Fertile_Crescent (general, not detailed) | |
| Northern Africa (Egypt-Lybia-Tunisia-Algeria-Morocco) | 137 |
| Eastern Africa (Ethiopia-durum landraces) | 172 |
| Eastern Africa (Ethiopia-T. *aethiopicum*) | 14 |
| Eastern Europe (Romania-Bulgaria) | 7 |
| Eastern Europe (Russian Federation-Ukraine) | 53 |
| Central Asia (Kazhakstan-Uzbekistan) | 7 |
| Southern Europe (Greece-Albania-Croatia-Macedonia-Malta-Serbia-Italy-Spain-Portugal) | 157 |
| Western Asia-Transcaucasia (Armenia-Georgia-Azerbaijan) | 20 |

| | |
|---|---|
| Southern Asia (Iran-India) | 29 |
| Eastern Asia (China) | 3 |
| North America (USA-Canada) | 10 |
| Unknown origin | 17 |
| **Durum wheat cultivars** | |
| CIMMYT | 48 |
| ICARDA | 83 |
| Southern Europe (Italy-Spain) | 140 |
| Northern Africa (Morocco-Algeria) | 17 |
| Northern America (Canada-North Dakota) | 46 |
| Northern America (Desert Durum, California-Arizona) | 10 |
| Western Europe (Austria-France) | 45 |
| South America (Argentina) | 5 |
| Ethiopia | 24 |
| Australia-New Zealand | 6 |
| Unknown | 1 |
| ***T. turgidum* subsp. *turgidum*** | 8 |
| ***T. turgidum* subsp. *turanicum*** | 74 |
| ***T. turgidum* subsp.polonicum** | 22 |

Overall, 2,558 tetraploid wheat accessions were used to produce the genotyping based on the Illumina iSelect 90K SNP genotyping platform. The panel contained additional 490 accessions from the main wheat domestication and cultivation areas like Fertile Crescent, The Mediterranean basin, Western Asia and Eastern Africa. The germplasm accessions were provided by AgriBio (Australia), CREA (Consiglio per la ricerca in agricolture e l'analisi dell'economia agraria, Italy), University of Bologna (Italy), University of Saskatchewan (Canada), USDA-ARS (United States Department of Agriculture – Agricultural Research Service), Dr. Benjamin Kilian (Global Crop Diversity Trust, Germany) and Dr. Hakan Ozkan (Cukurova University, Turkey) and by the U.S. National Plant Germplasm System.

Further, these accessions were filtered based on the passport information (i.e. accession name or international code) and genetic similarity ≥ 0.95, leaving only 1 representative accession from the group. Therefore, the final tetraploid wheat germplasm collection (GTC) contained 1,856 accessions.

The tetraploid germplasm collection resulted in total of 23,862 SNPs. Using the following three criteria 17,340 informative SNPs were selected for the further analyses: i) SNPs that had a matching genetic and physical positions were retained, ii) SNPs of null allele frequency ≤ 0.25 were

retained to minimize the ascertainment bias effects, iii) singleton or double singleton SNPs were removed. Thus, 17,340 SNPs represented the first set of SNPs used to detect the signatures of selection from wild emmer wheat to domesticated wheat. Whereas, the second set of 5,774 SNPs, which were LD-pruned ($r^2 > 0.5$) and filtered for minor allele frequency (MAF 0.02), was used to study the population genetic structure.

### 5.1.2. Population genetic structure of GTC

The population genetic structure of 1,856 accessions with 5,774 non-redundant SNPs was studied using several methods:

- Neighbor Joining (NJ) phylogenetic trees;
- Model and non-model based clustering analyses such as DAPC and ADMIXTURE;
- Hierarchical population structure at multiple levels using Fst and Nei's genetic distances.

Neighbor Joining (NJ) phylogenetic trees

We constructed the phylogenetic neighbor joining tree for 1,856 accessions in R using package *ape* aimed to study the evolution and phylogenetic relationships (Paradis et al., 2004). We calculated the genetic distances using the *dist.gene* function. We performed the 1,000 bootstrapping while calculating the genetic distances using the *boot.phylo* function of the *ape* package. All the results were then saved in nexus format using *write.nexus* function for further inspection.

Cluster analysis with ADMIXTURE and DAPC

ADMIXTURE is a model-based algorithm used to estimate the ancestry relationship of individuals. ADMIXTURE uses the same likelihood model used in structure, but is much faster compared to it (Alexander et al., 2009). We explored the clustering using ADMIXTURE for a number of K groups ranging from 2-20. The ADMIXTURE analysis was run based on 100 replications at different random seeds and with 10 cross-validations for a number of populations *K* ranging from 2-20. For each K the replicate with the highest log-likelihood was considered. When performing the ADMIXTURE analysis, cross validation (CV) values showed a continuous decrease with *K*, indicating the presence of a complex population structure, as expected. The ADMIXTURE results were retained for detailed investigation of the relationship among and within the subspecies and populations in the tetraploid wheat germplasm.

After obtaining the global population structure representations, ADMIXTURE analysis was carried out on the four principal germplasm groups separately based on the results on the number of K groups (K = 2-20).

A non-model based approach like Discriminant Analysis of Principal Components (DAPC) was used to characterize the population structure (Patterson et al., 2006; Jombart et al., 2010; Jombart and Collins, 2015). In order to perform the DAPC analysis, we need to identify the optimum number of clusters (k) useful to describe the data first. To do this we used the *find.clusters* function from the R package *adegenet* (Jombart et al., 2010; Jombart and Collins, 2015). We used the K-means method and run the procedure sequentially at increasing values of k from 2 to 20 and computed the BIC (Bayesian Information Criterion) statistics to measure the goodness of fit at each k with the following parameters: n.pca = 1500 (number of retained PC's), stat = "BIC", n.iter = 1000 (number of iterations). The optimal cluster number was evaluated from the BIC plot. Further, the *dapc* function was implemented to describe the genetic diversity between the clusters by retaining the 50 principal components and 7 discriminant functions (n.pca = 50, n.da = 7). In order to choose the number of retained PCs and discrimination functions we used the cross-validation procedure and the maximum of the α-score (*a.score*, the difference between the proportion of successful reassignment of the analysis: observed discrimination, and the values obtained using random groups: random discrimination). Also, the *dapc* function provided the membership probabilities of each individuals to belong to the different clusters. Additionally, the Ward clustering method was used as a valuable alternative to K-means clustering method to explore the grouping of accessions based on the same number of principal components and discriminant functions analysis. While K-means clustering is based on using a random centroid as a start, Ward's is based on the hierarchical clustering.

Hierarchical population structure at multiple levels

We performed a detailed investigation on the genetic relationship among and within the groups of subspecies and populations. We used the population structure identified in ADMIXTURE for this analysis.

- Two main WEW populations:
  - North Eastern Fertile Crescent (WEW-NE)
  - Southern Levant (WEW-SL)
- Six main DEW populations subdivided into Northern and Southern:
  - Northern FC
    - Turkey-to-Transcaucasia/Iran (DEW-T-TRC-IRN)

        - Turkey-to-Balkans (DEW-T-BLK)
    - Southern FC
        - Southern Europe (DEW-Sth-EU)
        - Southern Levant-to-Europe1 (DEW-SL-EU1)
        - Southern Levant-to-Europe2 (DEW-SL-EU2)
    - Indian, Omani and Ethiopian (DEW-ETH)
- Six main DWL populations:
    - Northern FC
        - Turkey-to-Fertile Crescent (DWL-T-FC)
        - Turkey-to-Transcaucasia (DWL-T-TRC)
    - Southern Levant FC
        - Southern Levant-to-North Africa (DWL-SL-NA)
        - Greece-to-Balkans (DWL-GRC-BLK)
    - Ethiopian Landraces (DWL-ETH)
    - *Triticum turanicum* (DWL-TRN)
- Five main DWC branched from North African Landraces
    - ICARDA dryland (DWC-DRY)
    - Italian germplasm (DWC-ITLY)
    - CIMMYT germplasm released in the '70's (DWC-CIM70)
    - CIMMYT germplasm released in the '80's (DWC-CIM80)
    - Germplasm adapted to Mediterranean environment (DWC-AMR)

In order to minimize the confounding effect of recent admixture, accessions with substantial admixture were removed and only accessions with $Q > 0.5$ for WEW and DEW and $Q > 0.4$ for DWL/DWC were retained (Luo et al., 2007).

We tested the hierarchical level of differentiation between the populations using the R package *hierFstat* (Goudet and Jombart, 2015) at three levels:

- level 1, among four main germplasm groups WEW, DEW, DWL and DWC;
- level 2, among domestication origins within subspecies/groups, including North East Fertile Crescent (NE), Southern Levant (SL), Ethiopia (ETH) ;

– level 3, among 19 populations within origins and subspecies: WEW-NE, WEW-SL, DEW-T-TRC-IRN, DEW-T-BLK, DEW-Sth-EU, DEW-SL-EU1, DEW-SL-EU2, DEW-ETH, DWL-SL-NA, DWL-GRC-BLK, DWL-T-TRC, DWL-T-FC, DWL-TRN, DWL-ETH, DWC-DRY, DWC-ITLY, DWC-CIM70, DWC-CIM80, DWC-AMR.

Using the above-described structures of population, we computed pairwise Fst values and Nei's genetics distances among populations and used the *boot.ppFst* function to calculate the 1,000 bootstrap upper and lower limits. We also calculated the expected heterozygosity within populations.

### 5.1.3. Diversity reduction and selection signals associated to main domestication and improvement factors

We used the three main cross-population transitions in order to study the genetic diversity reduction and detect the selection signals associated with wild emmer domestication and durum wheat evolution and breeding:

- WEW-DEW
- DEW-DWL
- DWL-DWC

We calculated the genetic Diversity Reduction Indexes (DRI) for the three transitions. Additionally, we assessed for the presence of selection signatures and their co-occurrence with the diversity reduction using four indexes: i) divergence index, Fst (Weir and Cockerham, 1984) ii) divergence on site frequency spectrum measured through cross-population composite likelihood score, XP-CLR (Chen et al., 2010) iii) haplotype-based metrics such as cross-population extended homozygosity, XP-EHH (Sabeti et al., 2007) iv) and the haplotype-based FLK test, hapFLK (https://github.com/bcm-uga/SSMPG2017/tree/master/Presentations/hapflk, Bonhomme *et al.*, 2010; Fariello *et al.*, 2013).

We used 17,340 SNPs for this analysis. We applied smoothing methods to reduce the erratic signals present for each SNP. For DI, Fst, DRI, hapFLK we used a rolling mean of 25-SNP window with single SNP step. Smoothing using the rolling mean had two advantages: the constant number of markers in the rolling window and managing the irregular marker density per Mb, particularly in the pericentromeric regions.

Highly admixed accessions, Ethiopian origin accessions and those that grouped with Ethiopian accessions were not included in this analysis. As a result, for the three cross-populations we end up with 104 accessions for WEW, 248 for DEW, 591 for DWL and 394 for DWC.

For diversity index (D) and DRI distributions the top and bottom 2.5 percentiles were defined as outliers. For Fst, XP-CLR, XP-EHH, hapFLK distributions instead the top 5.0 percentiles were considered. Adjacent outlier windows interrupted by one or few SNPs in less than 10 Mb distance were merged in one single window. Based on the evidence of strong and extended signals detected in the centromeric regions, the peri-centromeric regions were masked from the distributions and 2.5 and 5.0 percentile distributions were then re-calculated (Hufford et al., 2012).

Diversity Reduction Index

The cross-population Diversity Reduction Index (*DRI*) was calculated according to the following formula:

$$DRI = \frac{D_{wild} + 0.1}{D_{derived} + 0.1}$$

The constant value was added to cope with regions extremely low in diversity. A DRI of 2 means that the diversity in derived germplasm is half of that in the wild. Moreover, we computed DRI with 100,000 permutations in order to identify the thresholds and outliers. The calculations were done with a custom script in R produced by the author.

Fst

The F-statistics was computed for each locus according to Weir and Cockerham (Weir and Cockerham, 1984) using R package *pegas* (Paradis, 2010). We computed Fst with 100,000 permutations carried out in R custom script.

XP-CLR

XP-CLR is one of the powerful indexes in detecting the ancient sweeps (Chen et al., 2010). XP-CLR is based on the likelihood of multi-locus allele frequency distribution modelling between two populations. XP-CLR copes with the ascertainment bias, as it appears to be less sensitive to it. In the current analysis, we used 0.5 cM sliding window with 50kb steps across the whole genome. The results were smoothed over 1 Mb size intervals.

XP-EHH

The haplotype-based metric XP-EHH was estimated using software Selscan (Szpiech and Hernandez, 2014). The method is based on Extended Haplotype Homozygosity (EHH) and measures the reduction in haplotype diversity in cross-population comparisons. We normalized the XP-EHH values as recommended in the software manual. Then, from the normalized values, we got absolute

values. The normalized XP-EHH values were averaged across 50kb windows. The results were smoothed over 1 Mb size intervals.

HapFLK

We used fastPHASE v1.4.8 (Scheet and Stephens, 2006) and *R* package *imputeq* (Khvorykh, 2018) to reconstruct the haplotypes from SNP data in order to identify the optimum number of haplotype clusters. We created 5 test sets using *imputeq* and imputed the genotypes with fastPHASE using the following parameters:

```
fastPHASE -T10 -C25 -K{5:25} -H-1 -n –Z
```

Where, the main parameters are:

-T10, is the number of random starts of the Expectation-Maximization (EM) algorithm

-C25, is the number of EM iterations

-K, is the number of haplotype clusters. We imputed the genotypes at ranges of clusters K from 5, 10, 15, 20 and 25.

-H-1 -n –Z are the additional recommended tags (for details fastPHASE manual, http://scheet.org/code/fastphase_doc_1.4.pdf).

Further, we estimated imputation errors using *EstimateErrors()* function implemented in *imputeq*. The optimum *K* is the one that minimizes the error. Finally, HapFLK (Bonhomme et al., 2010; Fariello et al., 2013) was computed individually on each chromosome on all three sets of cross-populations using the following parameters: *K* (number of haplotypes) = 10, and nfit (number of iterations) = 50.

Genetic diversity investigation in the Global Tetraploid wheat Collection (GTC)

We investigated the genetic diversity of SNPs based on the Nei's genetic diversity, D (Nei, 1973) calculated using the following formula:

$$D = 1 - \sum_{i=1}^{k} p_i^2,$$

where,

$p_i$ = frequency of the $i^{th}$ allele in a locus.

The diversity index (DI) was calculated for 17,340 informative SNPs for all four principal germplasm groups (WEW, DEW, DWL, DWC). The SNP-wise DI values were further smoothed using the window of 25 SNPs with single SNP step. Moreover, we identified the top and bottom 2.5% quantile distributions based on the strong signal evidences.

Domestication genes

We selected a set of 41 wheat-cloned loci associated with domestication or improvement-selection from the literature (Table 10). We located these genes on the Svevo genome (BLASTX with E-value $10^{-10}$) and compared with the evidences of putative selection signals for three cross-populations. QTLs in the tetraploid wheat for the following trait categories, domestication, grain yield, phenology, disease resistance were also considered for evidences of overlap with the selection signals. QTLs were then projected on the relevant transition plots based on the correspondence with the subspecies/group of the crossing parents.

Table 10. Cloned genes relevant to durum wheat breeding, and based on literature known to be under selection during domestication, subspeciation or breeding with their position on the Svevo genome.

| **Locus acronym** | **Locus name** | **Chr** | **Sequence start** | **Sequence end** | **Reference** |
|---|---|---|---|---|---|
| Glu-A3 | Glutenins | chr1A | 5.047.553 | 5.048.723 | (Zhang et al., 2004) |
| TaSUT1A | Sucrose transporter | chr1A | 194.456.769 | 194.515.110 | (Aoki et al., 2002) |
| Glu-A1 | Glutenins | chr1A | 500.859.392 | 501.060.448 | (Salmanowicz and Dylewicz, 2007; Xu et al., 2008) |
| T6P | Trehalose-6-phosphate synthase | chr1A | 520.686.209 | 520.692.550 | (Xie et al., 2015) |
| ELF3-A1 | Early flowering 3 | chr1A | 582.981.365 | 582.985.067 | (Alvarez et al., 2016; Zikhali et al., 2016) |
| TaSUT1B | Sucrose transporter | chr1B | 230.604.772 | 230.650.472 | (Aoki et al., 2002) |
| ELF3-B1 | Early flowering 3 | chr1B | 676.974.612 | 676.978.342 | (Alvarez et al., 2016; Zikhali et al., 2016) |
| Ppd-A1 | Photoperiod respose | chr2A | 36.577.899 | 36.565.231 | (Wilhelm et al., 2009) |
| TaSus2-2A | Sucrose synthase | chr2A | 120.335.255 | 120.340.200 | (Jiang et al., 2011; Hou et al., 2014) |
| TaSdr-A1 | Seed dormancy | chr2A | 156.408.483 | 156.409.475 | (Zhang et al., 2014) |
| TaCwi-A1 | Cell wall invertase | chr2A | 501.893.554 | 501.897.261 | (Ma et al., 2012; Jiang et al., 2015) |
| Ppd-B1 | Photoperiod respose | chr2B | 56.297.789 | 56.294.941 | (Wilhelm et al., 2009; Takenaka |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | and Kawahara, 2012) |
| TaSus2-2B | Sucrose synthase | chr2B | 169.016.255 | 169.020.790 | (Jiang et al., 2011; Hou et al., 2014) |
| TaSdr-B1 | Seed dormancy | chr2B | 198.376.152 | 198.377.132 | (Zhang et al., 2014) |
| TaCwi-B1 | Cell wall invertase | chr2B | 439.343.754 | 439.347.239 | (Ma et al., 2012; Jiang et al., 2015) |
| BRT-3A | Brittle rachis | chr3A | 61.344.533 | 61.345.121 | (Avni et al., 2017) |
| BRT-3B | Brittle rachis | chr3B | 96.155.280 | 95.381.784 | (Avni et al., 2017) |
| Rht-A1 | Reduced height | chr4A | 575.088.221 | 575.090.083 | (Pearce et al., 2011) |
| Phs-A1 | Seed dormancy | chr4A | 598.755.842 | 598.762.987 | (Shorinola et al., 2016) |
| Rht-B1 | Reduced height | chr4B | 29.292.990 | 29.294.855 | (Pearce et al., 2011) |
| HMA3-A1 | Heavy metal ATPase | chr5A | 542.961.581 | 542.964.488 | (Maccaferri *et al.*, unpublished) |
| VRN-A1 | Vernalization | chr5A | 549.152.139 | 549.156.384 | (Yan et al., 2003) |
| Q-5A | Domestication | chr5A | 608.796.291 | 608.792.747 | (Zhang et al., 2011) |
| HMA3-B1 | Heavy metal ATPase | chr5B | 563.900.691 | 563.903.585 | (Maccaferri *et al.*, unpublished) |
| VRN-B1 | Vernalization | chr5B | 570.831.391 | 570.844.281 | (Chu et al., 2011) |
| Q-5B | Domestication | chr5B | 650.078.209 | 650.075.235 | (Zhang et al., 2011) |
| Phs-B1 | Seed dormancy | chr5B | 698.826.783 | 698.832.510 | (Shorinola et al., 2016) |
| Gli | Alpha-gliadins | chr6A | 24.341.990 | 24.342.853 | (Gu et al., 2004) |
| NAC-A1 | NAC domain-containing protein | chr6A | 75.453.416 | 75.454.973 | (Uauy et al., 2006) |
| TaGW2-A | Grain weight | chr6A | 235.270.703 | 235.295.537 | (Zhang et al., 2018) |
| Sr13-6A | Stem rust resistance | chr6A | 611.710.263 | 611.713.775 | (Zhang et al., 2017) |
| NAC-B1 | NAC domain-containing protein | chr6B | 130.826.078 | 130.826.755 | (Uauy et al., 2006) |
| TaGW2-B | Grain weight | chr6B | 300.791.272 | 300.808.374 | (Zhang et al., 2018) |
| Sr13-6B | Stem rust resistance | chr6B | 689.235.987 | 689.239.462 | (Zhang et al., 2017) |
| VRN-A3 | Vernalization | chr7A | 69.364.420 | 69.367.738 | (Yan et al., 2006) |
| TaTGW-7A | Grain weight | chr7A | 204.055.853 | 204.061.744 | (Hu et al., 2016) |
| TaCML20 | Calmodulin 20 | chr7A | 686.342.874 | 686.348.391 | (Kalaipandian et al., 2018) |
| VRN-B3 | Vernalization | chr7B | 9.128.364 | 9.124.817 | (Yan et al., 2006) |
| TaTGW-7B | Grain weight | chr7B | 168.949.495 | 168.955.358 | (Hu et al., 2016) |
| TaCML20 | Calmodulin 20 | chr7B | 663.786.935 | 663.788.698 | (Kalaipandian et al., 2018) |
| Psy-B1 | Phytoene synthase | chr7B | 714.361.446 | 714.362.281 | (Zhang and Dubcovsky, 2008) |

Multiple and overlapping selection signals with selection signal peaks located within 10 Mb or less were considered to be selection signal clusters, both within transition and across transitions.

## 5.2. Results

### 5.2.1. Population genetic structure of Global Tetraploid wheat Collection

Global Tetraploid Wheat Collection (GTC) was categorized in four principal germplasm groups: wild emmer wheat –WEW, domesticated emmer wheat – DEW, durum wheat landraces – DWL and durum wheat modern cultivars – DWC. We used a high quality Illumina iSelect 90K SNP genotyping platform to determine the genetic diversity of the germplasm collection (Wang et al., 2014; Maccaferri et al., 2015). There is a possible SNP-ascertainment bias present due to usage of a fixed SNP platform (Qanbari and Simianer, 2014; Malomane et al., 2018). A relatively wide discovery panel of both hexaploid and tetraploid wheats were used to develop and ascertain 90K wheat SNP platform (Wang et al., 2014). The SNPs that were ascertained in the A and B genomes of hexaploid wheat contributed a relatively non-biased representation of allele frequencies in the ancestral tetraploid A and B genome. This is probably explained by the gene flow occurred between wild and cultivated wheats. This feature was not observed for hexaploid wheat genome D (Haudry et al., 2007; Akhunov et al., 2010), which suggests the occurrence of the limited loss of diversity in the A and B genomes during the *T. aestivum* evolution from domesticated tetraploid wheat and diploid *A. taucshii*. As a consequence, using the 90K assay to build a consensus map for the tetraploid wheat, the SNP composition ascertained in the wheat 90K assay allowed to genetically map a high number of functional and evenly distributed SNPs in the mapping populations obtained from the ancestral tetraploids (WEW, DEW) × modern durum crosses (e.g. Svevo × Zavitan DWC × WEW mapping populations and three additional DEW × modern durum mapping populations). Due to use of 90K assay and mapping populations, the tetraploid consensus map had even marker density and genome coverage. Therefore, the wheat 90K is a valuable SNP array for mapping and genetic diversity studies in the whole tetraploid wheat genome.

Neighbor joining (NJ) phylogenetic trees

We assessed the phylogeny and population genetics structure using a representative set of LD-pruned 5,775 SNPs that had an $r^2 = 0.50$. The Neighbour Joining tree shows the genetic relationships among taxa and populations. The NJ tree analysis evidenced that the four main germplasm groups (WEW, DEW, DWL, DWC) appear clearly differentiated, suggesting strong and founder effects and little evidence for phylogenetic origin (Figure 22).

