nature genetics

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41588-019-0381-3

In the format provided by the authors and unedited.

Durum wheat genome highlights past domestication signatures and future improvement targets

Marco Maccaferri[1,2,28], Neil S. Harris[3,28], Sven O. Twardziok[4,28], Raj K. Pasam[5,28], Heidrun Gundlach[4], Manuel Spannagl[4], Danara Ormanbekova[1,4], Thomas Lux[4], Verena M. Prade[4], Sara G. Milner[6], Axel Himmelbach[6], Martin Mascher[6,7], Paolo Bagnaresi[8], Primetta Faccioli[8], Paolo Cozzi[9], Massimiliano Lauria[9], Barbara Lazzari[9], Alessandra Stella[9], Andrea Manconi[10], Matteo Gnocchi[10], Marco Moscatelli[10], Raz Avni[11], Jasline Deek[11], Sezgi Biyiklioglu[12], Elisabetta Frascaroli[1], Simona Corneti[1], Silvio Salvi[1], Gabriella Sonnante[13], Francesca Desiderio[8], Caterina Marè[8], Cristina Crosatti[8], Erica Mica[8], Hakan Özkan[14], Benjamin Kilian[15], Pasquale De Vita[2], Daniela Marone[2], Reem Joukhadar[5,16], Elisabetta Mazzucotelli[8], Domenica Nigro[17], Agata Gadaleta[18], Shiaoman Chao[19], Justin D. Faris[19], Arthur T. O. Melo[20], Mike Pumphrey[21], Nicola Pecchioni[2], Luciano Milanesi[10], Krystalee Wiebe[22], Jennifer Ens[22], Ron P. MacLachlan[22], John M. Clarke[22], Andrew G. Sharpe[23], Chu Shin Koh[23], Kevin Y. H. Liang[3], Gregory J. Taylor[3], Ron Knox[24], Hikmet Budak[12], Anna M. Mastrangelo[2,25], Steven S. Xu[19], Nils Stein[6], Iago Hale[20], Assaf Distelfeld[11], Matthew J. Hayden[5,26], Roberto Tuberosa[1], Sean Walkowiak[22], Klaus F. X. Mayer[4,27,29*], Aldo Ceriotti[9,29*], Curtis J. Pozniak[22,29*] and Luigi Cattivelli[8,29*]

[1]Department of Agricultural and Food Sciences, University of Bologna, Bologna, Italy. [2]CREA—Research Centre for Cereal and Industrial Crops, Foggia, Italy. [3]Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada. [4]Helmholtz Zentrum München, Plant Genome and Systems Biology, Neuherberg, Germany. [5]Agriculture Victoria, Agribio Centre for AgriBioscience, Bundoora, Victoria, Australia. [6]Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany. [7]German Centre for Integrative Biodiversity Research Halle-Jena-Leipzig, Leipzig, Germany. [8]CREA—Research Centre for Genomics and Bioinformatics, Fiorenzuola d'Arda, Italy. [9]National Research Council—Institute of Agricultural Biology and Biotechnology, Milano, Italy. [10]National Research Council—Institute of Biomedical Technologies, Segrate, Italy. [11]School of Plant Sciences and Food Security, Tel Aviv University, Tel Aviv, Israel. [12]Montana State University, Bozeman, MT, USA. [13]National Research Council—Institute of Biosciences and Bioresources, Bari, Italy. [14]Çukurova University, Faculty of Agriculture, Department of Field Crops, Adana, Turkey. [15]Global Crop Diversity Trust, Bonn, Germany. [16]Department of Animal, Plant and Soil Sciences, La Trobe University, Bundoora, Victoria, Australia. [17]Department of Soil, Plant and Food Sciences, University of Bari Aldo Moro, Bari, Italy. [18]Department of Agricultural and Environmental Science, University of Bari Aldo Moro, Bari, Italy. [19]United States Department of Agriculture, Agricultural Research Service, Edward T. Schafer Agricultural Research Center, Fargo, ND, USA. [20]Department of Agriculture, Nutrition, and Food Systems, University of New Hampshire, Durham, NH, USA. [21]Department of Crop and Soil Sciences, Washington State University, Pullman, WA, USA. [22]Crop Development Centre and Department of Plant Sciences, University of Saskatchewan, Saskatoon, Saskatchewan, Canada. [23]Global Institute for Food Security, University of Saskatchewan, Saskatoon, Saskatchewan, Canada. [24]Swift Current Research and Development Centre, Agriculture and Agri-Food Canada, Swift Current, Saskatchewan, Canada. [25]CREA—Research Centre for Cereal and Industrial Crops, Bergamo, Italy. [26]School of Applied Systems Biology, La Trobe University, Bundoora, Victoria, Australia. [27]School of Life Sciences Weihenstephan, Technical University Munich, Freising, Germany. [28]These authors contributed equally: M. Maccaferri, N. S. Harris, S. O. Twardziok, R. K. Pasam. [29]These authors jointly supervised this work: K. F. X. Mayer, A. Ceriotti, C. J. Pozniak, L. Cattivelli. *e-mail: k.mayer@helmholtz-muenchen.de; ceriotti@ibba.cnr.it; curtis.pozniak@usask.ca; luigi.cattivelli@crea.gov.it