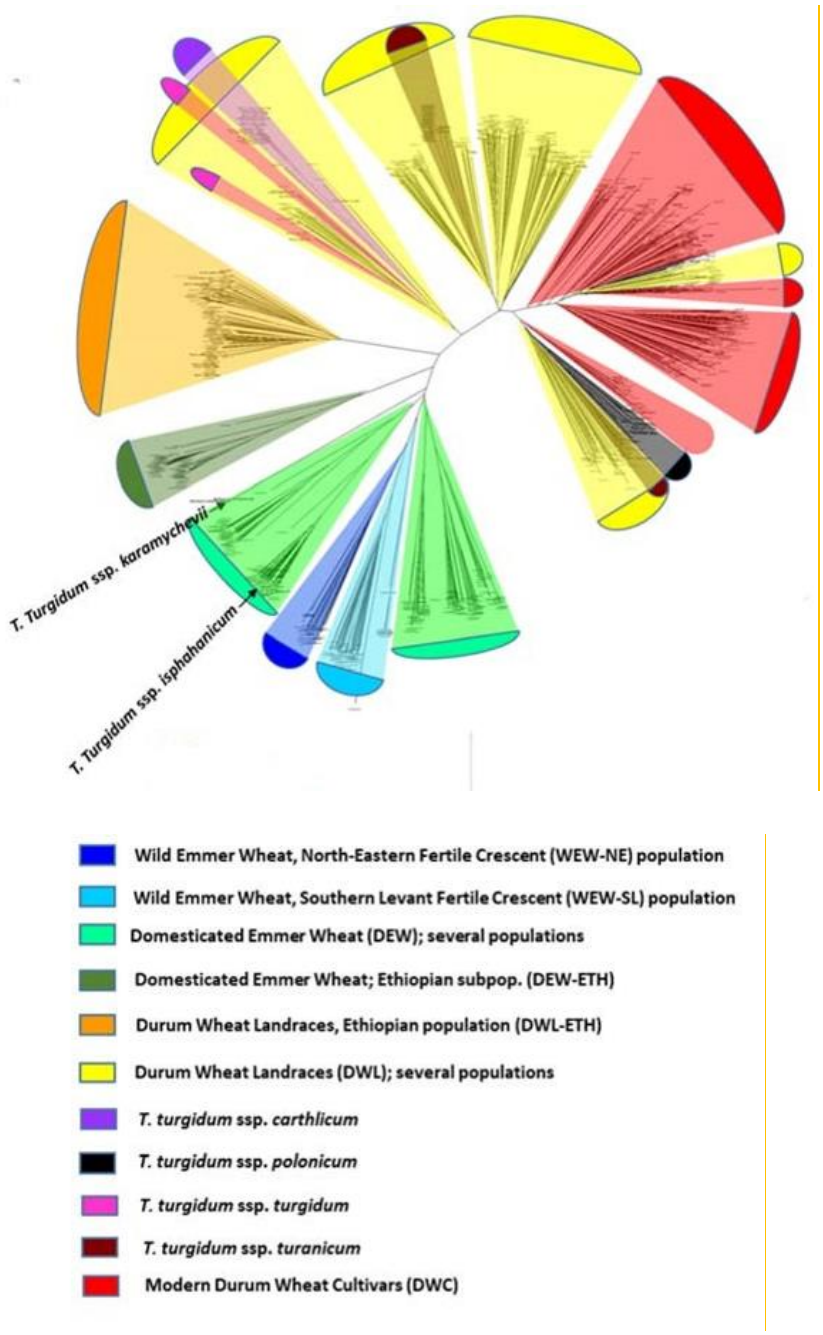



Figure 22. The Neighbour Joining tree from Nei's genetic distance that shows the genetic relationships among taxa and populations estimated at 1000 bootstrap. Four main germplasm groups were evidenced (WEW, DEW, DWL, DWC). The Neighbour Joining tree for A. WEW, B. DEW, C. DWL, D. DWC was computed as well.

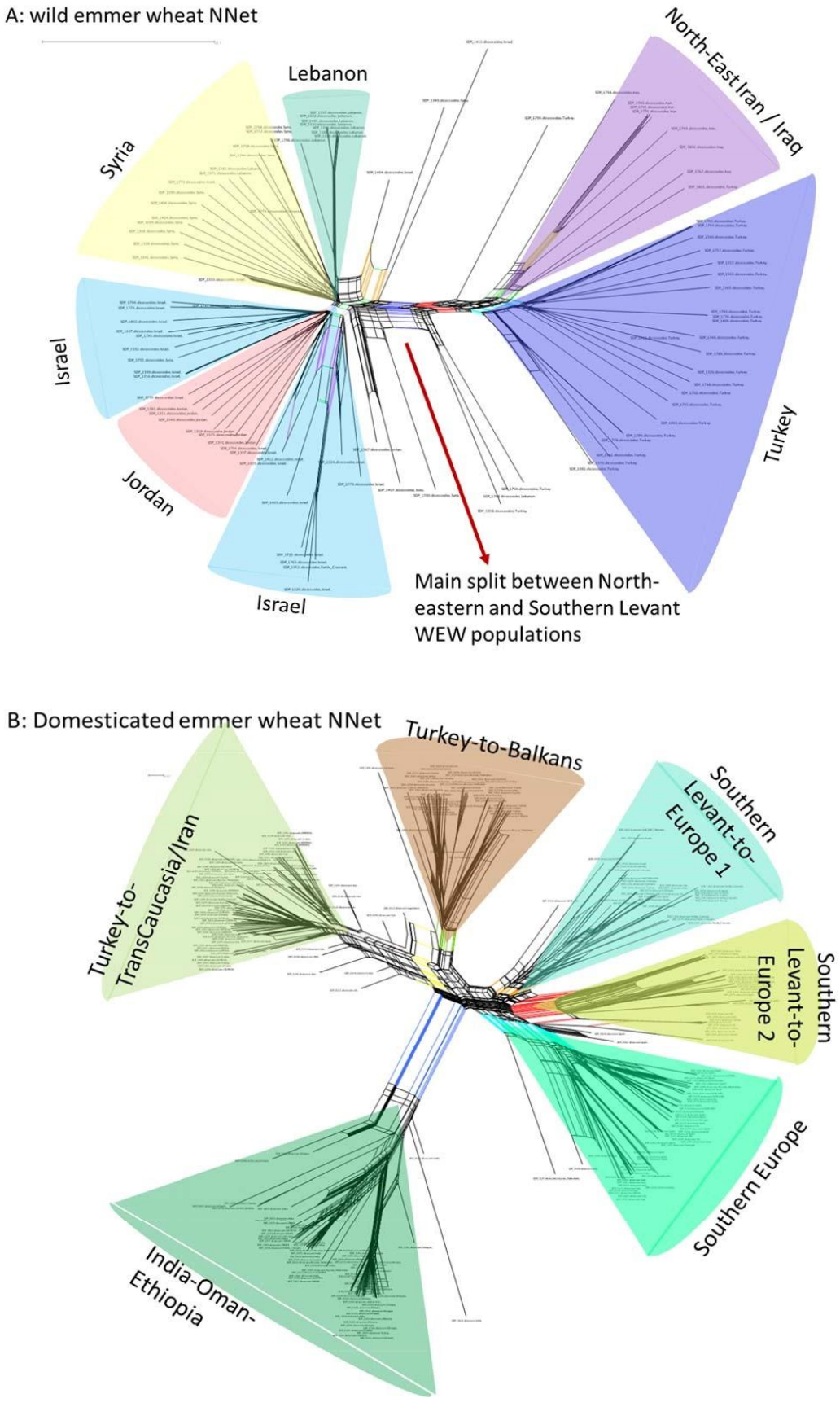



Figure 22. Continued.

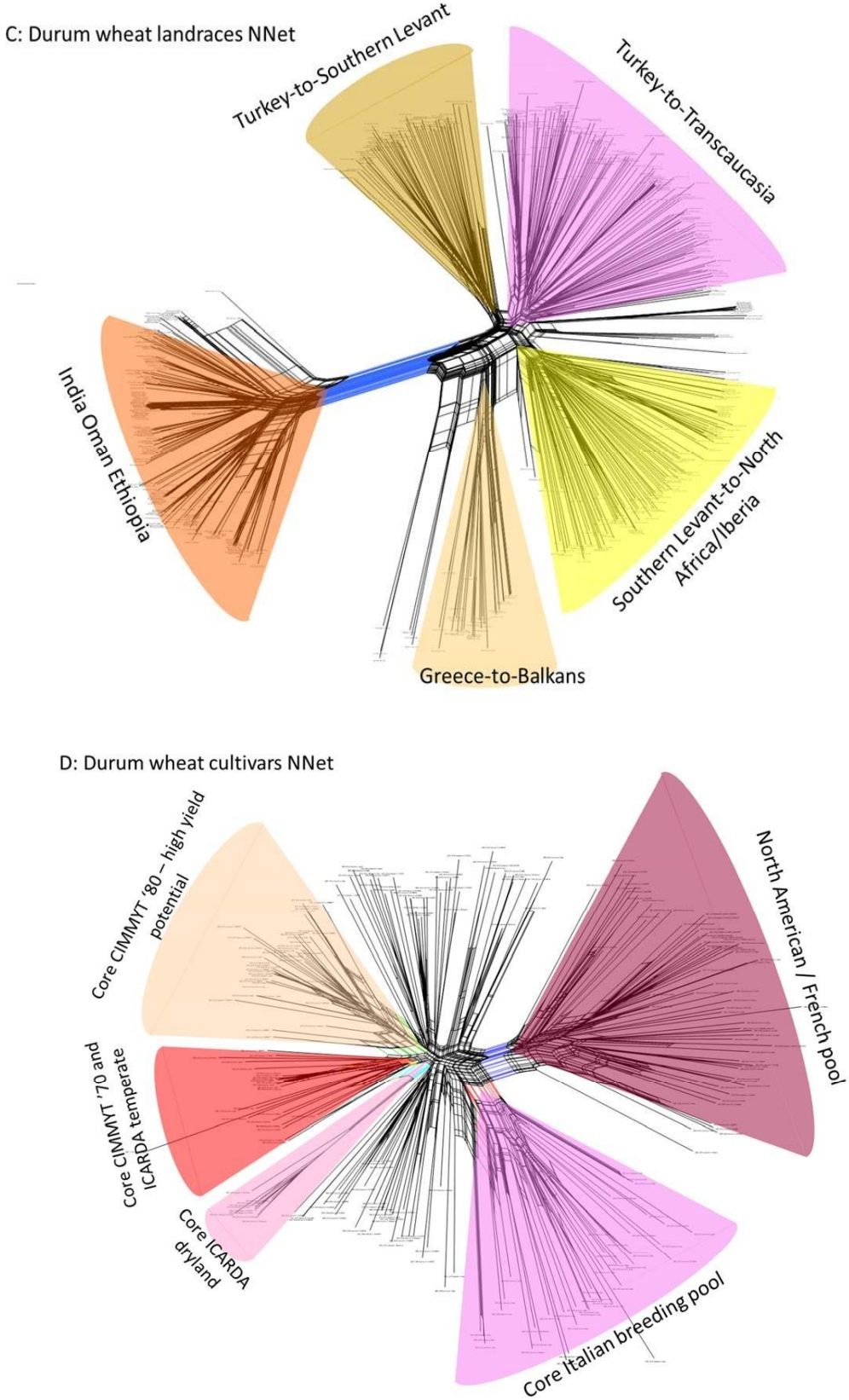



Figure 22. Continued.

Cluster analysis with ADMIXTURE and DAPC

The two independent ADMIXTURE and DAPC non-hierarchical clustering methods showed concordant and overall similar population structure representations.

The accessions classified into clusters by DAPC (both DAPC K-means and DAPC Ward's methods) allowed less quantitative admixture and cross relationship compared to ADMIXTURE cluster classification. Therefore, ADMIXTURE resulted to be relatively more informative and cross comparable than DAPC. Moreover, the correlation value between the two DAPC methods and with ADMIXTURE was relatively low (0.59). Therefore, based on these results the ADMIXTURE cluster classification was considered for further investigation at single taxon level.

The tetraploid wheat germplasm is highly structured and it is known that a well-defined population structure cannot be captured by a single K value. At lower K values WEW and DEW occurred to be highly structured based on main population structures and at higher K values additional well-defined populations emerged, mostly related to the geographical origin of the accessions. In case of DEW and DWL, well-defined populations are due to human-driven dispersal processes. In the DWL germplasm, admixture among the main populations brought by the Mediterranean cross exchange resulted to be an important component of the diversity.

ADMIXTURE analysis assumes a Hardy-Weinberg equilibrium, nonetheless, the results obtained by this analysis was highly valuable for the structure description of tetraploid wheat germplasm. While accurately assessing the admixture events the analysis could capture majority of the geographical based structures. We detected distinct population structure differentiation from K 2 to 10 and substantial relevant differentiations in agreement with passport-pedigree information, taxonomy and geographical area of origins were detected at higher K values (up to 20). At K = 20, 1,053 (56.58 %) accessions had membership Q values > 0.7, and 1,440 (77.38 %) accessions had Q values > 0.5. At K = 2, WEW, DEW, *T. turgidum* ssp. *carthlicum* and the Ethiopian DWL formed first group and, *T. turgidum* ssp. *durum* landraces, cultivated accessions and other taxa formed a second group. At K = 3 the Ethiopian DWL clearly separated from all the other accessions and at K = 17 subdivided further in two different populations. DEW differentiated from WEW at K = 5.

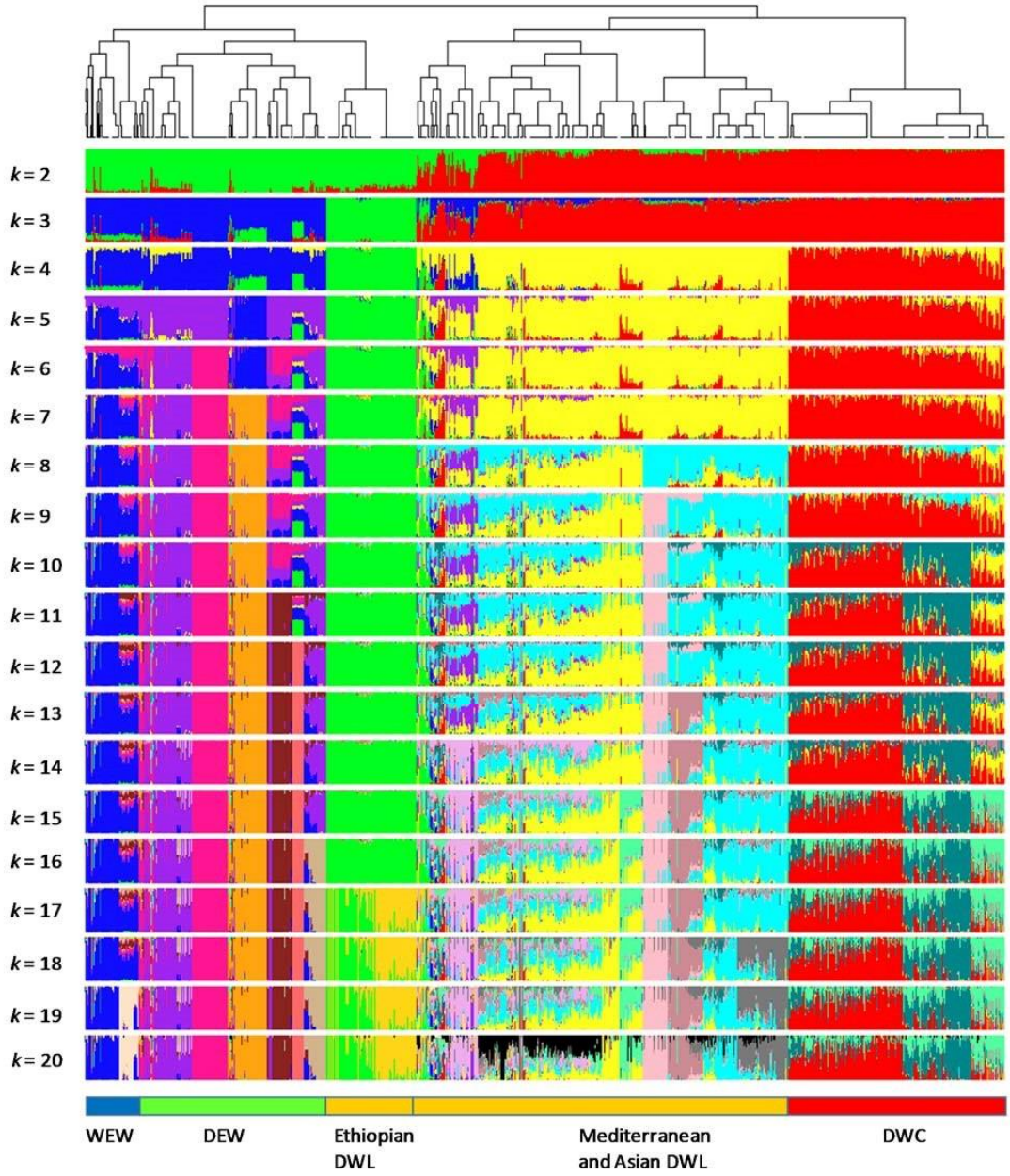



Figure 23. Population structure of the tetraploid diversity panel as assessed globally by ADMIXTURE. WEW: wild emmer wheat, DEW: domesticated emmer wheat, DWL: durum wheat landraces, DWC: durum wheat cultivars.

Western and Eastern DEW populations separated at K = 6, Ethiopian and Indian DEW separated at K = 7. DEW was consistently structured into five populations at K = 12 to 20. At K from 4 to 13 DEW germplasm clearly separated into five groups: i) Western populations from Fertile Crescent to Europe; ii) Western populations majority from Fertile Crescent with relation to European accessions; iii) Western populations from Turkey to Balkans and Russia; iv) Eastern population including Iran, Transcaucasia, Russia and Asia; v) Eastern population including Ethiopia and India.

The obtained results are highly similar to the results obtained from previous diversity studies (Badaeva et al., 2005). The. *T. turgidum* ssp. *carthlicum* clearly separated at K = 12 (Figure 23).

DWL from different origins and other durum wheat related *Triticum* taxa mainly separated from the modern elite durum wheat germplasm at K 4. At higher K values the DWL germplasm subdivided further to Western-Eastern spread and dispersal passages associated to humankind migration and trade. This confirms further observations in emmer and hexaploid wheats (Badaeva et al., 2005; Dubcovsky and Dvorak, 2007). At K = 6-8 the separation between Western (Mediterranean) and Eastern (Asian continental) populations became evident. One of the *T. turgidum* ssp. *turanicum* populations separated from DWL at K = 10, while the other classified with DWL. At K = 14 Western Mediterranean DWL classified to three populations originated from: i) Greece to Balkans; ii) Fertile Crescent including Southern Levant, Cyprus to North Africa and Iberian peninsula; iii) North Africa, Egypt to Morocco, to Iberian peninsula with relations to a group of *T. turgidum* ssp. *turanicum* accessions. Two populations from Russia and Turkey, Transcaucasia and Asia were included as Western Continental landraces. Although very little evidence of cross-talk events was observed between the germplasms of DWL and DEW, DWL resembled the genetic and geographical differentiation evidence similar to DEW.

DWC mainly grouped in three-five populations that correspond to three main germplasm pools bred worldwide:

- The first pool consisted of cultivars bred at CIMMYT (The International Maize and Wheat Improvement Center), or in Mediterranean breeding programs that rely on semi-dwarf or photoperiod insensitive CIMMYT germplasm;
- The second pools consisted of the cultivars bred in France and Austria and the North American germplasm (Canada and Northern USA);
- The third pool consisted of germplasm from ICARDA (The International Center for Agricultural Research in the Dry Areas), mainly originated from crosses between the native North African and Syrian landraces and modern semi-dwarf cultivars, and germplasm locally bred in Mediterranean countries (i.e. Italy).

In general, both durum wheat landraces and modern cultivars showed a greater admixture compared to DEW and WEW. This result was anticipated based on history of durum wheat breeding and studies of Mediterranean durum wheat landrace pool (Maccaferri et al., 2003, 2005; De Vita et al., 2007; Oliveira et al., 2014; Soriano et al., 2016; Kabbaj et al., 2017). Detailed ADMIXTURE results for WEW, DEW, DWL and DWC is illustrated in Figure 24.

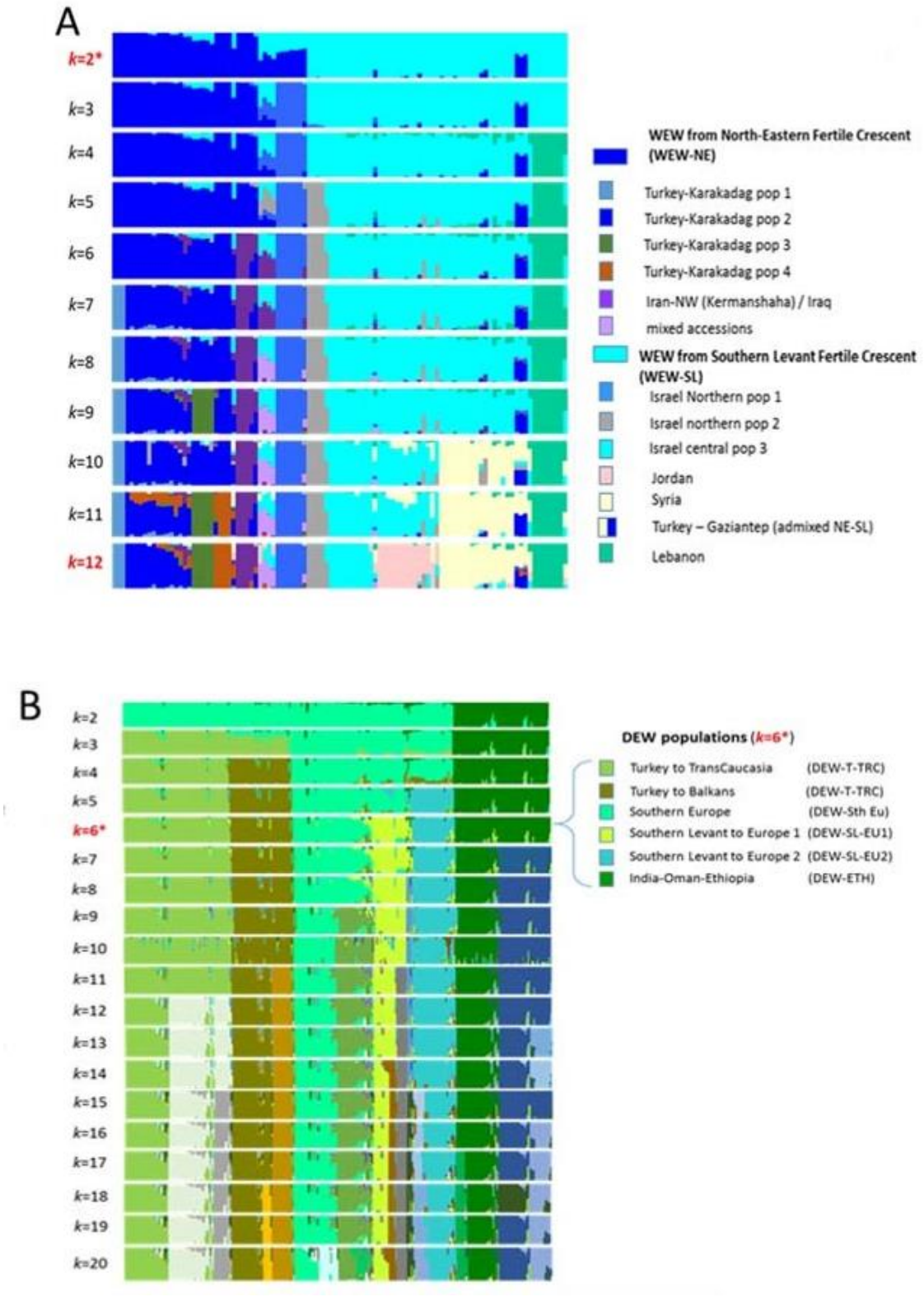



Figure 24. ADMIXTURE analysis results of wild emmer wheat (WEW) accessions with *K* from 2 to 12 (A), domesticated emmer wheat (DEW) accessions with *K* from 2 to 20 (B), durum wheat landrace (DWL) accessions with *K* from 2 to 12 (C), and durum wheat cultivars (DWC) with *K* from 2 to 5 (D). Results are represented as bar plots of *Q* membership coefficients.

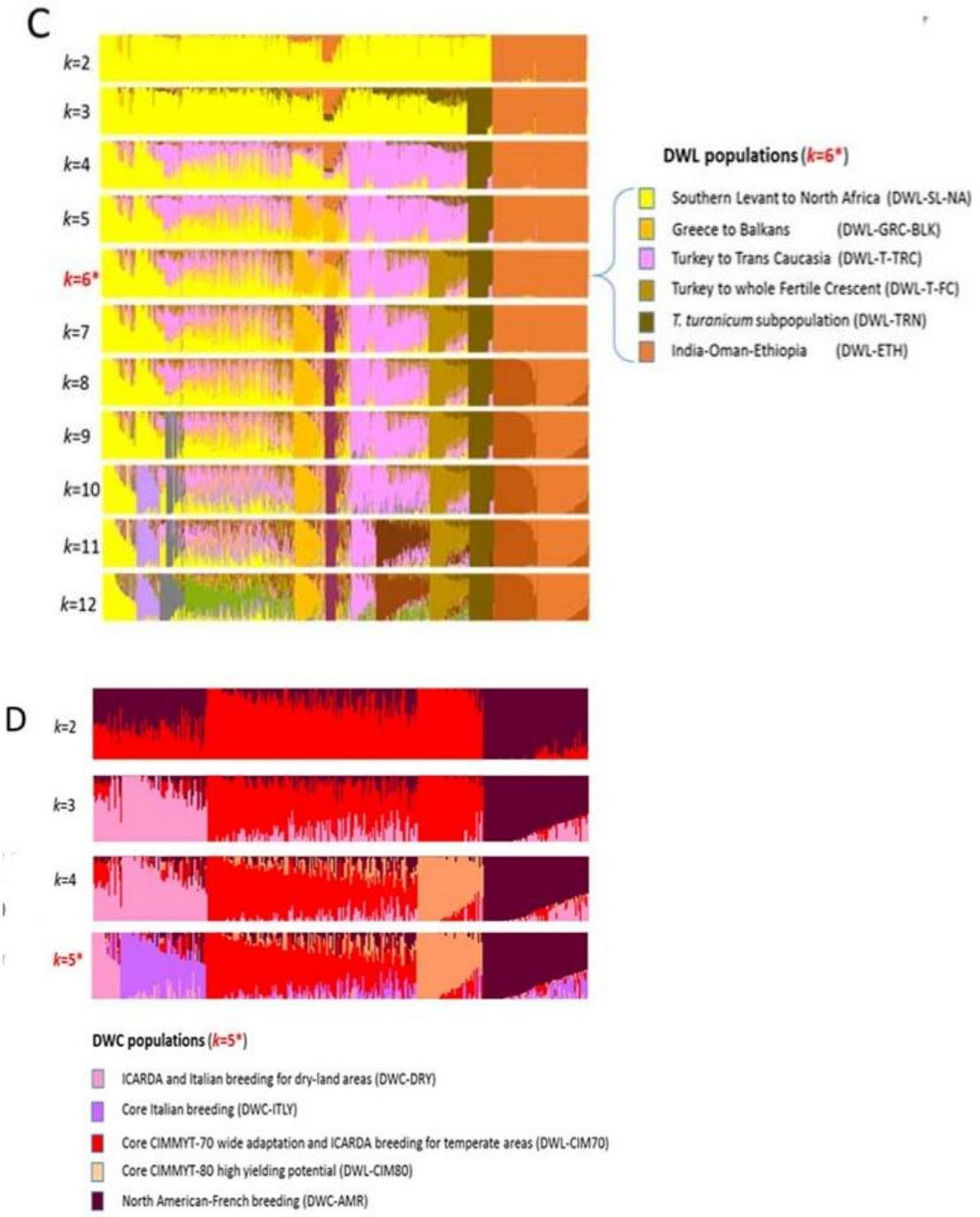



Figure 24. Continued.