nature genetics

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41588-019-0381-3

In the format provided by the authors and unedited.

Durum wheat genome highlights past domestication signatures and future improvement targets

Marco Maccaferri[1,2,28], Neil S. Harris[3,28], Sven O. Twardziok[4,28], Raj K. Pasam[5,28], Heidrun Gundlach[4], Manuel Spannagl[4], Danara Ormanbekova[1,4], Thomas Lux[4], Verena M. Prade[4], Sara G. Milner[6], Axel Himmelbach[6], Martin Mascher[6,7], Paolo Bagnaresi[8], Primetta Faccioli[8], Paolo Cozzi[9], Massimiliano Lauria[9], Barbara Lazzari[9], Alessandra Stella[9], Andrea Manconi[10], Matteo Gnocchi[10], Marco Moscatelli[10], Raz Avni[11], Jasline Deek[11], Sezgi Biyiklioglu[12], Elisabetta Frascaroli[1], Simona Corneti[1], Silvio Salvi[1], Gabriella Sonnante[13], Francesca Desiderio[8], Caterina Marè[8], Cristina Crosatti[8], Erica Mica[8], Hakan Özkan[14], Benjamin Kilian[15], Pasquale De Vita[2], Daniela Marone[2], Reem Joukhadar[5,16], Elisabetta Mazzucotelli[8], Domenica Nigro[17], Agata Gadaleta[18], Shiaoman Chao[19], Justin D. Faris[19], Arthur T. O. Melo[20], Mike Pumphrey[21], Nicola Pecchioni[2], Luciano Milanesi[10], Krystalee Wiebe[22], Jennifer Ens[22], Ron P. MacLachlan[22], John M. Clarke[22], Andrew G. Sharpe[23], Chu Shin Koh[23], Kevin Y. H. Liang[3], Gregory J. Taylor[3], Ron Knox[24], Hikmet Budak[12], Anna M. Mastrangelo[2,25], Steven S. Xu[19], Nils Stein[6], Iago Hale[20], Assaf Distelfeld[11], Matthew J. Hayden[5,26], Roberto Tuberosa[1], Sean Walkowiak[22], Klaus F. X. Mayer[4,27,29*], Aldo Ceriotti[9,29*], Curtis J. Pozniak[22,29*] and Luigi Cattivelli[8,29*]