Out of 1,856 accessions, only 1,755 were assigned to populations based on Q membership score and were used to further study the population structure in details. The detailed view of population structure and genetic relationships among the populations was obtained by running ADMIXTURE and NJ analysis within each germplasm group separately (Figure 24). The results obtained using both methods were similar. However, ADMIXTURE carried out at increasing K number of populations showed the most probable relationships among taxa and germplasm

populations at the historical level. WEW and DEW in contrast to drum wheat, demonstrated a highly structured genetic diversity; a high rate of population assignment at the increased K value of K = 12 and K = 20 for WEW and DEW, respectively. The WEW germplasm subdivided in two main populations such as from North-Eastern Fertile Crescent and Southern Levant (WEW-NE and WEW-SL, respectively). WEW-NE divided further into distinct populations such as Turkey, Iran and Iraq, and WEW-SL divided into populations from Israel, Jordan, Syria and Lebanon. These results were consistent both in ADMIXTURE and NJ analysis and in previous phylogenetic analysis conducted at lower density WEW with molecular markers (Ozkan et al., 2005; Luo et al., 2007).

DEW and DWL showed similar Northern-to-Southern Fertile Crescent and Eastern-to-Western radial dispersal patterns. DEW separated in six major populations:

- Two from the Northern Fertile Crescent, Turkey-to-Transcaucasia/Iran (DEW-T-TRC-IRN), Turkey-to-Balkans (DEW-T-BLK);
- Three from Southern Levant: Southern Europe (DEW-Sth-EU), Southern Levant-to-Europe1 (DEW-SL-EU1), Southern Levant-to-Europe2 (DEW-SL-EU2);
- One Indian, Omani and Ethiopian DEW (DEW-ETH).

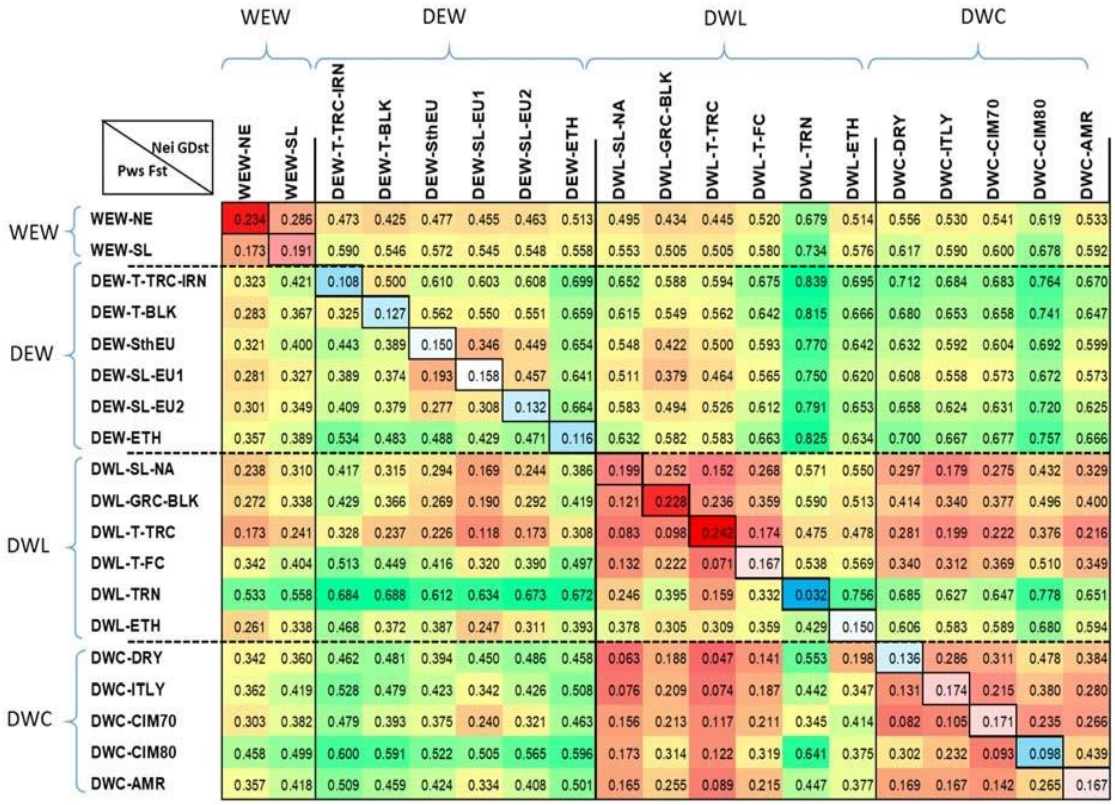
Similar results were obtained for population structure analysis in worldwide emmer (Badaeva et al., 2015) and our results further supported the evidence that emmer germplasm is delineated into four main subgroups based on geographical factors, namely Europeans, Balkans, Asians and Ethiopians. Six main populations emerged also for DWL:

- Two populations from the Northern Fertile Crescent, Turkey-to-Fertile Crescent (DWL-T-FC), Turkey-to-Transcaucasia (DWL-T-TRC);
- Two from the Southern Levant, Southern Levant-to-North Africa (DWL-SL-NA), Greece-to-Balkans (DWL-GRC-BLK);
- Two highly distinct populations consisting of the Ethiopian landraces (DWL-ETH) and the *T. turanicum* (DWL-TRN) accessions.

All of the cultivated durum accessions group into a germplasm that represent a wide branch of the durum North African landrace pool.

We investigated further the genetic relationship among and within the main tetraploid taxa and population after removing the accessions that showed high admixture level both across taxa and across populations within taxa. We studied the population differentiation using hierarchical ANOVA by computing the pairwise Fst and Nei's genetic distance among and within the populations (Figure

25). The results supported the Northern-to-Southern Fertile Crescent and Eastern to Western radial dispersal pattern and phylogeny.



| | Nei GDst / Pws Fst | WEW-NE | WEW-SL | DEW-T-TRC-IRN | DEW-T-BLK | DEW-SthEU | DEW-SL-EU1 | DEW-SL-EU2 | DEW-ETH | DWL-SL-NA | DWL-GRC-BLK | DWL-T-TRC | DWL-T-FC | DWL-TRN | DWL-ETH | DWC-DRY | DWC-ITLY | DWC-CIM70 | DWC-CIM80 | DWC-AMR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| WEW | WEW-NE | 0.234 | 0.286 | 0.473 | 0.425 | 0.477 | 0.455 | 0.463 | 0.513 | 0.495 | 0.434 | 0.445 | 0.520 | 0.679 | 0.514 | 0.556 | 0.530 | 0.541 | 0.619 | 0.533 |
| | WEW-SL | 0.173 | 0.191 | 0.590 | 0.546 | 0.572 | 0.545 | 0.548 | 0.558 | 0.553 | 0.505 | 0.505 | 0.580 | 0.734 | 0.576 | 0.617 | 0.590 | 0.600 | 0.678 | 0.592 |
| DEW | DEW-T-TRC-IRN | 0.323 | 0.421 | 0.108 | 0.500 | 0.610 | 0.603 | 0.608 | 0.699 | 0.652 | 0.588 | 0.594 | 0.675 | 0.839 | 0.695 | 0.712 | 0.684 | 0.683 | 0.764 | 0.670 |
| | DEW-T-BLK | 0.283 | 0.367 | 0.325 | 0.127 | 0.562 | 0.550 | 0.551 | 0.659 | 0.615 | 0.549 | 0.562 | 0.642 | 0.815 | 0.666 | 0.680 | 0.653 | 0.658 | 0.741 | 0.647 |
| | DEW-SthEU | 0.321 | 0.400 | 0.443 | 0.389 | 0.150 | 0.346 | 0.449 | 0.654 | 0.548 | 0.422 | 0.500 | 0.593 | 0.770 | 0.642 | 0.632 | 0.592 | 0.604 | 0.692 | 0.599 |
| | DEW-SL-EU1 | 0.281 | 0.327 | 0.389 | 0.374 | 0.193 | 0.158 | 0.457 | 0.641 | 0.511 | 0.379 | 0.464 | 0.565 | 0.750 | 0.620 | 0.608 | 0.558 | 0.573 | 0.672 | 0.573 |
| | DEW-SL-EU2 | 0.301 | 0.349 | 0.409 | 0.379 | 0.277 | 0.308 | 0.132 | 0.664 | 0.583 | 0.494 | 0.526 | 0.612 | 0.791 | 0.653 | 0.658 | 0.624 | 0.631 | 0.720 | 0.625 |
| | DEW-ETH | 0.357 | 0.389 | 0.534 | 0.483 | 0.488 | 0.429 | 0.471 | 0.116 | 0.632 | 0.582 | 0.583 | 0.663 | 0.825 | 0.634 | 0.700 | 0.667 | 0.677 | 0.757 | 0.666 |
| DWL | DWL-SL-NA | 0.238 | 0.310 | 0.417 | 0.315 | 0.294 | 0.169 | 0.244 | 0.386 | 0.199 | 0.252 | 0.152 | 0.268 | 0.571 | 0.550 | 0.297 | 0.179 | 0.275 | 0.432 | 0.329 |
| | DWL-GRC-BLK | 0.272 | 0.338 | 0.429 | 0.366 | 0.269 | 0.190 | 0.292 | 0.419 | 0.121 | 0.228 | 0.236 | 0.359 | 0.590 | 0.513 | 0.414 | 0.340 | 0.377 | 0.496 | 0.400 |
| | DWL-T-TRC | 0.173 | 0.241 | 0.328 | 0.237 | 0.226 | 0.118 | 0.173 | 0.308 | 0.083 | 0.098 | 0.242 | 0.174 | 0.475 | 0.478 | 0.281 | 0.199 | 0.222 | 0.376 | 0.216 |
| | DWL-T-FC | 0.342 | 0.404 | 0.513 | 0.449 | 0.416 | 0.320 | 0.390 | 0.497 | 0.132 | 0.222 | 0.071 | 0.167 | 0.538 | 0.569 | 0.340 | 0.312 | 0.369 | 0.510 | 0.349 |
| | DWL-TRN | 0.533 | 0.558 | 0.684 | 0.688 | 0.612 | 0.634 | 0.673 | 0.672 | 0.246 | 0.395 | 0.159 | 0.332 | 0.032 | 0.756 | 0.685 | 0.627 | 0.647 | 0.778 | 0.651 |
| | DWL-ETH | 0.261 | 0.338 | 0.468 | 0.372 | 0.387 | 0.247 | 0.311 | 0.393 | 0.378 | 0.305 | 0.309 | 0.359 | 0.429 | 0.150 | 0.606 | 0.583 | 0.589 | 0.680 | 0.594 |
| DWC | DWC-DRY | 0.342 | 0.360 | 0.462 | 0.481 | 0.394 | 0.450 | 0.486 | 0.458 | 0.063 | 0.188 | 0.047 | 0.141 | 0.553 | 0.198 | 0.136 | 0.286 | 0.311 | 0.478 | 0.384 |
| | DWC-ITLY | 0.362 | 0.419 | 0.528 | 0.479 | 0.423 | 0.342 | 0.426 | 0.508 | 0.076 | 0.209 | 0.074 | 0.187 | 0.442 | 0.347 | 0.131 | 0.174 | 0.215 | 0.380 | 0.280 |
| | DWC-CIM70 | 0.303 | 0.382 | 0.479 | 0.393 | 0.375 | 0.240 | 0.321 | 0.463 | 0.156 | 0.213 | 0.117 | 0.211 | 0.345 | 0.414 | 0.082 | 0.105 | 0.171 | 0.235 | 0.266 |
| | DWC-CIM80 | 0.458 | 0.499 | 0.600 | 0.591 | 0.522 | 0.505 | 0.565 | 0.596 | 0.173 | 0.314 | 0.122 | 0.319 | 0.641 | 0.375 | 0.302 | 0.232 | 0.093 | 0.098 | 0.439 |
| | DWC-AMR | 0.357 | 0.418 | 0.509 | 0.459 | 0.424 | 0.334 | 0.408 | 0.501 | 0.165 | 0.255 | 0.089 | 0.215 | 0.447 | 0.377 | 0.169 | 0.167 | 0.142 | 0.265 | 0.167 |

Figure 25. Nei's genetic distances, above diagonal, and pairwise Fst, below diagonal, between main tetraploid wheat populations. Diagonal represents expected heterozygosity, values within populations.

As a result, WEW-NE from Turkey, Iran and Iraq resulted to be the most possible ancestor of all the DEW populations and durum wheat germplasm, in contrast to WEW-SL populations that had Fst and Nei's genetic distance values lower for all WEW-DEW and WEW-DW pairs. Additionally, in the DEW-DWL transition, the two *T. turgidum* ssp. *dicoccum* populations from Southern Levant Fertile Crescent that demonstrated a primary relationships with the European accessions (DEW-SL-EU1 and DEW-SL-EU2) and low genetic distance and hence relationship to all DWL populations, except *T. turgidum* ssp. *turanicum* populations.

The modern durum wheat germplasm showed the highest relationship to the two DWL populations from North Africa (DWL-SL-NA) and Turkey to Transcaucasia (DWL-T-TRC). Instead, DWL-GRC-BLK and DWL-T-FC were greatly related to the modern durum varieties bred for the Dryland areas located at ICARDA and to the Italian germplasm adapted to Mediterranean environments. Among DWL populations, the Ethiopian and *T. turanicum* durum populations were

the most differentiated and their contribution to the modern durum wheat germplasm was minimal. The high-yield and successful CIMMYT germplasm released in the 80's (Altar84) resulted to be the most differentiated from all DEW and DWL germplasm pools.

### 5.2.2. Demography and selection signals in the Global Tetraploid wheat Collection

We applied several metrics to determine selection signals. Among them a haplotype-based methods (Qanbari and Simianer, 2014) which reduces the effect of ascertainment bias, population structure and demographic factors. Additionally, a greatly divergent from the rest of the populations, Ethiopian DEW and DWL germplasms were excluded from the selection signal analysis. Initially, the SNP-based diversity index (D) that was averaged over a 10Mb non-overlapping window was used to understand the extent of diversity loss along the three transitions. Further, we expanded the analysis using five different selection signal/divergence signal indexes, including the diversity index, using a rolling mean of 25 SNPs as shown in Figure 26A and detailed in Figure 32. The following four indexes were applied to estimate the selection signatures: i) the diversity reduction index, ii) the divergence estimated using both a single site index (Fst) and a haplotype-based frequency differentiation index, hapFLK, corrected for population structure, iii) the haplotype structure, using the Cross-population Extended Haplotype Homozygosity (XP-EHH), iv) the spatial pattern of site frequency spectrum XP-CLR.

Using the results of selection sweep indexes, various chromosome regions putatively under selection/differentiation sweep were detected. In order to reduce the single cite erratic behavior and uncover the strong selective sweep signals comprising wide genomic regions we applied a rolling mean of 25-SNPs or an average window of 1 Mb to all indexes.

The overlapping regions with two or more indexes showing outlier signals were considered as a single selection region. Hereafter, either all selection regions identified by a one-selection index (singleton) or by multiple selection indexes are referred as unique selection clusters.

In total, we count 454 unique clusters, out of which 104 are pericentromeric and 350 are distal putative selection signal clusters. On average, the pericentromeric clusters had a size of 107.7 Mb (95% size distribution: 2.7 to 369.1 Mb), while the distal clusters had an average size of 11.4 Mb (95% size distribution: 0.37 and 42.2 Mb). The average cluster physical size progressively increased from WEW-DEW cross-comparison to DWL-DWC cross-comparison, from 10.2 to 15.3 Mb for distal regions.

Diversity Reduction Index

Among the four groups of germplasm, WEW demonstrated the highest average gene diversity and a uniformly distributed diversity pattern across the whole genome, except for the pericentromeric regions of chromosomes 2A and 4A that showed a local reduction in diversity (Marone et al., 2012). Therefore, the diversity reduction pattern contributed as a valuable reference when comparing with domesticated/improved germplasm groups DEW, DWL and DWC. We observed a strong depletion in diversity across the genome that occurred independently and consolidated progressively during the crop improvement process in all of three germplasm groups. We observed that the regions where a depletion of diversity took place (during WEW-DEW or DEW-DWL cross-comparisons), no recovery in diversity was observed in the subsequent derived germplasm (Figure 26 B, C, D and Figure 31), apart some few exceptions. At the uttermost of the evolution, domestication and breeding processes, the genome of the elite DWC progressively accumulated a lot near-fixation of diversity regions. Except the pericentromeric region of chromosomes 2A and 3A where there was observed an increased diversity in DWC compared to DWL and DEW. The rolling mean across fixed 25 SNPs for DRI index confirmed the high rate of domestication-related diversity depletions in the pericentromeric regions compared to the distal regions of chromosomes. The chromosome regions with adjacent non-interrupted SNPs that had a DRI value >2 (equivalent to a diversity reduction of 50% or more), were 65 for WEW-DEW accounting for 1,999.2 Mb, 111 for DEW-DWL for 2,138.5 Mb and 75 for DWL-DWC transition accounting for 1,086.6 Mb. As a result, the modern durum germplasm cumulated on average 5 Gb of sequence undergoing less than half diversity compared to the ancestral WEW.

The projection and mapping of the 41 cloned genes known to be under selection during emmer domestication, durum wheat evolution and breeding on the Svevo genome revealed an explanation for several clusters. Most of the strongest, pericentromeric diversity depletions (*DRI* > 4) was already noticed in the first WEW-DEW transition: chromosomes 2A (282.7 Mb), 4A (341.8 Mb) 4B (211.5 Mb), 5A (two regions of 61.4 and 48.4 Mb), 5B (two regions of 24.9 and 144.7Mb) and 6A (334.0 Mb). A distal region on chromosome 5A with DRI > 4 of 5.4 Mb was concurring with the location of a vernalisation gene *VRNA1* (Yan et al., 2003) . Moreover, one of the two brittle rachis loci associated with the early domestication process, harboring *BRT3-B1* at 96.2Mb (Avni et al., 2017) showed a localized reduction in diversity highlighted by *Fst* and *XP-CLR* metrics (Figure 26 B, C, D and Figure 27). The same region, subsequently, underwent a more extreme diversity reduction in the DEW-DWL transition ($DRI_{max}$ = 3.4 in a region of 79.1-125.8 Mb).

The transition from domesticated emmer to durum was spotted by two main depletions ($DRI_{max}$ >4) in pericentromeric regions on chromosomes 1A (one single region of 185 Mb) and 2B

(two regions of 12.5 and 34.9 Mb). We observed plentiful other depletions in the non-pericentromeric regions as well, including the chromosome 2B harboring one of the major *tough-glumes* QTL (*Tg-2B*) governing threshability and marking the emmer to durum transition locus. *Tg-2B* mapped between 31.9 and 36.1 Mb on Svevo genome (Faris et al., 2014b, 2014a) and we observed two severe diversity depletions ($DRI_{max}$ = 4.3) on chromosome 2BS in the following regions: 25.1-26.4 Mb and 33.3-49.1 Mb. The *Tg-2A* homoeolog genetically mapped between 21.2 and 31.7 Mb on Svevo genome, and was associated to the threshing-related traits in the 27.9-32.0 Mb region. The gene *Glu-1*, coding for glutenin subunits and located at 500.8 Mb on chromosome 1A, is reported to be nearly fixed in modern germplasm for null allele *Glu-A1c* (Xu et al., 2008), was associated to a local strong $DRI_{max}$ signal = 3.2 in 8 Mb window. None strong diversity selection signal was associated to the domestication-related *Q-5A* locus positioned at 608.8 Mb, except for a local peak of diversity in DWC. Instead, the *Q-5B* harboring region (Zhang et al., 2011) positioned at 650.1 Mb, showed D, DRI and XP-EHH signals (Figure 29). Further extreme reduction in diversity associated to the DWL-DWC transition was observed on chromosomes 2B, 5B, 6A and 7B. The latter overlaps with several disease resistance (*Lr14a*) and grain yellow pigment content loci, including *Psy-B1*.

Divergence and haplotype based metric signals

We used Fst index to investigate the allele frequency differentiation complemented by XP-EHH, XP-CLR and hapFLK methods that are based on multiple-SNPs linked regions/haplotypes, and therefore more buffered against influence of demography and population structure. Extensive signal of divergence/selection were observed at the pericentromeric regions pointed out by overlapping peaks present in two or more indexes, particularly in WEW-DEW and DEW-DWL cross-population transitions. These results underline and confirm that most of the loss-of-diversity and divergence signatures took place during domestication and selection processes.

Prioritization of selective signatures was managed by selecting the top ranking 1 % distribution and by investigating for co-occurred signal clusters. Out of 454 selection signals identified by at least one metric, 96 were identified by DRI, 184 by Fst 167 by XP-EHH and 153 by XP-CLR. Moreover, 68 DRI signals co-occurred with at least one of the other metrics (71 % of all DRI signals), particularly with Fst index, followed by XP-CLR and XPEHH. These signal clusters, in combination with both diversity reduction and divergence effects, can be prioritized as the most interesting putative selection clusters.

We also compared the occurrence of selection signals and with wheat genes relevant for domestication and improvement. Among a set of 41 previously cloned loci, which are associated with the selection process, many loci co-located with regions where a strong selection occurred. *TaGW2-*

*A1* (Zhang et al., 2018) on chromosome 6A (235.3 Mb) was associated with a selection signal detected by all metrics in the WEW-DEW transition, and was also associated with a sharp decrease of diversity in DEW. A grain weight gene *TaGW2-B1* (Zhang et al., 2018) on chromosome 2B at 300.8Mb coincided with the region with top Fst and hapFLK at WEW-DEW transition and XP-EHH (DEW-DWL) signals. In addition, *TaSus2-A1*, *TaSdr-A1*, and *TaCWI-A1* on chromosome 2A and their homoeologs on chromosome 2B were associated to multiple extended signals in WEW-DEW and in DEW-DWL transitions, while the durum germplasm showed extended regions of low diversity.

Among the loci mapped to non-pericentromeric regions, the following genes were associated with selection signal peaks:

- On chromosome group 3, *BRT-A1* was associated to XP-CLR and hapFLK signals in WEW-DEW transition, while *BRT-B1* was associated to Fst and XP-CLR;

- On chromosome 5B, *Q-5B* was associated to a XP-CLR signal in WEW-DEW transition, although for *Q-5A* there was no evident selection signal found, probably due to interactions with other regulatory elements, such as *miR172*, that could have weakened the signal;

- On chromosome 1A, *Glu-A1* was associated to XP-CLR signal in DWL-DWC transition;

- On chromosome 5A, *VRN-A1* was associated to an Fst signal;

- On chromosome 2A, *Ppd-A1* was associated to XP-EHH signal in DWL-DWC cross-population transition;

- On chromosome group 7, *TaTGW-7A* mapped central to a XP-EHH signal in WEW-DEW transition, and to a region proximal to multiple DRI, Fst, XP-EHH, XP-CLR signals in DEW-WEW transition. The surrounding region had high-depleted diversity in durum. Moreover, *TaTGW-7B* mapped to XP-EHH signal in both DEW-DWL and DWL-DWC transitions and in a DRI region in DWL-DWC;

- On chromosome 7B, the *Psy-B1* region was coincident with two signals: XP-CLR in DEW-DWL and Fst in DWL-DWC transitions;

- On chromosome 4B, there was no single associated with *Rht-B1*, probably because *Rht-B1* has not yet reached fixation in the elite germplasm of durum wheat and the North American germplasm is mostly composed of cultivars of conventional height. Nevertheless, *Rht-B1* gene

mapped closely (<2 Mb) to an extended region with strong increase in diversity in DWC compared to DWL.

A

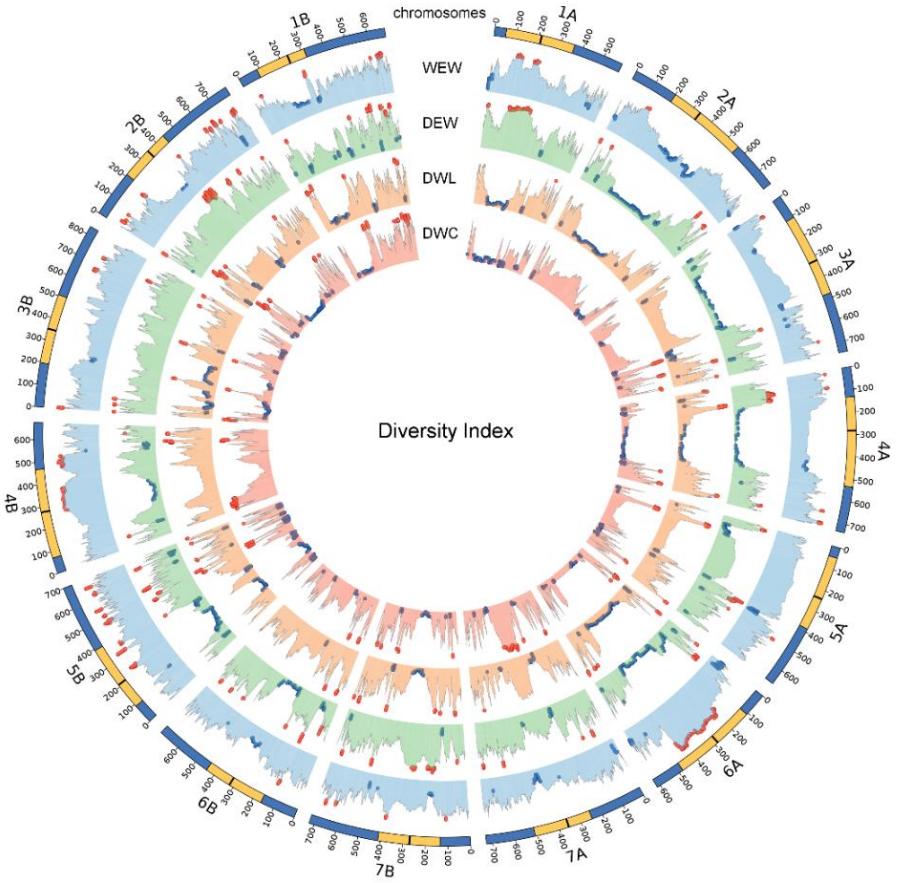
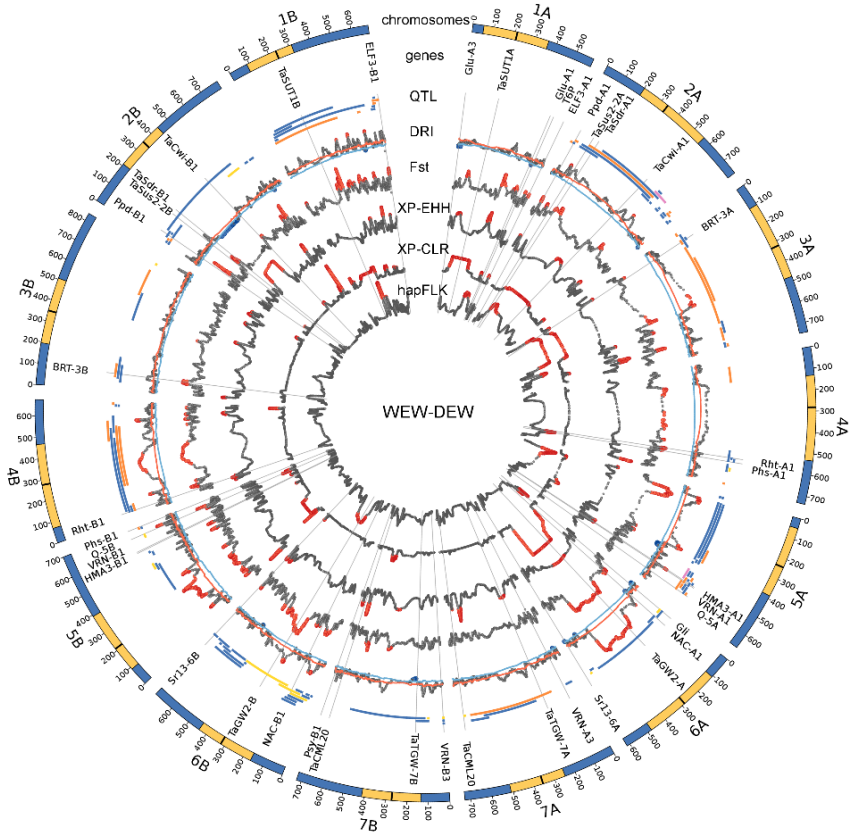



Figure 26. Genome wide analysis of diversity and selection signatures in tetraploid wheat based on 17,340 informative SNPs. (A) SNP-based Diversity Index (DI) for the main germplasm groups identified in the Global Tetraploid wheat Collection: wild emmer wheat (WEW), domesticated emmer wheat (DEW), durum wheat landraces (DWL), and durum wheat cultivars (DWC). DI is reported as a centered 25 SNP-based rolling mean with single SNP step. Top and bottom 2.5% DI quantile distributions are highlighted as red- and blue-filled dots, respectively. (B) Cross-population selection index metrics for the comparison between WEW and DEW. Selection metrics are provided for: Diversity Reduction Index (DRI), divergence index (Fst), cross population Extended Haplotype Homozygosity (XP-EHH), multilocus test for allele frequency differentiation (XP-CLR), and haplotype-based differentiation test (hapFLK). For DRI, top and bottom 2.5% DI quantile distributions are highlighted as red- and blue-filled dots, respectively, while for the other selection metrics top 5% quantile distributions are highlighted as red-filled dots. The physical location of genes (Table 10) and QTLs relevant to domestication and breeding is reported. (C) As in panel B for the comparison between DEW-DWL cross-population. (D) As in panel B for the comparison between DWL-DWC cross-populations.

B




C

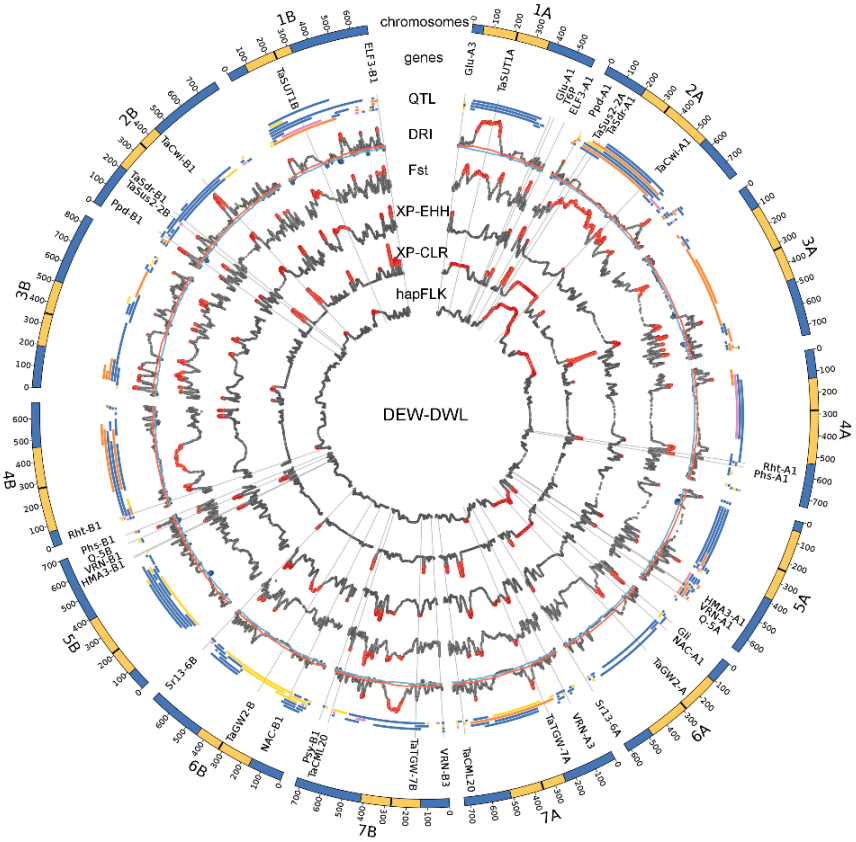
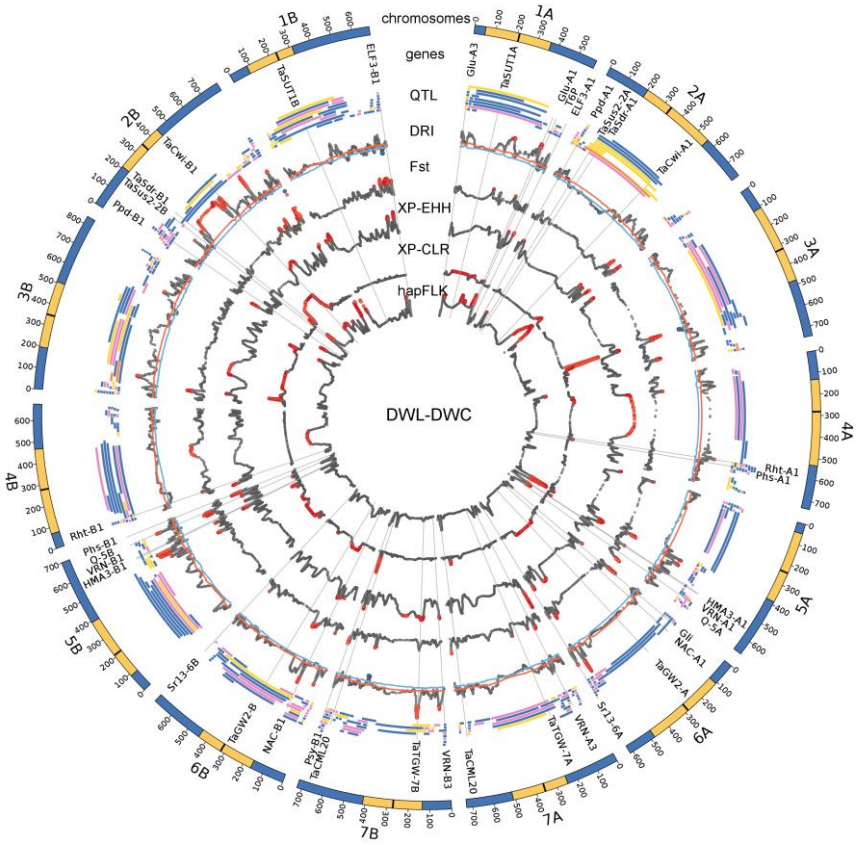



Figure 26. Continued.

D




Genes
*Glu-A3,*Glu-A1 – Glutenins
*Gli – Alpha-gliadins
*BRT-3A,*BRT-3B - Brittle Rachis
*ELF3-A1,*ELF3-B1 - Early flowering 3
*HMA3-A1,*HMA3-B1 - Heavy metal ATPase
*NAC-A1,*NAC-B1 - NAC domain-containing gene
*Phs-A1,Phs-B1,*TaSdr-A1,*TaSdr-B1 - Seed dormancy
*Ppd-A1,*Ppd-B1 - Photoperiod respose
*Psy-B1 - Phytoene synthase
*Q-5A,*Q-5B - Domestication
*Rht-A1,*Rht-B1 - Reduced height
*Sr13-6A,Sr13-6B - Stem rust resistance
*T6P - Trehalose-6-phosphate synthase
*TaCML20-A,TaCML20-B - Calmodulin 20
*TaCwi-A1,TaCwi-B1 - Cell wall invertase
*TaGW2-A,TaGW2-B - Grain weight
*TaSus2-2A,*TaSus2-2B - Sucrose synthase
*TaSUT1A,*TaSUT1B - Sucrose transporter
*TaTGW-7A,TaTGW-7B - Grain weight
*VRN-A1,*VRN-A3,*VRN-B1,*VRN-B3 - Vernalization
*cloned genes

QTLs
Yield
Domestication
Quality
Phenology

Figure 26. Continued.

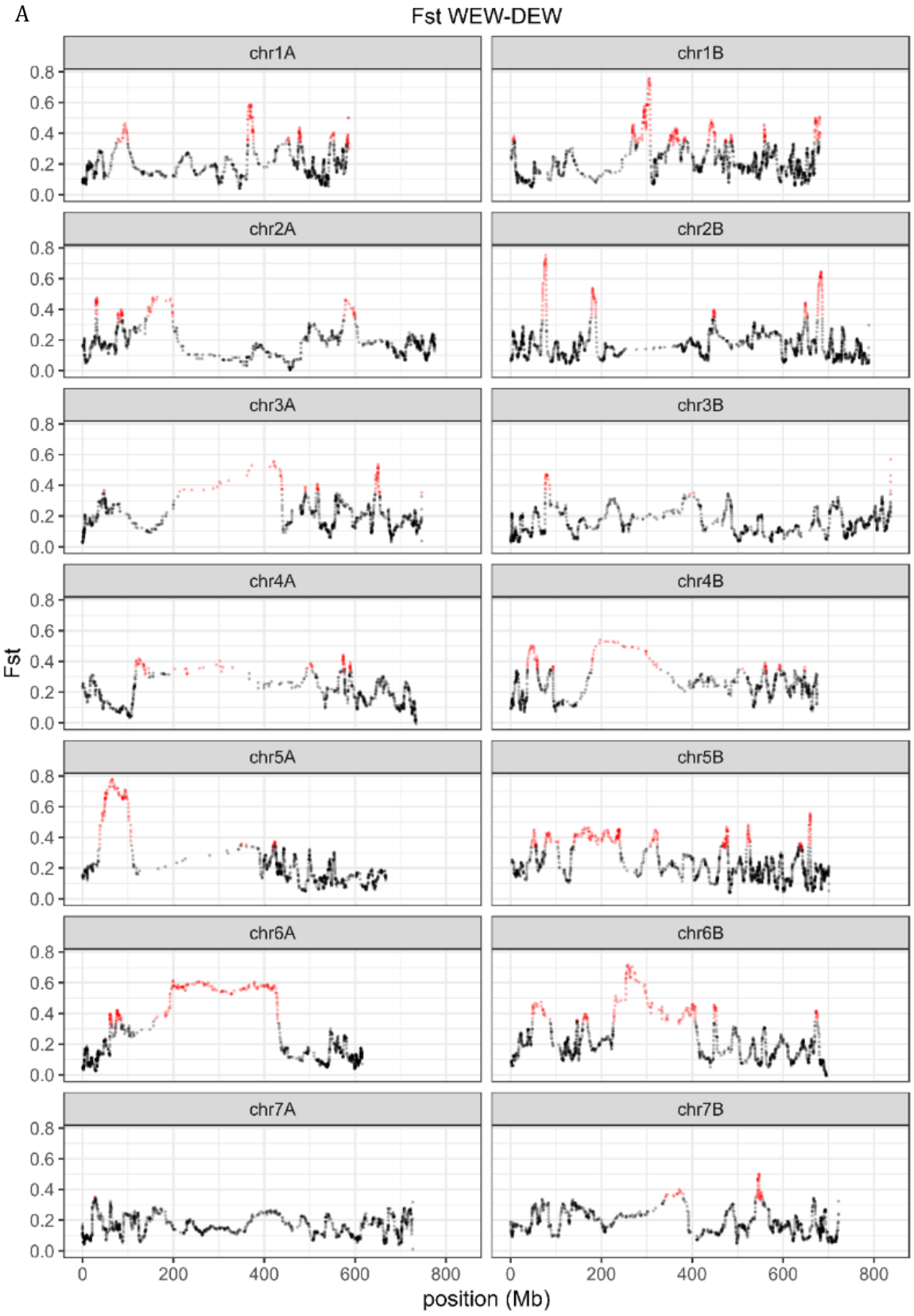



Figure 27. Fst divergence index for A. WEW-DEW, B. DEW-DWL, C. DWL-DWC. Top 5% quantile distributions are highlighted as red-filled dots.

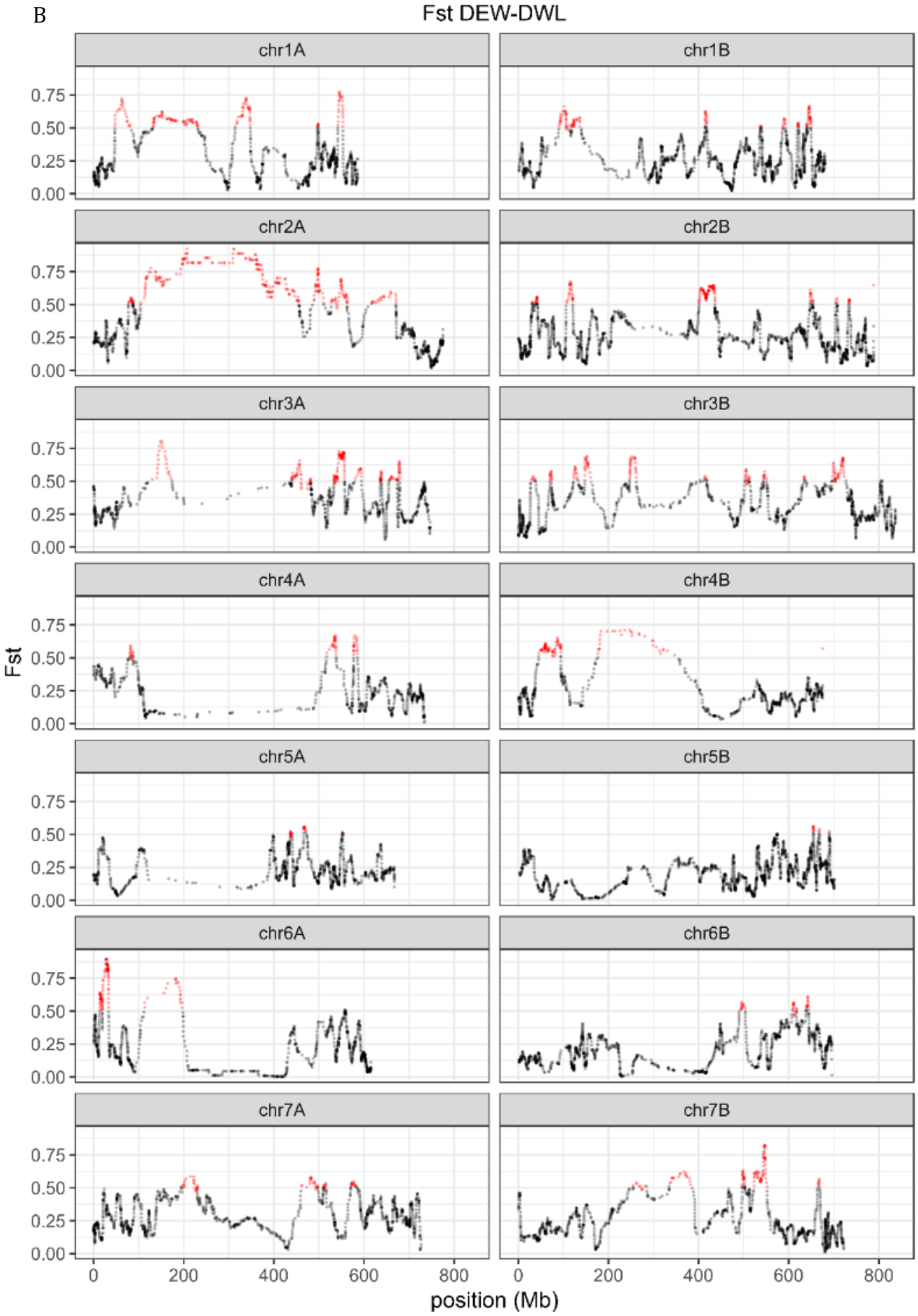



Figure 27. Continued.

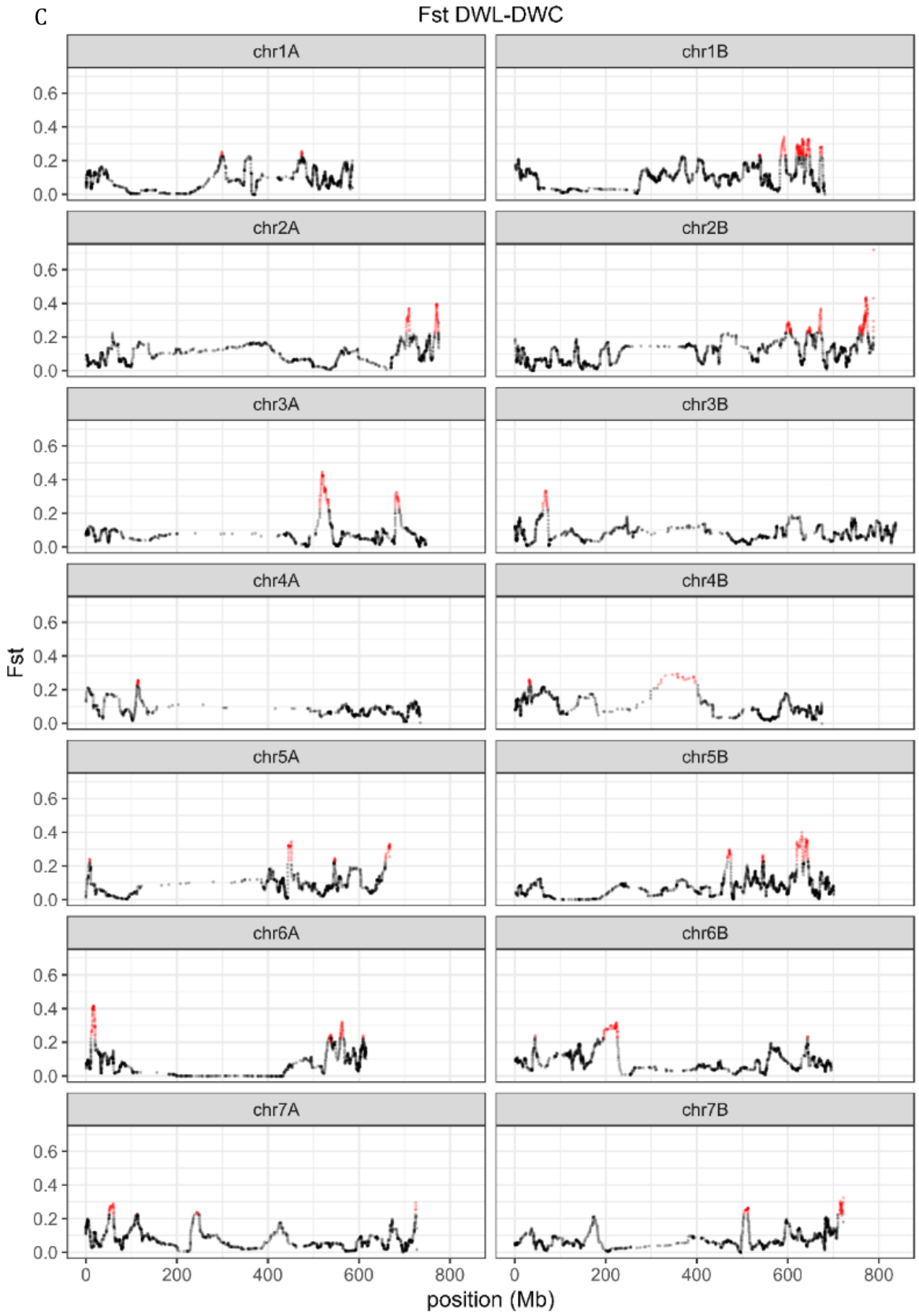



Figure 27. Continued.

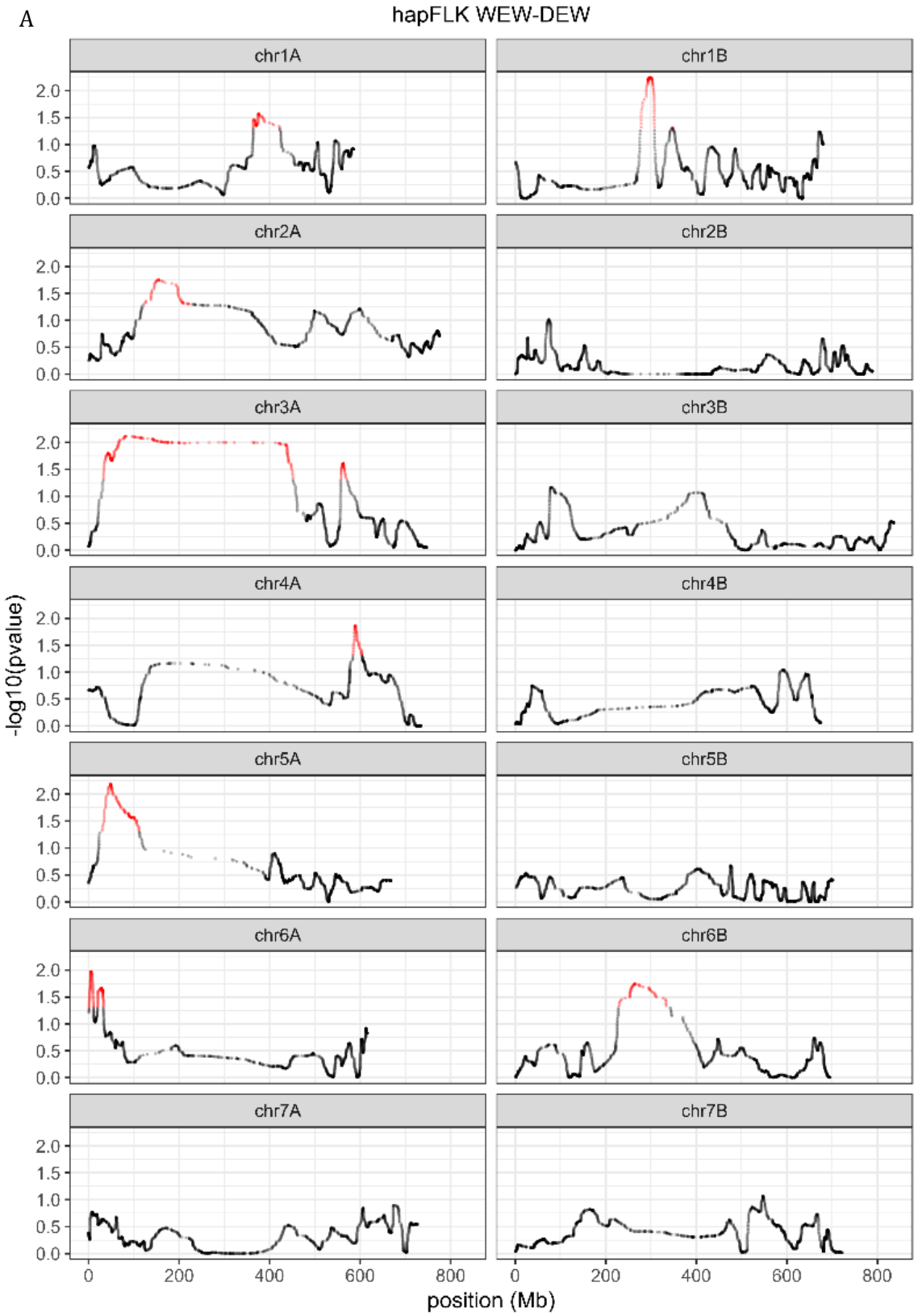



Figure 28. HapFLK haplotype-based metric for A. WEW-DEW, B. DEW-DWL, C. DWL-DWC. Top 5% quantile distributions are highlighted as red-filled dots.

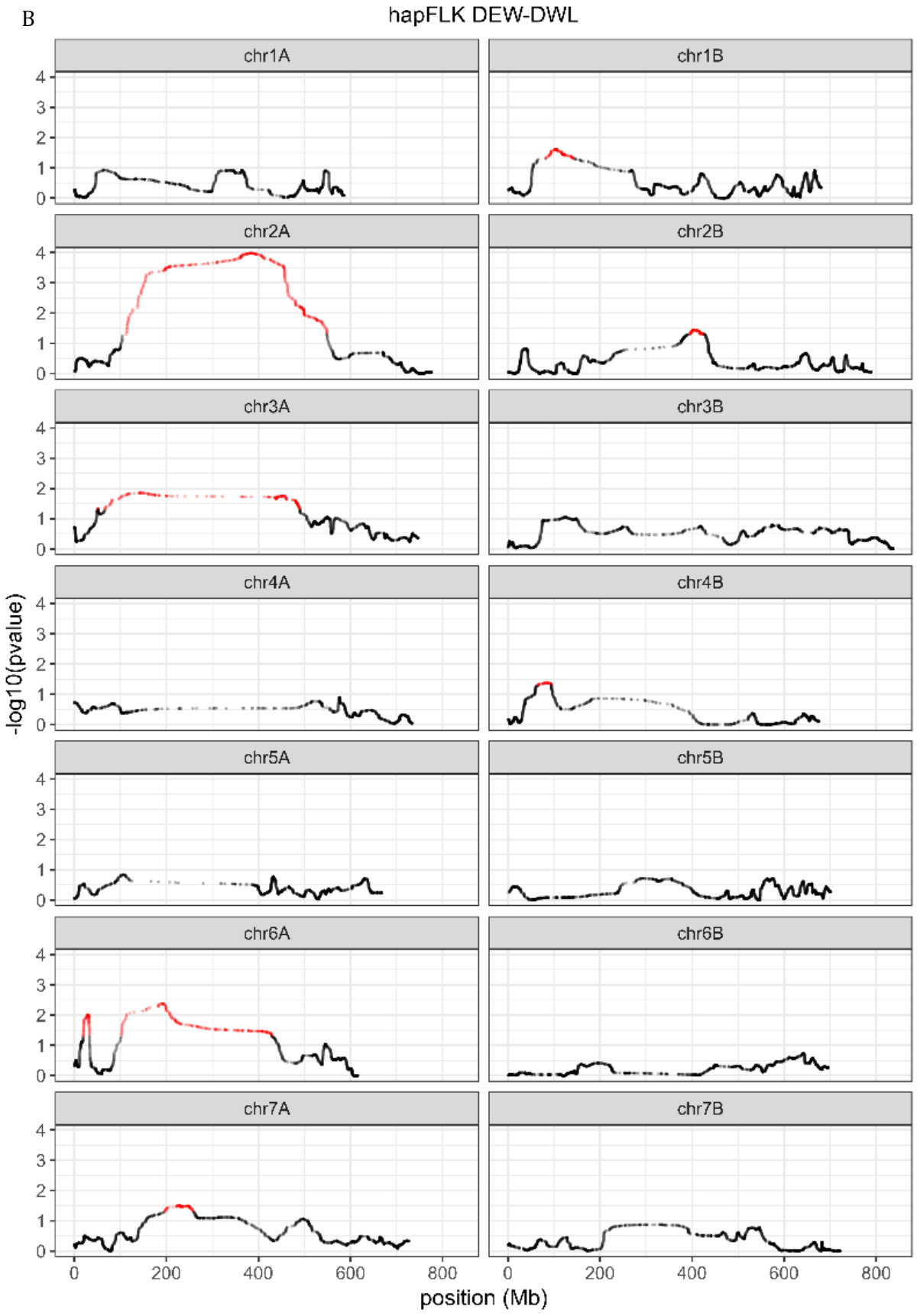



Figure 28. Continued.

C

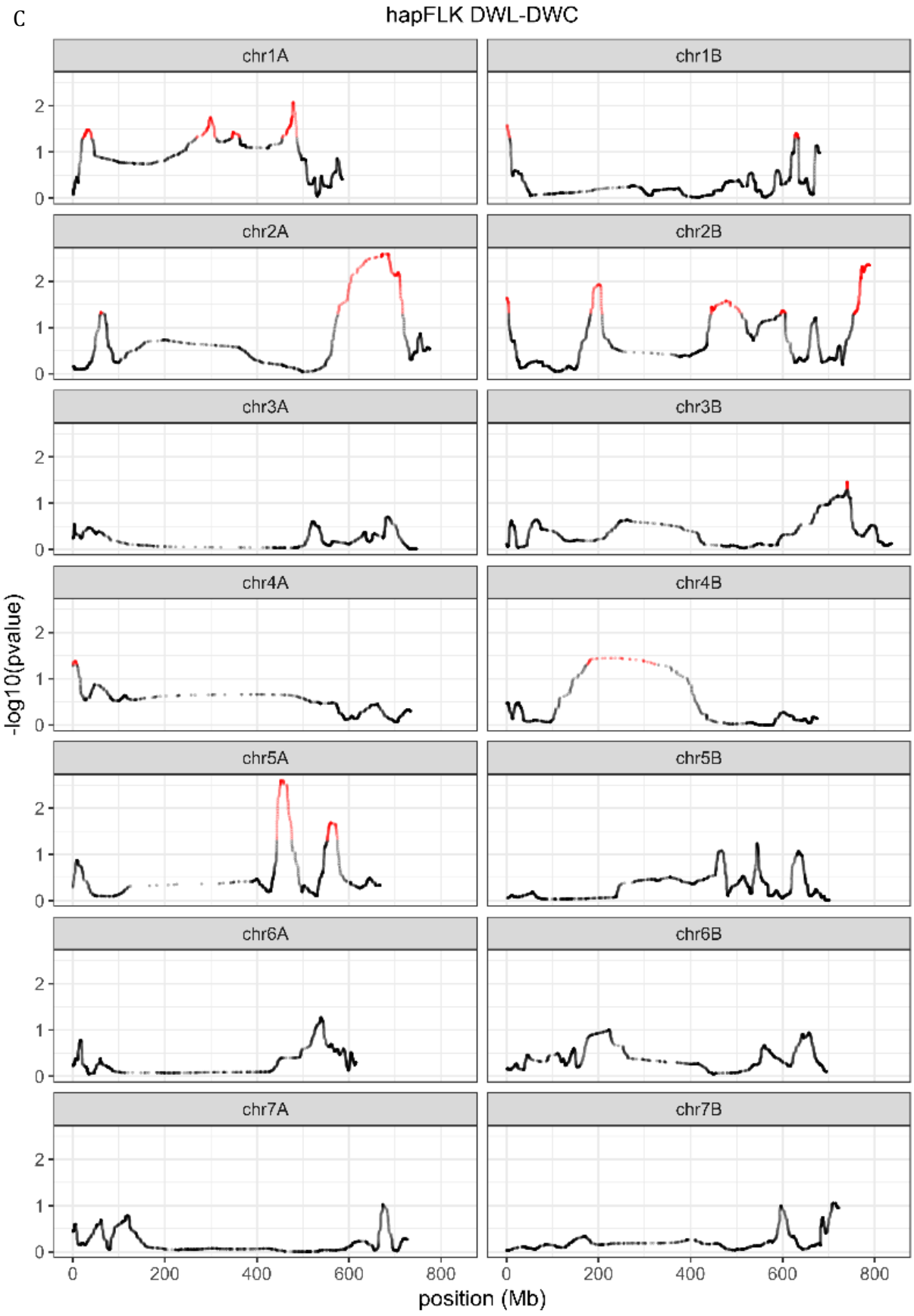



Figure 28. Continued.

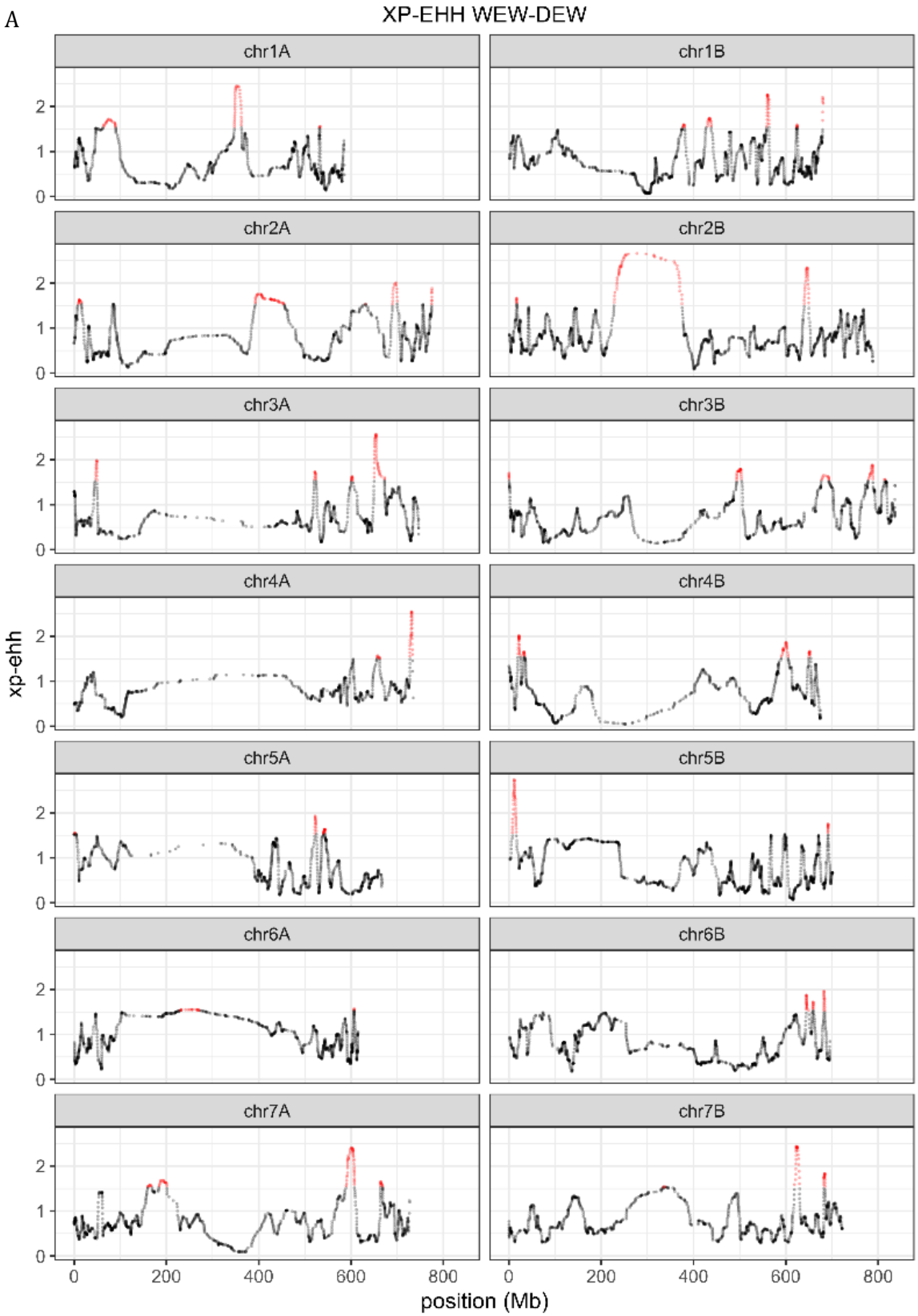



Figure 29. XP-EHH, cross population Extended Haplotype Homozygosity for A. WEW-DEW, B. DEW-DWL, C. DWL-DWC. Top 5% quantile distributions are highlighted as red-filled dots.

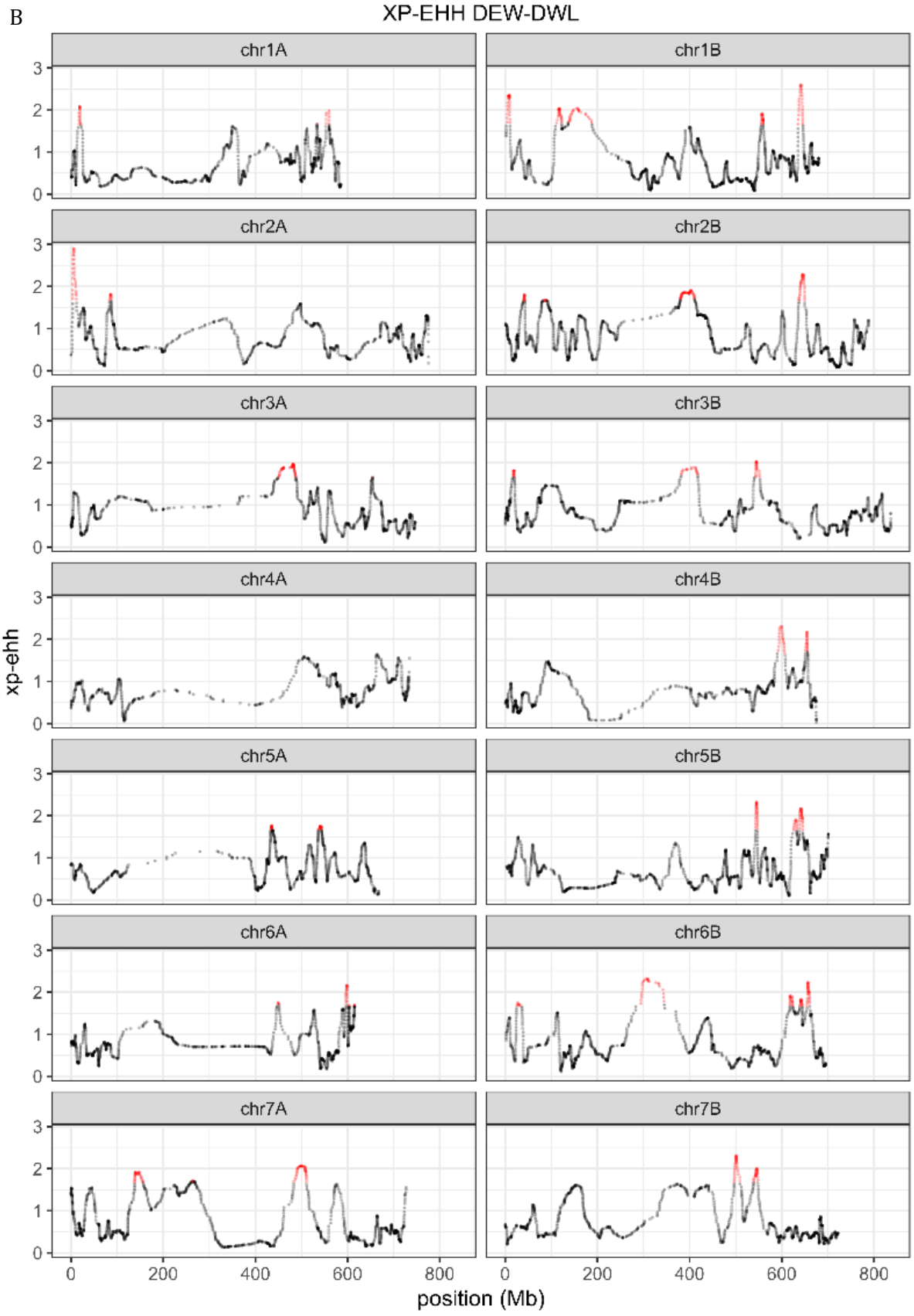



Figure 29. Continued.

C

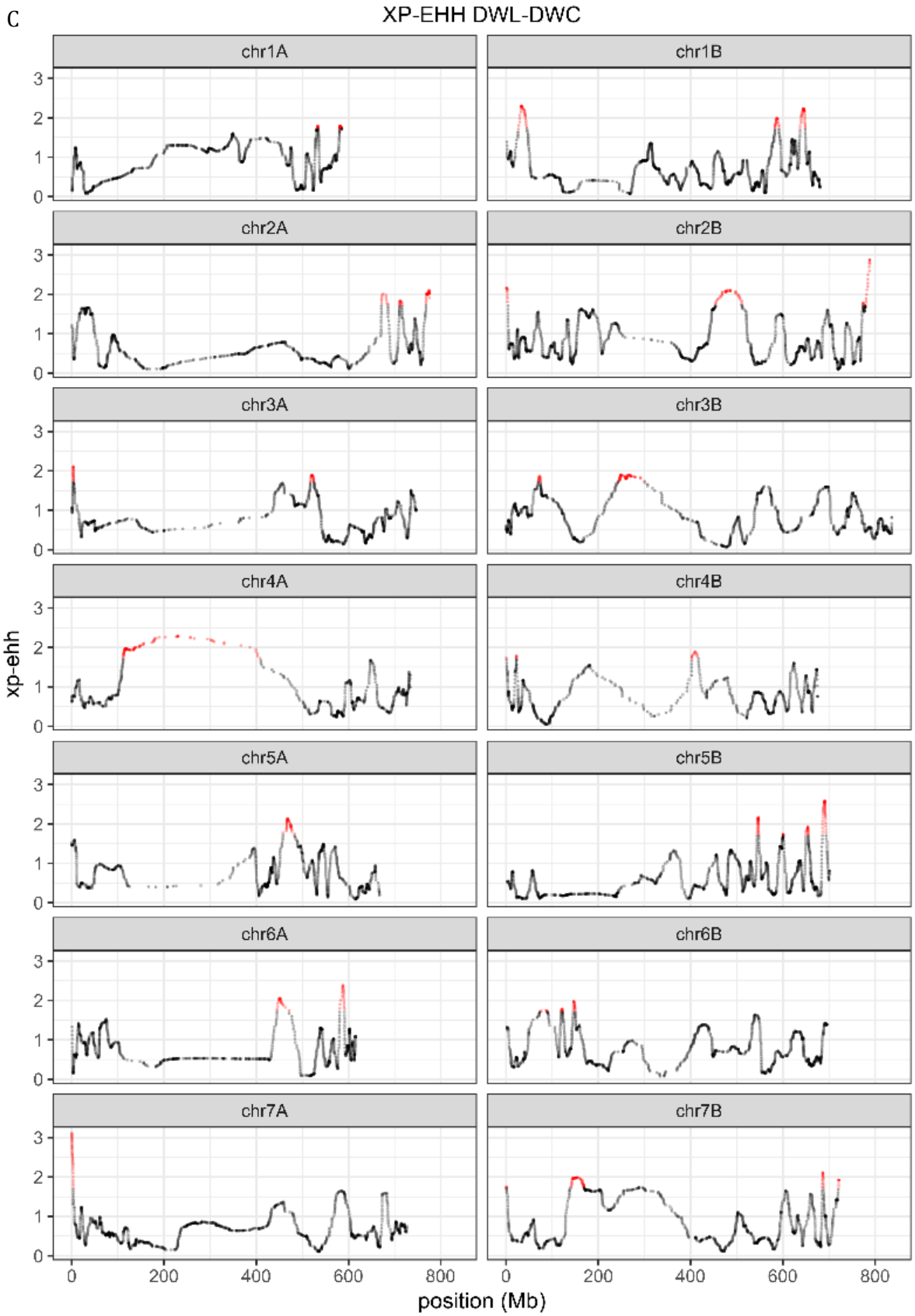



Figure 29. Continued.

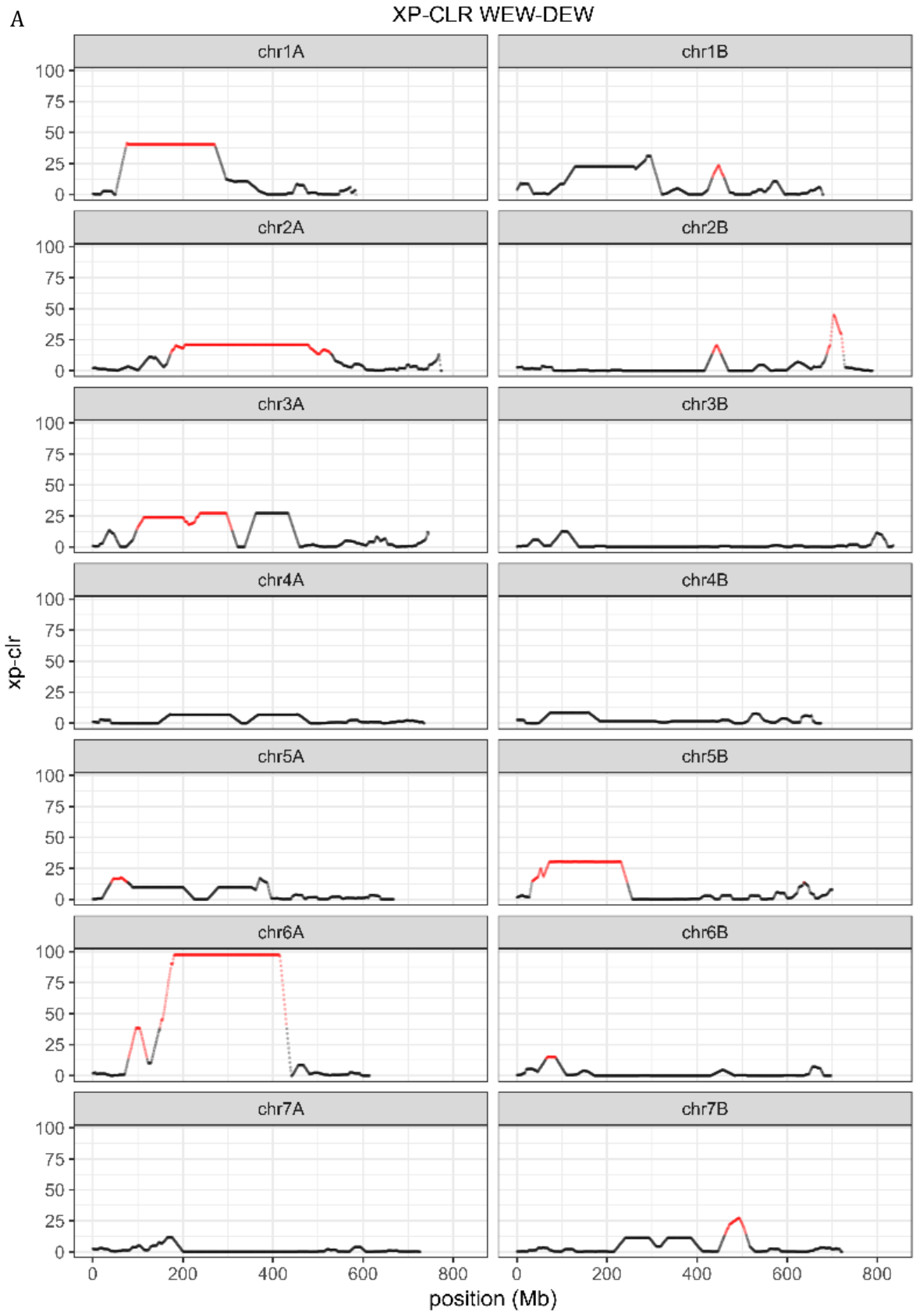



Figure 30. XP-CLR, multilocus test for allele frequency differentiation for A. WEW-DEW, B. DEW-DWL, C. DWL-DWC. Top 5% quantile distributions are highlighted as red-filled dots.

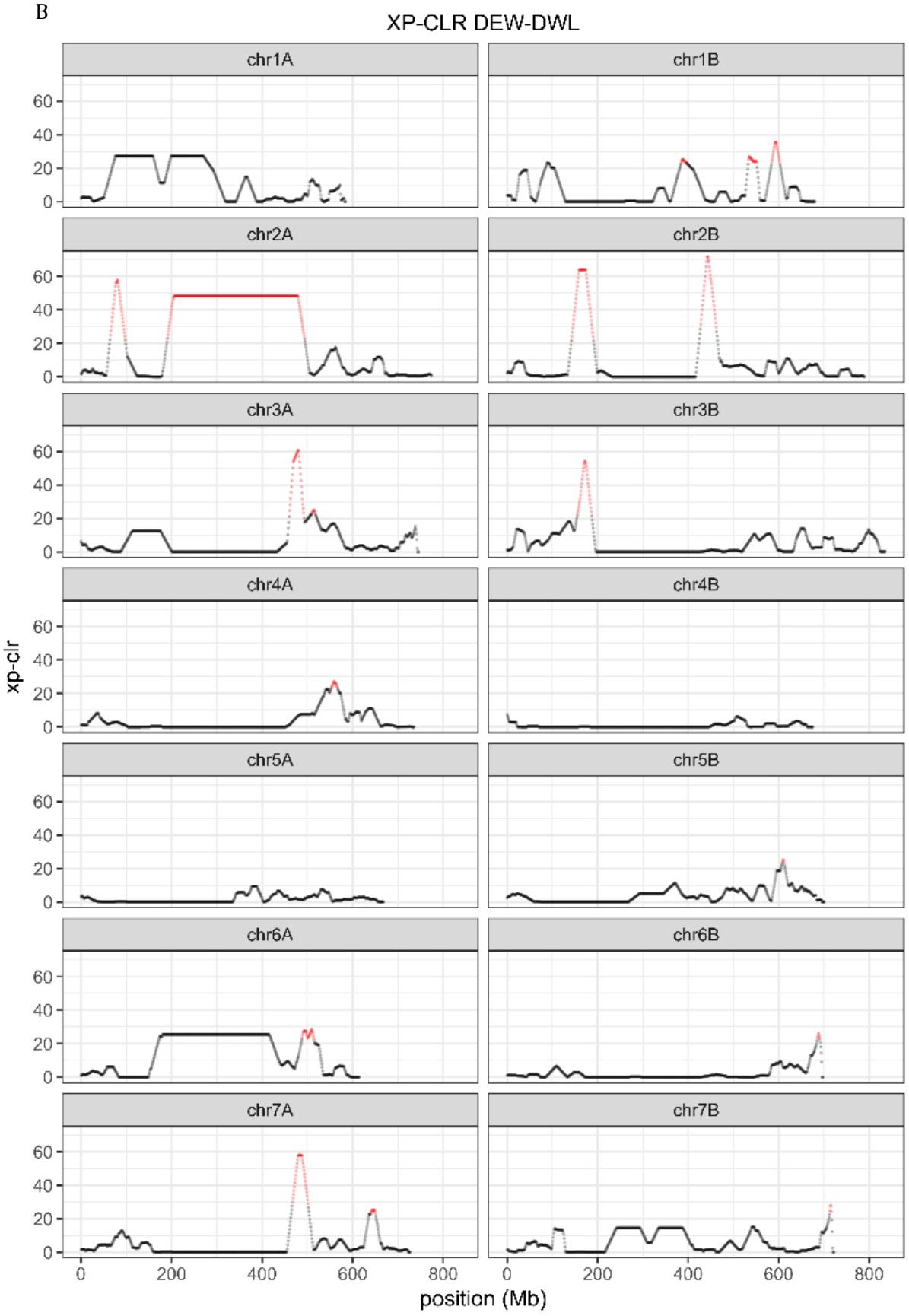



Figure 30. Continued.

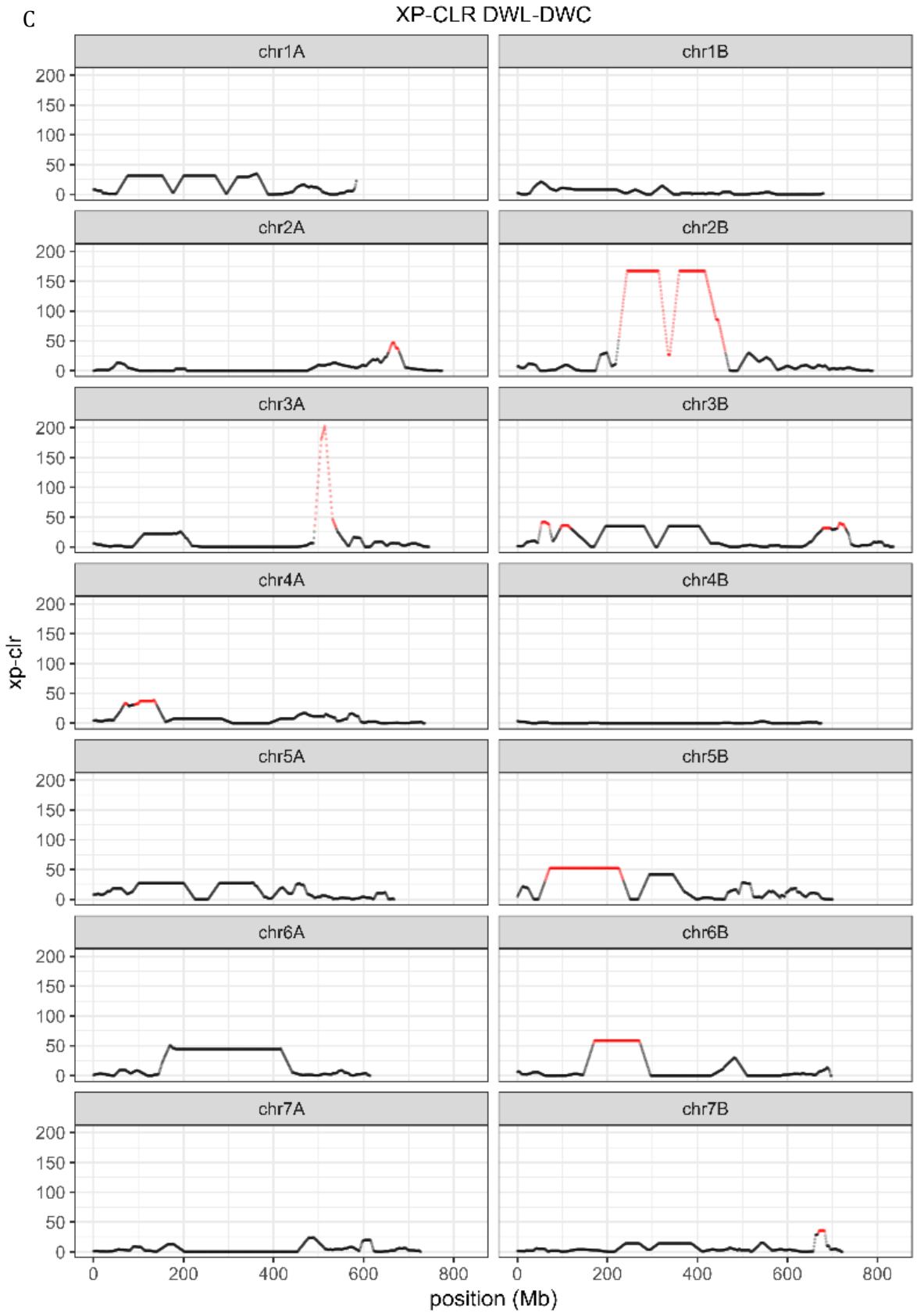



Figure 30. Continued.

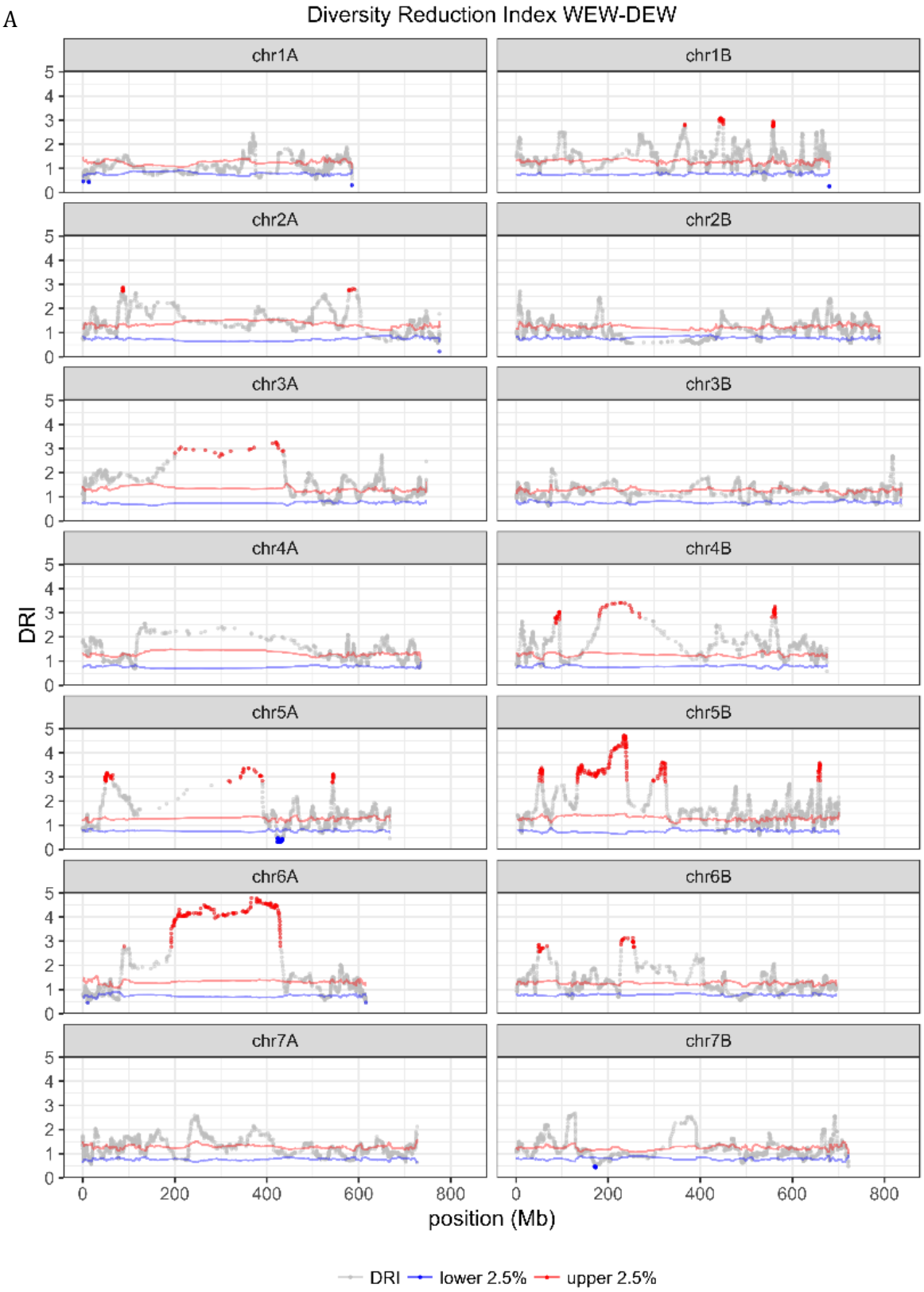



Figure 31. Diversity Reduction Index (DRI) for A. WEW-DEW, B. DEW-DWL, C. DWL-DWC. Top and bottom 2.5% DI quantile distributions are highlighted as red- and blue-filled dots, respectively.

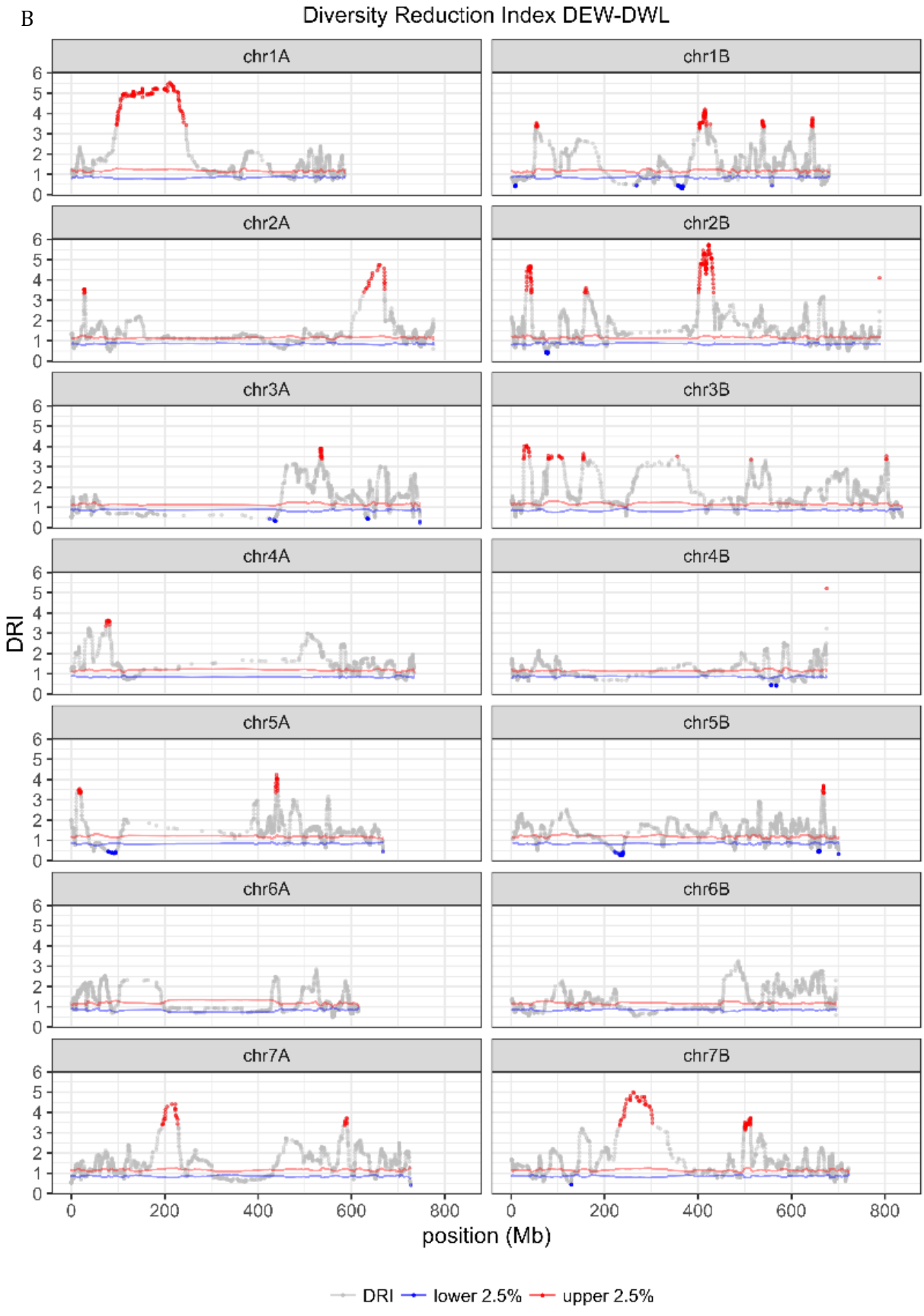



Figure 31. Continued.

C

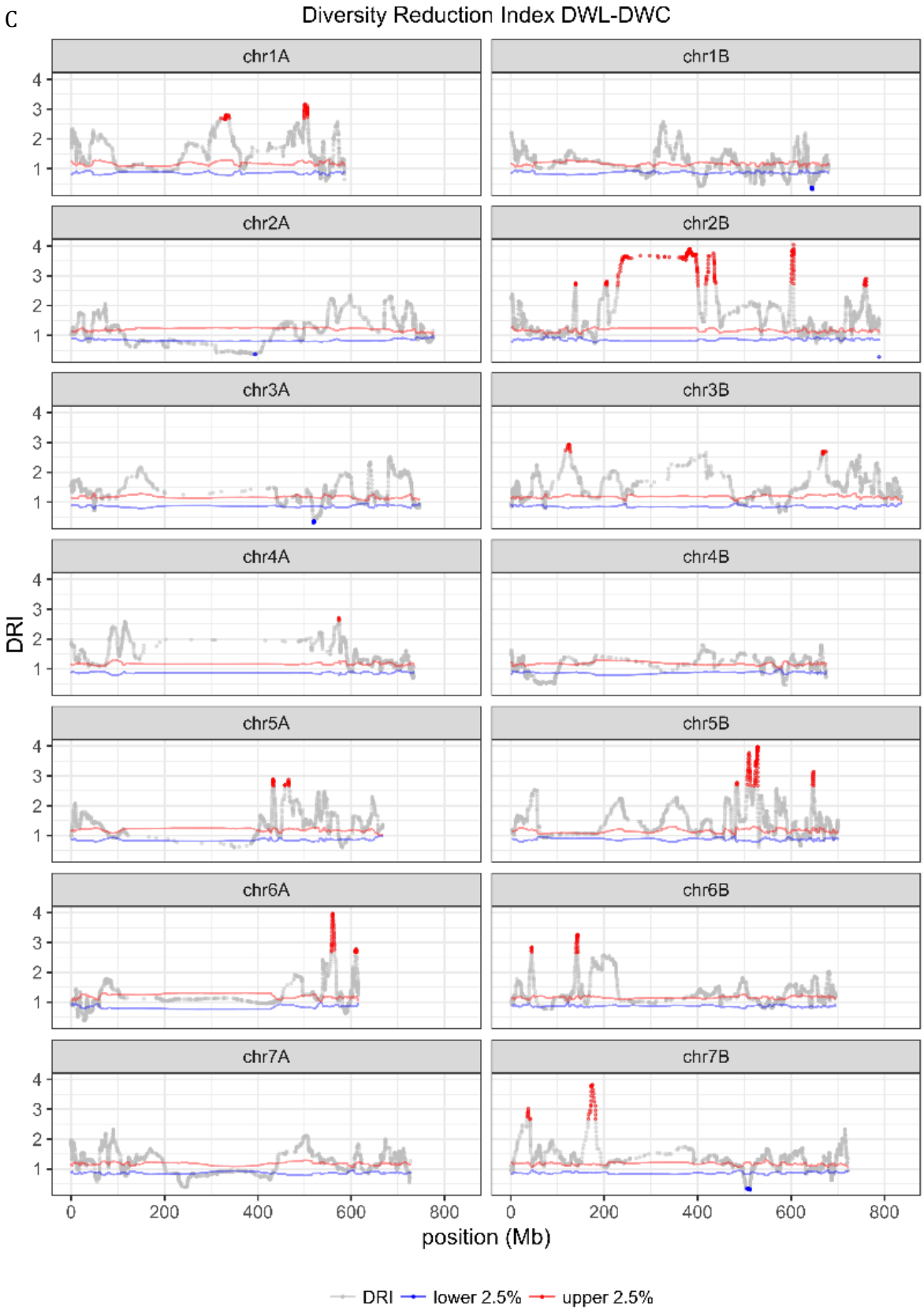



Figure 31. Continued.

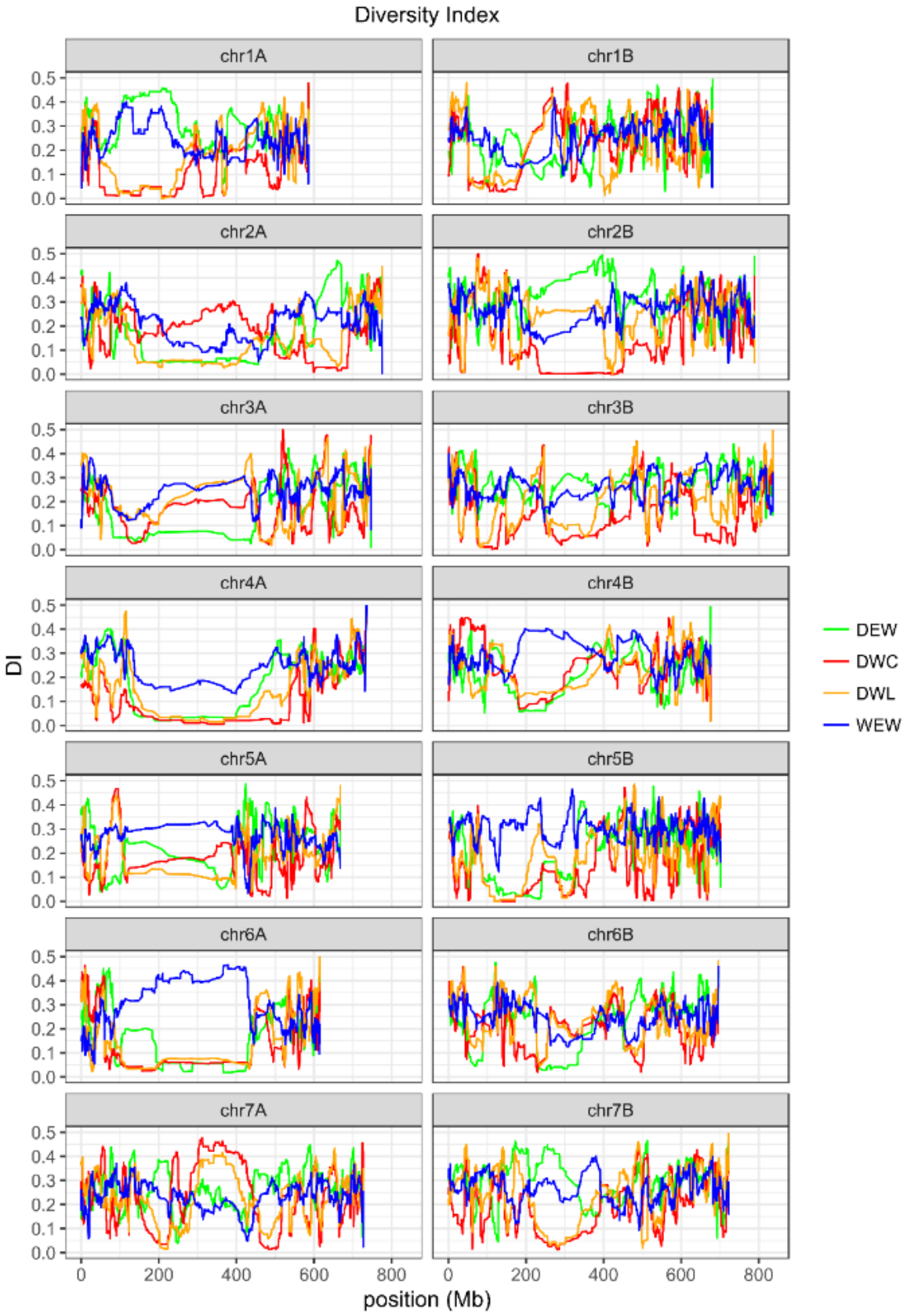



Figure 32. Diversity Index. Green - domesticated emmer wheat, red – durum wheat cultivar, orange - durum wheat landrace, blue - wild emmer wheat.

# 6. Discussion

The International Durum Wheat Genome Sequencing Consortium assembled a high-quality draft genome sequence of the durum wheat cultivar Svevo, which has been a quality and productivity durum variety in Italy for more than a decade. A DenovoMAGIC2 (NRGene, NesZiona, Israel) software using a novel 3D chromosome-conformation capture coupled with high-throughput sequencing (Hi-C) data produced a high quality genome assembly as for the wild emmer wheat accession Zavitan (Avni et al., 2017), bread wheat cultivar Chinese Spring (International Wheat Genome Sequencing Consortium, 2018) and barley cultivar Morex (Mascher et al., 2017). Subsequently, following the gene annotation pipeline established by Plant Genome and Systems biology (PGSB, Helmholtz Zentrum Muenchen) described previously (Avni et al., 2017) and various annotation data sources, such as cDNA and RNA-seq from different tissues and libraries, we investigated the gene content and organization of the durum wheat cv. Svevo genome. As a result, using the gene annotation pipeline we predicted a total number of 369,963 genes: 66,559 high-confidence (HC) and 303,404 low-confidence (LC) genes. Additionally, we predicted the v2 gene models of the wild emmer wheat (WEW) accession Zavitan following the same pipeline and using the same annotation data sources in order to investigate the divergence between the wild and domesticated tetraploid wheat genomes. Thus, wild emmer wheat had 67,182 high-confidence genes and 271,179 low-confidence genes. In addition, the quality of the predicted genes was tested using BUSCO and 216 experimentally determined genes, and for Svevo genome, 98.1 and 97.7% of genes were found, respectively. These high values indicate that the assembly represents an almost complete fraction of the gene space. The number of predicted HC transcripts of Svevo genome was slightly higher on subgenome B compared to subgenome A, except for chromosomes 4 and 7, which have undergone some translocation events. The density of the HC genes was higher in the distal regions of the chromosome arms compared to the pericentromeric regions, confirming previous findings on the gene density arrangement along the chromosome arms in other wheat species (International Wheat Genome Sequencing Consortium, 2014; Avni et al., 2017; International Wheat Genome Sequencing Consortium, 2018).

We used the transcriptomic data to investigate the gene expression and the pattern of gene expression on subgenomes and chromosomes. The analysis revealed that 61,269 (95.8%) genes were expressed at least in one of the 57 samples and 2,724 (4.3%) of genes were not expressed at all and 21,878 genes (34.2%) were expressed in all 57 samples. Moreover, while analyzing the pattern of expressed genes along the chromosomes, we noted that the number of the expressed genes was higher in the distal regions of the chromosome arms, and the number of libraries under which these genes were expressed was lower in the distal regions and higher in the pericentromeric region. This suggests that there are more condition specific genes rather than housekeeping genes. This result is

similar to the gene expression pattern observed for wild emmer wheat genome (Avni et al., 2017). Moreover, in Chinese Spring the genes located in the proximal regions correlated with Gene Ontology (GO) terms such as “cell cycle”, “photosynthesis” and “translation”, while in the highly recombinant distal regions the genes correlated with GO terms such as “response to stress” and “external stimuli” (International Wheat Genome Sequencing Consortium, 2018). The mean expression level per gene, hence the mean expression value of all conditions and tissues, was higher in the distal regions compared to pericentromeric regions. Same result was obtained for the bread wheat cultivar Chinese Spring chromosome 3B (Choulet et al., 2014) and wild emmer wheat accession Zavitan (Avni et al., 2017). Therefore, this observed pattern could be a whole genome phenomenon typical to wheat species (Avni et al., 2017). The assembly allows studying the gene expression analysis in genome-wide and subgenome level and using a wide range of tissues and development stages of wheat transcriptomic data it is possible to produce a co-expression network analysis of the genes.

In addition to these analyses, we analyzed the expression pattern of thirteen varieties of durum wheat that represent the worldwide elite durum germplasm and span a breeding period from 1940 to 2004. These varieties include a wide range of germplasm such as Italian, CIMMYT and North American. Out of 63,993 high-confidence genes 86.6, 75.0, 70.5 and 74.5% of genes showed expression evidence in all tissues and all varieties, grains, leaves and roots, respectively. Interestingly, the PCA analysis showed a stronger variation between tissues than between the varieties and we identified several differentiated up- and down-regulated gene expression clusters based on both tissues and varieties. Moreover, the expression profiles of the cultivars revealed some ancestral related clustering. This kind of expression pattern databases could be useful to identify genes regulated by expression QTL (eQTL) and to elucidate the function of candidate genes. Moreover, characterizing the gene expression presence-absence variation (ePAV) in tetraploid durum wheat enables to investigate the association between genotypic and phenotypic variation.

NLR gene family is one of the gene families that represent a high importance in wheat breeding and improvement. We identified several loci that could be associated with disease resistance genes (NLRs) both in durum wheat cv. Svevo and in wild emmer wheat accession Zavitan. NLR loci clustered principally at the distal regions of the chromosome arms and we observed several regions overlapping with the confidence intervals of disease resistance QTLs known from literature. In general, the number of predicted loci was the same: 1,487 for durum wheat and 1,462 for wild emmer wheat. Compared to the RNA-seq based gene models we were able to predict additional 390 NLR loci for durum wheat and 417 for wild emmer wheat. Moreover, we identified that 172 loci were specific for Svevo and 136 loci were specific for Zavitan genome. NB-LRRs are plant disease resistance genes that form one of the biggest gene families in plants and have an important role in

plant resistance mechanisms and plant innate immune systems (Jupe et al., 2012; Marone et al., 2013; Bouktila et al., 2015; Lee and Yeom, 2015). Therefore, identification of their locations on the genome is of a primary importance for scientific and breeding community and the availability of the wheat assemblies, including durum wheat assembly, greatly accelerates the investigation of gene family analysis. Knowledge of the genome-wide distribution (recombination and allele diversity distribution) of NLR and other gene families highly relevant for wheat breeding programs is essential for wheat improvement.

The availability of wild emmer and durum wheat genomes allow us to understand better the evolution and domestication of tetraploid wheat. Besides, we used a Global Tetraploid wheat germplasm Collection (GTC), composed of 1,854 accessions of up to ten different species and subspecies such as Persian wheat, wild emmer wheat, durum wheat landraces and cultivars, cultivated emmer wheat, Polish wheat, Khorasan wheat, Miracle wheat, Karamyschev's wheat and Ethiopian wheat. These accessions were collected from a wide range of areas including Fertile Crescent, Northern Africa, Europe, India, Ethiopia, Transcaucasia, Central Asia, North and South America. These accessions represent the four principal germplasm groups that are involved in the history of tetraploid domestication and selection processes: Wild Emmer Wheat - WEW, Domesticated Emmer Wheat - DEW, Durum Wheat Landraces - DWL and Durum Wheat Cultivars - DWC. The accessions were genotyped using the wheat iSelect 90K SNP Infinium assay (Wang et al., 2014). As a result, 17,340 SNPs represented the first set of SNPs used to detect the signatures of selection from wild emmer wheat to domesticated wheat. Whereas, 5,784 SNPs represented the second LD-pruned ($r^2 > 0.5$) and minor allele frequency (MAF 0.02) filtered set and was used to study the population genetic structure. The NJ tree analysis clearly differentiated the four main germplasm groups (WEW, DEW, DWL, DWC), which suggest the presence of a strong founder effect and only slight evidence for the phylogenetic origin. The results of population structure representation observed from two independent ADMIXTURE and DAPC non-hierarchical clustering methods were similar. However, ADMIXTURE cluster classification was relatively more informative compared to DAPC. Based on ADMIXTURE analysis at K = 2, WEW, DEW, *T. turgidum* ssp. *carthlicum* and the Ethiopian DWL formed the first group, and *T. turgidum* ssp. *durum* landraces, cultivated accessions and other taxa formed a second group. At K = 3, the Ethiopian DWL separated from the other accessions and at K = 17 subdivided further in two different populations. At K = 5, DEW separated from WEW. At K = 6, DEW Western and Eastern populations separated, and at K = 7 DEW Ethiopian and Indian accessions separated. At K = 4 - 13 DEW germplasm clearly separated into five groups. K = 12 the. *T. turgidum* ssp. *carthlicum* clearly separated as well. This result was comparable to the results obtained from previous diversity studies (Badaeva et al., 2005).

The genetic diversity of DEW was rather expansive compared to a little genetic diversity of DWC. At higher K values, the genetic diversities of DEW as well as WEW were well structured. While DWL was less structured, and at low K values showed high admixture rate between the populations. Ethiopian, *T. turanicum* and *T. carthlicum* subpopulations were genetically distant from other populations and probably have low contribution for the modern tetraploid germplasm. For WEW germplasm, we identified two main populations: North Eastern Fertile Crescent and Southern Levant. DEW germplasm subdivided into six main populations: two Northern (Turkey-to-Transcaucasia/Iran and Turkey-to-Balkans) and four Southern populations (Southern Europe, Southern Levant-to-Europe, Southern Levant-to-Europe2, Indian, Omani and Ethiopian). Further, DWL germplasm composed of six main populations as well: two Northern FC (Turkey-to-Fertile Crescent, Turkey-to-Transcaucasia) and four Southern (Southern Levant FC, Southern Levant-to-North Africa, Greece-to-Balkans, Ethiopian Landraces and *Triticum turanicum*). Finally, DWC composed of five branches from North African Landraces such as ICARDA dryland, Italian germplasm, CIMMYT germplasm released in the '70's and '80's and germplasm adapted to Mediterranean environment.

According to our results, Northern populations of WEW germplasm is the possible ancestor of DEW populations. DWC germplasm showed closer relationship to the DWL North-African and Transcaucasian subpopulations. The Balkan and Syrian subpopulations of DWL were closely related to the DWC subpopulations for the ICARDA dryland and the Italian germplasm adapted to Mediterranean basin.

The tetraploid wheat germplasm is highly structured and it is difficult to capture a well-defined population structure using a single K value. At lower K values WEW and DEW occurred to be highly structured based on main population structures and at higher K values additional well-defined populations emerged. This population differentiation is mostly related to the geographical origin of the accessions. For the DEW and DWL germplasm groups the differentiation is caused mainly by the human-driven dispersal processes. In the DWL germplasm, admixture among the main populations brought by the Mediterranean cross exchange resulted to be an important component of the diversity.

We used the following three main cross-population transitions in order to study the genetic diversity reduction and detect the selection signals associated with wild emmer domestication and durum wheat evolution and breeding: WEW-DEW, DEW-DWL and DWL-DWC. Further, the three main cross-population transitions (WEW-DEW, DEW-DWL, and DWL-DWC), to study the genetic diversity reduction and detect the selection signals associated with wild emmer domestication and durum wheat evolution and breeding, showed some interesting regions. We observed lower diversity

for WEW germplasms only on pericentromeric regions of chromosomes 2A and 4A. This allowed us to use WEW as a reference to study the diversity reduction caused by domestication and breeding in the durum wheat. Subsequently, we observed several diversity reduction regions in the domesticated and improved germplasm groups. Namely, on chromosomes 1A and 7B we observed DEW-DWL specific diversity depletion and on chromosomes 4A, 5B, 6A a WEW-DEW specific diversity reduction emerged.

The selection sweep detection highlighted several genomic regions with putative signatures of selection supported by one or more indexes. At the first transition step of domestication of wild emmer wheat, we already observed diversity reductions on several chromosomes such as chromosome 2A, group 4, group 5 and chromosome 6B. Moreover, the region of brittle rachis loci location on chromosome group 3 showed a strong reduction in diversity supported by several indexes. Subsequently, in cross-population transition from DEW to DWL the region was subjected to a further reduction in diversity. In addition, several diversity reduction and selection signature peaks overlapped with loci associated with domestication, improvement or breeding that lead to the formation of the present-day durum wheat cultivars. These loci are associated with disease resistance, domestication, yellow pigment content, seed dormancy.

The analysis of selective signature detection provides an insight into the dynamics of the changes occurred during domestication and selection processes. Therefore, the detailed knowledge of the genetic information of durum wheat provides a valuable source directed to increase the accuracy of breeding process, which, in turn, will bring the advancement of pasta wheat.

Durum wheat cv. Svevo assembly is a valuable resource for wheat improvement. For breeders, the knowledge of the whole-genome of the durum wheat is advantageous since new crosses implicate genome-wide changes in the gene networks that control the expression of complex traits. Availability of the assembly and annotation of the durum wheat genome allows breeders and scientists to have access to the changes occurring at genome level. Plenty of QTLs have been identified for wheat; however, only in few cases the causal gene has been cloned. Durum wheat annotation gives access to the regulatory regions and can be used to conduct a meta-QTL analysis by anchoring all published QTLs against the reference genome. This knowledge aids researchers and breeders to deal with various selection challenges.

The selection and improvement of important agronomic traits is complicated since such traits expression greatly depends on the environmental conditions and management practices. This problem can be overcome by using a forward genetic approach (by identifying the DNA markers, which are in strong linkage disequilibrium with the phenotype) or by targeting the genes through

the genome editing.

Avni et al. (Avni et al., 2017) in order to demonstrate the potential of the polyploid wheat assembly targeted the domestication trait of non-shattering spike, Brittle Rachis 1, using a population derived from a cross between accession Zavitan and domesticated durum wheat cultivar Svevo. The authors identified and demonstrated that the alleles of domesticated wheat carried mutations, which destroyed the structures of the encoded proteins, and hence are loss-of function alleles. Therefore, the reference assembly of durum wheat serve as a valuable resource directed to accelerate the progress and accuracy of genome-assisted advancement of modern wheat.

The availability of the durum wheat assembly and annotation accelerates the identification of candidate genes in QTLs. The assemblies and annotations of wild emmer wheat (Avni et al., 2017), bread wheat (International Wheat Genome Sequencing Consortium, 2018) and durum wheat could be implemented in order to identify the QTL confidence intervals, find the genes in the confidence interval and prioritize the candidate genes based on the annotation. Moreover, the three assemblies could be used in order to compare the gene models in the confidence intervals of the three genomes, the presence of the scaffold ordering errors and the annotation of the assemblies.

In general, knowledge of the durum wheat genome and its organization and function of the genes is fundamental. It provides a basis for the evolutionary reveal from wild to domesticated tetraploid wheat as well as for the investigation of the major important traits directed for durum wheat improvement. Therefore, the availability of genome assembly of durum wheat, their predicted gene models, annotations and gene expressions represent a valuable resource for scientists as well as breeders to speed up and advance the process of genetic improvement of durum wheat.

**References**


Akhunov, E. D., Akhunova, A. R., Anderson, O. D., Anderson, J. A., Blake, N., Clegg, M. T., et al. (2010). Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. *BMC Genomics* 11, 702. doi:10.1186/1471-2164-11-702.

Akhunova, A. R., Matniyazov, R. T., Liang, H., and Akhunov, E. D. (2010). Homoeolog-specific transcriptional bias in allopolyploid wheat. *BMC Genomics* 11, 505. doi:10.1186/1471-2164-11-505.

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–64. doi:10.1101/gr.094052.109.

Alvarez, M. A., Tranquilli, G., Lewis, S., Kippes, N., and Dubcovsky, J. (2016). Genetic and physical mapping of the earliness per se locus Eps-A m 1 in Triticum monococcum identifies EARLY FLOWERING 3 (ELF3) as a candidate gene. *Funct. Integr. Genomics* 16, 365–382. doi:10.1007/s10142-016-0490-3.

Aoki, N., Whitfeld, P., Hoeren, F., Scofield, G., Newell, K., Patrick, J., et al. (2002). Three sucrose transporter genes are expressed in the developing grain of hexaploid wheat. *Plant Mol. Biol.* 50, 453–462. doi:10.1023/A:1019846832163.

Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, Sven, O., Gundlach, H., et al. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* 357, 93–97. doi:10.1126/science.aan0032.

Avni, R., Nave, M., Eilam, T., Sela, H., Alekperov, C., Peleg, Z., et al. (2014). Ultra-dense genetic map of durum wheat × wild emmer wheat developed using the 90K iSelect SNP genotyping assay. *Mol. Breed.* 34, 1549–1562. doi:10.1007/s11032-014-0176-2.

Badaeva, E. D., Keilwagen, J., Knüpffer, H., Waßermann, L., Dedkova, O. S., Mitrofanova, O. P., et al. (2015). Chromosomal passports provide new insights into diffusion of emmer wheat. *PLoS One* 10, 1–25. doi:10.1371/journal.pone.0128556.

Badaeva, E. D., Loskutov, I. G., Shelukhina, O. Y., and Pukhalsky, V. A. (2005). Cytogenetic Analysis of Diploid Avena L. Species Containing the As Genome. *Russ. J. Genet.* 41, 1428–1433. doi:10.1007/s11177-006-0018-3.

Bálint, A. F., Kovács, G., and Sutka, J. (2000). ORIGIN AND TAXONOMY OF WHEAT IN THE LIGHT OF RECENT RESEARCH. *Acta Agron. Hungarica* 48, 301–313. doi:10.1556/AAgr.48.2000.3.11.

Beier, S., Himmelbach, A., Colmsee, C., Zhang, X.-Q., Barrero, R. A., Zhang, Q., et al. (2017). Construction of a map-based reference genome sequence for barley, Hordeum vulgare L. *Sci. Data* 4, 170044. doi:10.1038/sdata.2017.44.

Belea, A. (1971). Are there other possibilities for the origin of T. spelta or T. aestivum? *Belea, A.*, 153–206.

Belea, A., and Fejér, O. (1980). Evolution of the wheat (Triticum L.) in respect to recent research. *Acta Agron. Hung.*, 305–315.

Bennetzen, J. L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A. C., et al. (2012). Reference genome sequence of the model plant Setaria. *Nat. Biotechnol.* 30, 555–561. doi:10.1038/nbt.2196.

Bennici, A. (1986). "Durum Wheat (Triticum durum Desf.)," in (Springer, Berlin, Heidelberg), 89–104. doi:10.1007/978-3-642-61625-9_5.

Blake, N. K., Lehfeldt, B. R., Lavin, M., and Talbert, L. E. (1999). Phylogenetic reconstruction based on low copy DNA sequence data in an allopolyploid: The B genome of wheat. *Genome* 42, 351–

360. doi:10.1139/g98-136.

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., et al. (2010). Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics* 186, 241–62. doi:10.1534/genetics.104.117275.

Bouktila, D., Khalfallah, Y., Habachi-Houimli, Y., Mezghani-Khemakhem, M., Makni, M., and Makni, H. (2015). Full-genome identification and characterization of NBS-encoding disease resistance genes in wheat. *Mol. Genet. Genomics* 290, 257–271. doi:10.1007/s00438-014-0909-2.

Bozzini, A. (1988). Origin, distribution, and production of durum wheat in the world. *Fabriani G Lintas C (ed). Durum Chem. Technol. AACC, Minnesota, USA*, 1–16.

Braidwood, R. J., Cambel, H., and Watson, P. J. (1969). Prehistoric investigations in southeastern Turkey. *Science* 164, 1275–6. doi:10.1126/science.164.3885.1275.

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G. L. A., D'Amore, R., Allen, A. M., et al. (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* 491, 705–710. doi:10.1038/nature11650.

Cantu, D., Vanzetti, L. S., Sumner, A., Dubcovsky, M., Matvienko, M., Distelfeld, A., et al. (2010). Small RNAs, DNA methylation and transposable elements in wheat. *BMC Genomics* 11, 408. doi:10.1186/1471-2164-11-408.

Chapman, J. A., Mascher, M., Buluç, A., Barry, K., Georganas, E., Session, A., et al. (2015). A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biol.* 16, 26. doi:10.1186/s13059-015-0582-8.

Chapman, V., Miller, T. E., and Riley, R. (1976). Equivalence of the A genome of bread wheat with that of Triticum urartu. *Genet. Res.*, 69–76.

Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps Population differentiation as a test for selective sweeps. 393–402. doi:10.1101/gr.100545.109.

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 13, 1050–1054. doi:10.1038/nmeth.4035.

Choulet, F., Alberti, A., Theil, S., Glover, N., Barbe, V., Daron, J., et al. (2014). Structural and functional partitioning of bread wheat chromosome 3B. *Science* 345, 1249721. doi:10.1126/science.1249721.

Choulet, F., Wicker, T., Rustenholz, C., Paux, E., Salse, J., Leroy, P., et al. (2010). Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell* 22, 1686–701. doi:10.1105/tpc.110.074187.

Chu, C.-G., Tan, C. T., Yu, G.-T., Zhong, S., Xu, S. S., and Yan, L. (2011). A Novel Retrotransposon Inserted in the Dominant Vrn-B1 Allele Confers Spring Growth Habit in Tetraploid Wheat (Triticum turgidum L.). *G3 (Bethesda).* 1, 637–45. doi:10.1534/g3.111.001131.

Clarke, J. ., McCaig, T. ., DePauw, R. M., Knox, R. E., Ames, N. P., Clarke, F. R., et al. (2005). Commander Durum Wheat. *Can. J. Plant Sci*, 901–904.

Clavijo, B. J., Venturini, L., Schudoma, C., Accinelli, G. G., Kaithakottil, G., Wright, J., et al. (2017). An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Res.* 27, 885–896. doi:10.1101/gr.217117.116.

De Vita, P., Li, O., Nicosia, D., Nigro, F., Platani, C., Riefolo, C., et al. (2007). Breeding progress in morpho-physiological, agronomical and qualitative traits of durum wheat cultivars released in

Italy during the 20th century. *Eur. J. Agron.* 26, 39–53. doi:10.1016/j.eja.2006.08.009.

Donmez, E., Sears, R., Shroyer, J., and Paulsen, J. (2000). Evaluation of Winter Durum Wheat for Kansas. *Kansas State Univ. Agric. Exp. Stn. Coop. Ext. Serv.*

Dorofeev, V., Filatenko, A., Mighushova, E., Udaczin, R., and Jakubziner, M. (1979). *Wheat*. In: Dorofe. Leningrad, Russia.

Dubcovsky, J., and Dvorak, J. (2007). Genome Plasticity a Key Factor in the Success of Polyploid Wheat Under Domestication. Available at: http://science.sciencemag.org/ [Accessed August 28, 2018].

Durbin, R. (1998). *Biological sequence analysis : probabilistic models of proteins and nucleic acids*.

Dvořák, J. (2001). Triticum Species (Wheat). *Encycl. Genet.*, 2060–2068. doi:10.1006/RWGN.2001.1672.

Dvorak, J., and Zhang, H. B. (1990). Variation in repeated nucleotide sequences sheds light on phylogeny of the wheat B and G genomes. *Proc. Natl. Acad. Sci. USA*, 9640–9644.

Edwards, D., Batley, J., and Snowdon, R. J. (2013). Accessing complex crop genomes with next-generation sequencing. *Theor. Appl. Genet.*, 1–11. doi:10.1007/s00122-012-1964-x.

Eversole, K., Feuillet, C., Mayer, K. F. X., and Rogers, J. (2014). Slicing the wheat genome. Introduction. *Science* 345, 285–7. doi:10.1126/science.1257983.

Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M., and Servin, B. (2013). Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations. *Genetics* 193, 929–941. doi:10.1534/genetics.112.147231.

Faris, J. D., Zhang, Q., Chao, S., Zhang, Z., and Xu, S. S. (2014a). Analysis of agronomic and domestication traits in a durum × cultivated emmer wheat population using a high-density single nucleotide polymorphism-based linkage map. *Theor. Appl. Genet.* 127, 2333–2348. doi:10.1007/s00122-014-2380-1.

Faris, J. D., Zhang, Z., and Chao, S. (2014b). Map-based analysis of the tenacious glume gene Tg-B1 of wild emmer and its role in wheat domestication ☆. *Gene* 542, 198–208. doi:10.1016/j.gene.2014.03.034.

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. doi:10.1016/J.TPLANTS.2010.10.005.

Gill, B. S., Appels, R., Botha-Oberholster, A.-M., Buell, C. R., Bennetzen, J. L., Chalhoub, B., et al. (2004). A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium. *Genetics* 168, 1087–96. doi:10.1534/genetics.104.034769.

Goncharov, N. P. (2011). Genus Triticum L. taxonomy: the present and the future. *Plant Syst. Evol.* 295, 1–11. doi:10.1007/s00606-011-0480-9.

Goudet, J., and Jombart, T. (2015). Estimation and Tests of Hierarchical F-Statistics. *R Core Team*, 58. doi:10.2307/2411227>.

Gremme, G., Brendel, V., Sparks, M. E., and Kurtz, S. (2005). Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.* 47, 965–978. doi:10.1016/j.infsof.2005.09.005.

Gu, Y. Q., Crossman, C., Kong, X., Luo, M., You, F. M., Coleman-Derr, D., et al. (2004). Genomic organization of the complex alpha-gliadin gene loci in wheat. *Theor. Appl. Genet.* 109, 648–57. doi:10.1007/s00122-004-1672-2.

Gupta, P. K. (1972). Cytogenetic evolution in the Triticinae: Homoeologous relationships. *Genetica* 43, 504–530. doi:10.1007/BF00115595.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nat Protoc* 8. doi:10.1038/nprot.

Hammer, K. (1984). Das Domestikationssyndrom. *Die Kult.* 32, 11–34. doi:10.1007/BF02098682.

Haudry, A., Cenci, A., Ravel, C., Bataillon, T., Brunel, D., Poncet, C., et al. (2007). Grinding up Wheat: A Massive Loss of Nucleotide Diversity Since Domestication. *Mol. Biol. Evol.* 24, 1506–1517. doi:10.1093/molbev/msm077.

Heun, M., Schäfer-Pregl, R., Klawan, D., Castagna, R., Accerbi, M., Borghi, B., et al. (1997). Site of Einkorn Wheat Domestication Identified by DNA Fingerprinting. *Science (80-. ).* 278, 1312–1314. doi:10.1126/science.278.5341.1312.

Hou, J., Jiang, Q., Hao, C., Wang, Y., Zhang, H., and Zhang, X. (2014). Global Selection on Sucrose Synthase Haplotypes during a Century of Wheat Breeding. *Plant Physiol.* 164, 1918–1929. doi:10.1104/pp.113.232454.

Hu, M.-J., Zhang, H.-P., Liu, K., Cao, J.-J., Wang, S.-X., Jiang, H., et al. (2016). Cloning and Characterization of TaTGW-7A Gene Associated with Grain Weight in Wheat via SLAF-seq-BSA. *Front. Plant Sci.* 7, 1902. doi:10.3389/fpls.2016.01902.

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R., et al. (2002). Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. *Proc. Natl. Acad. Sci.* 99, 8133–8138. doi:10.1073/pnas.072223799.

Hufford, M. B., Xu, X., van Heerwaarden, J., Pyhäjärvi, T., Chia, J.-M., Cartwright, R. A., et al. (2012). Comparative population genomics of maize domestication and improvement. *Nat. Genet.* 44, 808–811. doi:10.1038/ng.2309.

International Grain Council (2017). Available at: https://www.igc.int/en/default.aspx.

International Wheat Genome Sequencing Consortium (IWGSC), T. I. W. G. S. C. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science* 345, 1251788. doi:10.1126/science.1251788.

International Wheat Genome Sequencing Consortium (IWGSC), T. I. W. G. S. C., IWGSC RefSeq principal investigators:, I. R. principal, Appels, R., Eversole, K., Feuillet, C., Keller, B., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361, eaar7191. doi:10.1126/science.aar7191.

Jiang, Q., Hou, J., Hao, C., Wang, L., Ge, H., Dong, Y., et al. (2011). The wheat (T. aestivum) sucrose synthase 2 gene (TaSus2) active in endosperm development is associated with yield traits. *Funct. Integr. Genomics* 11, 49–61. doi:10.1007/s10142-010-0188-x.

Jiang, Y., Jiang, Q., Hao, C., Hou, J., Wang, L., Zhang, H., et al. (2015). A yield-associated gene TaCWI, in wheat: its function, selection and evolution in global breeding revealed by haplotype analysis. *Theor. Appl. Genet.* 128, 131–143. doi:10.1007/s00122-014-2417-5.

Jombart, T., and Collins, C. (2015). A tutorial for Discriminant Analysis of Principal Components (DAPC) using adegenet 2.0.0. Available at: http://adegenet.r-forge.r-project.org/files/tutorial-dapc.pdf [Accessed September 6, 2018].

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* 11, 94. doi:10.1186/1471-2156-11-94.

Jupe, F., Pritchard, L., Etherington, G. J., MacKenzie, K., Cock, P. J., Wright, F., et al. (2012). Identification and localisation of the NB-LRR gene family within the potato genome. doi:10.1186/1471-2164-13-75.

Kabbaj, H., Sall, A. T., Al-Abdallat, A., Geleta, M., Amri, A., Filali-Maltouf, A., et al. (2017). Genetic Diversity within a Global Panel of Durum Wheat (Triticum durum) Landraces and Modern Germplasm Reveals the History of Alleles Exchange. *Front. Plant Sci.* 8, 1277. doi:10.3389/fpls.2017.01277.

Kalaipandian, S., Xue, G., Rae, A. L., Glassop, D., Bonnett, G. D., and McIntyre, L. C. (2018). Overexpression of *TaCML20* , a calmodulin-like gene, enhances water soluble carbohydrate accumulation and yield in wheat. *Physiol. Plant.* doi:10.1111/ppl.12786.

Kerby, K. Kuspira, J. (1987). The phylogeny of the polyploid wheats Triticum aestivum (bread wheat) and Triticum turgidum (macaroni wheat). *Genome*, 722–737.

Khvorykh, G. (2018). inzilico/imputeqc v1.0.0. GitHub repository, https://github.com/inzilico/imputeqc. GitHub repository,https://github.com/inzilico/impu.

Kihara, H. (1924). Cytologische und genetische Studien bei wichtigen Getreidearten mit besonderer Rucksicht auf das Verhalten der Chromosomen und die Sterilitat in den Bastarden. *Mem. Coll. Sci., Kyoto Imp. Univ.* 1, 1–200. Available at: https://ci.nii.ac.jp/naid/10006568257/ [Accessed August 11, 2018].

Kihara, H. (1937). Genomanalyse bei Triticum und Aegilops VII. Kurze Übersicht über die Ergebnisse der Jahre 1934–36. *Mem. Coll. Sci. Kyoto Imp. Univ*, 1–61.

Kilian, B. (2007). Genetic diversity, evolution and domestication of Triticeae in the Fertile Crescent. Available at: https://docserv.uni-duesseldorf.de/servlets/DerivateServlet/Derivate-6913/doktorarbeit4_2008.pdf [Accessed August 12, 2018].

Kilian, B., Martin, W., and Salamini, F. (2010). "Genetic diversity, evolution and domestication of wheat and barley in the Fertile Crescent," in *Evolution in Action: Case studies in Adaptive Radiation, Speciation and the Origin of Biodiversity* (Berlin, Heidelberg: Springer Berlin Heidelberg), 137–166. doi:10.1007/978-3-642-12425-9_8.

Kolde, R. (2018). Package "pheatmap." Available at: https://cran.r-project.org/web/packages/pheatmap/pheatmap.pdf [Accessed August 31, 2018].

Krasileva, K. V, Buffalo, V., Bailey, P., Pearce, S., Ayling, S., Tabbita, F., et al. (2013). Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biol.* 14, R66. doi:10.1186/gb-2013-14-6-r66.

Krishnakumar, V., Hanlon, M. R., Contrino, S., Ferlanti, E. S., Karamycheva, S., Kim, M., et al. (2015). Araport: The Arabidopsis Information Portal. *Nucleic Acids Res.* 43, D1003–D1009. doi:10.1093/nar/gku1200.

Kuckuck, H., and Schiemann, E. (1957). Über das Vorkommen von Speltz und Emmer (Triticum spelta L. und Tr. dicoccum Schübl.) im Iran. Z. Pflazenzücht. 383–396.

Lee, H.-A., and Yeom, S.-I. (2015). Plant NB-LRR proteins: tightly regulated sensors in a complex manner. *Brief. Funct. Genomics* 14, 233–242. doi:10.1093/bfgp/elv012.

Lev-Yadun, S., Gopher, A., and Abbo, S. (2000). The Cradle of Agriculture. *Science (80-. ).* 288.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi:10.1093/bioinformatics/btr509.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352.

Ling, H.-Q., Ma, B., Shi, X., Liu, H., Dong, L., Sun, H., et al. (2018). Genome sequence of the progenitor of wheat A subgenome Triticum urartu. *Nature* 557, 424–428. doi:10.1038/s41586-018-0108-

0.

Linné, C. von, Linné, C. von, and Salvius, L. (1753). *Caroli Linnaei ... Species plantarum :exhibentes plantas rite cognitas, ad genera relatas, cum differentiis specificis, nominibus trivialibus, synonymis selectis, locis natalibus, secundum systema sexuale digestas...* Holmiae : Impensis Laurentii Salvii, doi:10.5962/bhl.title.669.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8.

Luo, M.-C., Gu, Y. Q., Puiu, D., Wang, H., Twardziok, S. O., Deal, K. R., et al. (2017). Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. *Nature* 551, 498. doi:10.1038/nature24486.

Luo, M.-C., Yang, Z.-L., You, F. M., Kawahara, T., Waines, J. G., and Dvorak, J. (2007). The structure of wild and domesticated emmer wheat populations, gene flow between them, and the site of emmer domestication. *Theor. Appl. Genet.* 114, 947–959. doi:10.1007/s00122-006-0474-0.

Ma, D., Yan, J., He, Z., Wu, L., and Xia, X. (2012). Characterization of a cell wall invertase gene TaCwi-A1 on common wheat chromosome 2A and development of functional markers. *Mol. Breed.* 29, 43–52. doi:10.1007/s11032-010-9524-z.

Maccaferri, M., Cane', M., Sanguineti, M. C., Salvi, S., Colalongo, M. C., Massi, A., et al. (2014). A consensus framework map of durum wheat (Triticum durum Desf.) suitable for linkage disequilibrium analysis and genome-wide association mapping. *BMC Genomics* 15, 873. doi:10.1186/1471-2164-15-873.

Maccaferri, M., Ricci, A., Salvi, S., Milner, S. G., Noli, E., Martelli, P. L., et al. (2015). A high-density, SNP-based consensus map of tetraploid wheat as a bridge to integrate durum and bread wheat genomics and breeding. *Plant Biotechnol. J.* 13. doi:10.1111/pbi.12288.

Maccaferri, M., Sanguineti, M. C., Donini, P., and Tuberosa, R. (2003). Microsatellite analysis reveals a progressive widening of the genetic basis in the elite durum wheat germplasm. *TAG Theor. Appl. Genet.* 107, 783–797. doi:10.1007/s00122-003-1319-8.

Maccaferri, M., Sanguineti, M. C., Noli, E., and Tuberosa, R. (2005). Population structure and long-range linkage disequilibrium in a durum wheat elite collection. *Mol. Breed.* 15, 271–290. doi:10.1007/s11032-004-7012-z.

MacKey, J. (1966). Species relationships in Triticum. *Proc. 2nd Int. Wheat Genet. Symp., Hered.* 2, 237–276.

MacKey, J. (1988). A plant breeder's perspective on taxonomy of cultivated plants. *Biol. Zent. Bl.* 107, 369–379.

Malomane, D. K., Reimer, C., Weigend, S., Weigend, A., Sharifi, A. R., and Simianer, H. (2018). Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. *BMC Genomics* 19, 22. doi:10.1186/s12864-017-4416-9.

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., et al. (2017). CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* 45, D200–D203. doi:10.1093/nar/gkw1129.

Marone, D., Laidò, G., Gadaleta, A., Colasuonno, P., Ficco, D. B. M., Giancaspro, A., et al. (2012). A high-density consensus map of A and B wheat genomes. *Theor. Appl. Genet.* 125, 1619–38. doi:10.1007/s00122-012-1939-y.

Marone, D., Russo, M., Laidò, G., De Leonardis, A., Mastrangelo, A., Marone, D., et al. (2013). Plant Nucleotide Binding Site–Leucine-Rich Repeat (NBS-LRR) Genes: Active Guardians in Host Defense Responses. *Int. J. Mol. Sci.* 14, 7302–7326. doi:10.3390/ijms14047302.

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S. O., Wicker, T., et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. *Nature* 544, 427–433. doi:10.1038/nature22043.

McFadden, E. S., and Sears, E. R. (1944). The artificial synthesis of Triticum spelta. *Rec. Genet. Soc. Am.*, 26–27.

Mochida, K., Yoshida, T., Sakurai, T., Ogihara, Y., and Shinozaki, K. (2009). TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics. *Plant Physiol.* 150, 1135–46. doi:10.1104/pp.109.138214.

Mori, N., Miyashita, N. T., Terachi, T., and Nakamura, C. (1997). Variation in coxII intron in the wild ancestral species of wheat. *Hereditas*, 281–288.

Murtagh, F. (1985). *Multidimensional clustering algorithms. in COMPSTAT Lectures 4*. Physica-Verlag Available at: https://books.google.it/books/about/Multidimensional_clustering_algorithms.html?id=1RbvAAAAMAAJ&redir_esc=y [Accessed August 31, 2018].

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U. S. A.* 70, 3321–3. doi:10.1073/PNAS.70.12.3321.

Nesbitt, M., and Samuel, D. (1996). From Staple Crop to Extinction? The Archaeology and History of the Hulled Wheats. *Hulled Wheats*, 41–100.

Nishikawa, K. (1983). Species relationship of wheat and its putative ancestors as viewed from isozyme variation. *Proc. 6th Int. Wheat Genet. Symp.*, 59–63.

Oliveira, H. R., Hagenblad, J., Leino, M. W., Leigh, F. J., Lister, D. L., Penã-Chocarro, L., et al. (2014). Wheat in the Mediterranean revisited – tetraploid wheat landraces assessed with elite bread wheat Single Nucleotide Polymorphism markers. *BMC Genet.* 15, 54. doi:10.1186/1471-2156-15-54.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., et al. (2007). The TIGR Rice Genome Annotation Resource: Improvements and new features. *Nucleic Acids Res.* 35. doi:10.1093/nar/gkl976.

Ozkan, H., Brandolini, A., Pozzi, C., Effgen, S., Wunder, J., and Salamini, F. (2005). A reconsideration of the domestication geography of tetraploid wheats. *Theor. Appl. Genet.* 110, 1052–1060. doi:10.1007/s00122-005-1925-8.

Özkan, H., Willcox, G., Graner, A., Salamini, F., and Kilian, B. (2010). Geographic distribution and domestication of wild emmer wheat (Triticum dicoccoides). *Genet. Resour. Crop Evol.* 58, 11–53. doi:10.1007/s10722-010-9581-5.

Paradis, E. (2010). pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* 26, 419–420. doi:10.1093/bioinformatics/btp696.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20, 289–290. doi:10.1093/bioinformatics/btg412.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., et al. (2009). The Sorghum bicolor genome and the diversification of grasses. *Nature* 457, 551–556. doi:10.1038/nature07723.

Patterson, N., Price, A. L., and Reich, D. (2006). Population Structure and Eigenanalysis. *PLoS Genet.* 2, e190. doi:10.1371/journal.pgen.0020190.

Paux, E., Legeai, F., Guilhot, N., Adam-Blondon, A. F., Alaux, M., Salse, J., et al. (2008). Physical mapping in large genomes: Accelerating anchoring of BAC contigs to genetic maps through in silico analysis. *Funct. Integr. Genomics* 8, 29–32. doi:10.1007/s10142-007-0068-1.

Pearce, S., Saville, R., Vaughan, S. P., Chandler, P. M., Wilhelm, E. P., Sparks, C. A., et al. (2011). Molecular characterization of Rht-1 dwarfing genes in hexaploid wheat. *Plant Physiol.* 157, 1820–31. doi:10.1104/pp.111.183657.

Peng, J. H., Sun, D., and Nevo, E. (2011). Domestication evolution, genetics and genomics in wheat. *Mol. Breed.* 28, 281–301. doi:10.1007/s11032-011-9608-4.

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., and Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* 11, 1650–1667. doi:10.1038/nprot.2016.095.

Pingault, L., Choulet, F., Alberti, A., Glover, N., Wincker, P., Feuillet, C., et al. (2015). Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome. *Genome Biol.* 16, 29. doi:10.1186/s13059-015-0601-9.

Qanbari, S., and Simianer, H. (2014). Mapping signatures of positive selection in the genome of livestock. *Livest. Sci.* 166, 133–143. doi:10.1016/j.livsci.2014.05.003.

R Core Team (2013). R: A Language and Environment for Statistical Computing. *R Found. Stat. Comput. Vienna, Austria*. Available at: http://www.r-project.org/ [Accessed August 29, 2018].

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913–918. doi:10.1038/nature06250.

Sabot, F., Guyot, R., Wicker, T., Chantret, N., Bastien, A. E., Ae, L., et al. (2005). Updating of transposable element annotations from large wheat genomic sequences reveals diverse activities and gene associations. doi:10.1007/s00438-005.

Sakamura, T. (1918). Kurze Mitteilung über die Chromosomenzahlen und die Verwandtschaftsverhältnisse der Triticum-Arten. *Bot Mag Tokyo* 31, 151–154.

Salamini, F., Özkan, H., Brandolini, A., Schäfer-Pregl, R., and Martin, W. (2002). Genetics and geography of wild cereal domestication in the near east. *Nat. Rev. Genet.* 3, 429–441. doi:10.1038/nrg817.

Sallares, R., and Brown, T. A. (1999). PCR-based analysis of the intergenic spacers of the Nor loci on the genomes of Triticum diploids and polyploids. *Genome*, 116–128.

Salmanowicz, B. P., and Dylewicz, M. (2007). Identification and characterization of high-molecular-weight glutenin genes in Polish triticale cultivars by PCR-based DNA markers. *J. Appl. Genet.* 48, 347–357. doi:10.1007/BF03195231.

Salse, J., Chagué, V., Bolot, S., Magdelenat, G., Huneau, C., Pont, C., et al. (2008). New insights into the origin of the B genome of hexaploid wheat: Evolutionary relationships at the SPA genomic region with the S genome of the diploid relative Aegilops speltoides. *BMC Genomics* 9, 555. doi:10.1186/1471-2164-9-555.

Sax, K., and Sax, J. (1924). Chromosome behavior in a genus cross. *Genetics*, 954–464.

Scheet, P., and Stephens, M. (2006). A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. Available at: www.ajhg.org [Accessed September 7, 2018].

Schilling, A., Abaye, A., Griffeya, C., Branna, D., Alleya, M., and Pridgena, T. (2003). Adaptation and Performance of Winter Durum Wheat in Virginia. *Agron J.* 95, 642–651.

Schulz, A. (1913). *Die geschichte der kultivierten getreide.* Dogma.

Shi, X., and Ling, H.-Q. (2018). Current advances in genome sequencing of common wheat and its ancestral species. *Crop J.* 6, 15–21. doi:10.1016/J.CJ.2017.11.001.

Shorinola, O., Bird, N., Simmonds, J., Berry, S., Henriksson, T., Jack, P., et al. (2016). The wheat Phs-A1 pre-harvest sprouting resistance locus delays the rate of seed dormancy loss and maps 0.3 cM distal to the PM19 genes in UK germplasm. *J. Exp. Bot.* 67, 4169–78. doi:10.1093/jxb/erw194.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7, 539. doi:10.1038/msb.2011.75.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi:10.1093/bioinformatics/btv351.

Soriano, J. M., Villegas, D., Aranzana, M. J., García del Moral, L. F., and Royo, C. (2016). Genetic Structure of Modern Durum Wheat Cultivars and Mediterranean Landraces Matches with Their Agronomic Performance. *PLoS One* 11, e0160983. doi:10.1371/journal.pone.0160983.

Steuernagel, B., Jupe, F., Witek, K., Jones, J. D. G., and Wulff, B. B. H. (2015). NLR-parser: rapid annotation of plant NLR complements. *Bioinformatics* 31, 1665–1667. doi:10.1093/bioinformatics/btv005.

Szpiech, Z. A., and Hernandez, R. D. (2014). selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. *Mol. Biol. Evol.* 31, 2824–2827. doi:10.1093/molbev/msu211.

Takenaka, S., and Kawahara, T. (2012). Evolution and dispersal of emmer wheat (Triticum sp.) from novel haplotypes of Ppd-1 (photoperiod response) genes and their surrounding DNA sequences. *Theor. Appl. Genet.* 125, 999–1014. doi:10.1007/s00122-012-1890-y.

The International Brachypodium Initiative (2010). Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* 463, 763–768. doi:10.1038/nature08747.

The UniProt Consortium (2015). UniProt: a hub for protein information. *Nucleic Acids Res.* 43, D204-12. doi:10.1093/nar/gku989.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi:10.1038/nprot.2012.016.Differential.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28. doi:10.1038/nbt.1621.

Uauy, C. (2017). Wheat genomics comes of age. *Curr. Opin. Plant Biol.* 36, 142–148. doi:10.1016/j.pbi.2017.01.007.

Uauy, C., Distelfeld, A., Fahima, T., Blechl, A., and Dubcovsky, J. (2006). A NAC Gene regulating senescence improves grain protein, zinc, and iron content in wheat. *Science* 314, 1298–301. doi:10.1126/science.1133649.

van Slageren, M. W. (1994). Wild wheats: a monograph of Aegilops L. and Amblyopyrum (Jaub. & Spach). *Wageningen Agric. Univ. Pap* 7.

Wang, G.-Z., Matsuoka, Y., and Tsunewaki, K. (2000). Evolutionary features of chondriome divergence in Triticum (wheat) and Aegilops shown by RFLP analysis of mitochondrial DNAs. *TAG Theor. Appl. Genet.* 100, 221–231. doi:10.1007/s001220050030.

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., et al. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol. J.* 12, 787–796. doi:10.1111/pbi.12183.

Weir, B. S., and Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure.

Wilhelm, E. P., Turner, A. S., and Laurie, D. A. (2009). Photoperiod insensitive Ppd-A1a mutations in tetraploid wheat (Triticum durum Desf.). *Theor. Appl. Genet.* 118, 285–294. doi:10.1007/s00122-008-0898-9.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10, 515–34. doi:10.1093/biostatistics/kxp008.

Wu, T. D., Reeder, J., Lawrence, M., Becker, G., and Brauer, M. J. (2016). "GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality," in (Humana Press, New York, NY), 283–334. doi:10.1007/978-1-4939-3578-9_15.

Wu, T. D., and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21, 1859–1875. doi:10.1093/bioinformatics/bti310.

Xie, D. W., Wang, X. N., Fu, L. S., Sun, J., Zheng, W., and Li, Z. F. (2015). Identification of the trehalose-6-phosphate synthase gene family in winter wheat and expression analysis under conditions of freezing stress. Available at: http://www.ncbi.nlm.nih.gov/gorf/ [Accessed September 7, 2018].

Xu, Q., Xu, J., Liu, C. L., Chang, C., Wang, C. P., You, M. S., et al. (2008). PCR-based markers for identification of HMW-GS at Glu-B1x loci in common wheat. *J. Cereal Sci.* 47, 394–398. doi:10.1016/j.jcs.2007.05.002.

Yan, L., Fu, D., Li, C., Blechl, A., Tranquilli, G., Bonafede, M., et al. (2006). The wheat and barley vernalization gene VRN3 is an orthologue of FT. *Proc. Natl. Acad. Sci. U. S. A.* 103, 19581–6. doi:10.1073/pnas.0607142103.

Yan, L., Loukoianov, A., Tranquilli, G., Helguera, M., Fahima, T., and Dubcovsky, J. (2003). Positional cloning of the wheat vernalization gene VRN1. *Proc. Natl. Acad. Sci. U. S. A.* 100, 6263–8. doi:10.1073/pnas.0937399100.

ZADOKS, J. C., CHANG, T. T., and KONZAK, C. F. (1974). A decimal code for the growth stages of cereals. *Weed Res.* 14, 415–421. doi:10.1111/j.1365-3180.1974.tb01084.x.

Zhang, W., Chen, S., Abate, Z., Nirmala, J., Rouse, M. N., and Dubcovsky, J. (2017). Identification and characterization of Sr13, a tetraploid wheat gene that confers resistance to the Ug99 stem rust race group. *Proc. Natl. Acad. Sci. U. S. A.* 114, E9483–E9492. doi:10.1073/pnas.1706277114.

Zhang, W., and Dubcovsky, J. (2008). Association between allelic variation at the Phytoene synthase 1 gene and yellow pigment content in the wheat grain. *Theor. Appl. Genet.* 116, 635–645. doi:10.1007/s00122-007-0697-8.

Zhang, W., Gianibelli, M. C., Rampling, L. R., and Gale, K. R. (2004). Characterisation and marker development for low molecular weight glutenin genes from Glu-A3 alleles of bread wheat ( Triticum aestivum. L). *TAG Theor. Appl. Genet.* 108, 1409–1419. doi:10.1007/s00122-003-1558-8.

Zhang, Y., Li, D., Zhang, D., Zhao, X., Cao, X., Dong, L., et al. (2018). Analysis of the functions of *TaGW2* homoeologs in wheat grain weight and protein content traits. *Plant J.* 94, 857–866. doi:10.1111/tpj.13903.

Zhang, Y., Miao, X., Xia, X., and He, Z. (2014). Cloning of seed dormancy genes (TaSdr) associated with tolerance to pre-harvest sprouting in common wheat and development of a functional marker. *Theor. Appl. Genet.* 127, 855–866. doi:10.1007/s00122-014-2262-6.

Zhang, Z., Belcram, H., Gornicki, P., Charles, M., Just, J., Huneau, C., et al. (2011). Duplication and partitioning in evolution and function of homoeologous Q loci governing domestication

characters in polyploid wheat. *Proc. Natl. Acad. Sci. U. S. A.* 108, 18737–42. doi:10.1073/pnas.1110552108.

Zhao, G., Zou, C., Li, K., Wang, K., Li, T., Gao, L., et al. (2017). The Aegilops tauschii genome reveals multiple impacts of transposons. *Nat. Plants* 3, 946–955. doi:10.1038/s41477-017-0067-8.

Zikhali, M., Wingen, L. U., and Griffiths, S. (2016). Delimitation of the Earliness per se D1 (Eps-D1) flowering gene to a subtelomeric chromosomal deletion in bread wheat (Triticum aestivum). *J. Exp. Bot.* 67, 287–99. doi:10.1093/jxb/erv458.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics* 29, 2669–2677. doi:10.1093/bioinformatics/btt476.

Zimin, A. V., Puiu, D., Hall, R., Kingan, S., Clavijo, B. J., and Salzberg, S. L. (2017a). The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum. *bioRxiv*, 159111. doi:10.1101/159111.

Zimin, A. V., Puiu, D., Luo, M.-C., Zhu, T., Koren, S., Marçais, G., et al. (2017b). Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* 27, 787–792. doi:10.1101/gr.213405.116.

Zohary, D., Hopf, M., and Weiss, E. (2012). *Domestication of plants in the Old World : the origin and spread of domesticated plants in Southwest Asia, Europe, and the Mediterranean Basin*. Oxford University Press.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse Principal Component Analysis. doi:10.1198/106186006X113430.

Zou, H., Tzarfati, R., Hübner, S., Krugman, T., Fahima, T., Abbo, S., et al. (2015). Transcriptome profiling of wheat glumes in wild emmer, hulled landraces and modern cultivars. *BMC Genomics* 16, 777. doi:10.1186/s12864-015-1996-0.