[1]Department of Agricultural and Food Sciences, University of Bologna, Bologna, Italy. [2]CREA—Research Centre for Cereal and Industrial Crops, Foggia, Italy. [3]Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada. [4]Helmholtz Zentrum München, Plant Genome and Systems Biology, Neuherberg, Germany. [5]Agriculture Victoria, Agribio Centre for AgriBioscience, Bundoora, Victoria, Australia. [6]Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany. [7]German Centre for Integrative Biodiversity Research Halle-Jena-Leipzig, Leipzig, Germany. [8]CREA—Research Centre for Genomics and Bioinformatics, Fiorenzuola d'Arda, Italy. [9]National Research Council—Institute of Agricultural Biology and Biotechnology, Milano, Italy. [10]National Research Council—Institute of Biomedical Technologies, Segrate, Italy. [11]School of Plant Sciences and Food Security, Tel Aviv University, Tel Aviv, Israel. [12]Montana State University, Bozeman, MT, USA. [13]National Research Council—Institute of Biosciences and Bioresources, Bari, Italy. [14]Çukurova University, Faculty of Agriculture, Department of Field Crops, Adana, Turkey. [15]Global Crop Diversity Trust, Bonn, Germany. [16]Department of Animal, Plant and Soil Sciences, La Trobe University, Bundoora, Victoria, Australia. [17]Department of Soil, Plant and Food Sciences, University of Bari Aldo Moro, Bari, Italy. [18]Department of Agricultural and Environmental Science, University of Bari Aldo Moro, Bari, Italy. [19]United States Department of Agriculture, Agricultural Research Service, Edward T. Schafer Agricultural Research Center, Fargo, ND, USA. [20]Department of Agriculture, Nutrition, and Food Systems, University of New Hampshire, Durham, NH, USA. [21]Department of Crop and Soil Sciences, Washington State University, Pullman, WA, USA. [22]Crop Development Centre and Department of Plant Sciences, University of Saskatchewan, Saskatoon, Saskatchewan, Canada. [23]Global Institute for Food Security, University of Saskatchewan, Saskatoon, Saskatchewan, Canada. [24]Swift Current Research and Development Centre, Agriculture and Agri-Food Canada, Swift Current, Saskatchewan, Canada. [25]CREA—Research Centre for Cereal and Industrial Crops, Bergamo, Italy. [26]School of Applied Systems Biology, La Trobe University, Bundoora, Victoria, Australia. [27]School of Life Sciences Weihenstephan, Technical University Munich, Freising, Germany. [28]These authors contributed equally: M. Maccaferri, N. S. Harris, S. O. Twardziok, R. K. Pasam. [29]These authors jointly supervised this work: K. F. X. Mayer, A. Ceriotti, C. J. Pozniak, L. Cattivelli. *e-mail: k.mayer@helmholtz-muenchen.de; ceriotti@ibba.cnr.it; curtis.pozniak@usask.ca; luigi.cattivelli@crea.gov.it

Supplementary Note for:

Durum wheat genome highlights past domestication signatures and future improvement targets

Marco Maccaferri[1,2,^], Neil S. Harris[3,^], Sven O. Twardziok[4,^], Raj K. Pasam[5,^], Heidrun Gundlach[4], Manuel Spannagl[4], Danara Ormanbekova[1,4], Thomas Lux[4], Verena M. Prade[4], Sara G. Milner[6], Axel Himmelbach[6], Martin Mascher[6,7], Paolo Bagnaresi[8], Primetta Faccioli[8], Paolo Cozzi[9], Massimiliano Lauria[9], Barbara Lazzari[9], Alessandra Stella[9], Andrea Manconi[10], Matteo Gnocchi[10], Marco Moscatelli[10], Raz Avni[11], Jasline Deek[11], Sezgi Biyiklioglu[12], Elisabetta Frascaroli[1], Simona Corneti[1], Silvio Salvi[1], Gabriella Sonnante[13], Francesca Desiderio[8], Caterina Marè[8], Cristina Crosatti[8], Erica Mica[8], Hakan Özkan[14], Benjamin Kilian[15], Pasquale De Vita[2], Daniela Marone[2], Reem Joukhadar[5,16], Elisabetta Mazzucotelli[8], Domenica Nigro[17], Agata Gadaleta[18], Shiaoman Chao[19], Justin D. Faris[19], Arthur T. O. Melo[20], Mike Pumphrey[21], Nicola Pecchioni[2], Luciano Milanesi[10], Krysta Wiebe[22], Jennifer Ens[22], Ron P. MacLachlan[22], John M. Clarke[22], Andrew G. Sharpe[23], Chu Shin Koh[23], Kevin Y. H. Liang[3], Gregory J. Taylor[3], Ron Knox[24], Hikmet Budak[12], Anna M. Mastrangelo[2,25], Steven S. Xu[19], Nils Stein[6], Iago Hale[20], Assaf Distelfeld[11], Matthew J. Hayden[5,26], Roberto Tuberosa[1], Sean Walkowiak[22], Klaus F. X. Mayer[4,27,§], Aldo Ceriotti[9,§], Curtis J. Pozniak[22,§], Luigi Cattivelli[8,§,ç]

[1] Department of Agricultural and Food Sciences, University of Bologna, Bologna, Italy

[2] CREA-Research Centre for Cereal and Industrial Crops, Foggia, Italy

[3] Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada

[4] Helmholtz Zentrum München, Plant Genome and Systems Biology, Neuherberg, Germany

[5] Agriculture Victoria, Agribio Centre for AgriBioscience, Bundoora, Vic 3083, Australia

[6] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, 06466, Seeland, Germany

[7] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany

[8] CREA-Research Centre for Genomics and Bioinformatics, Fiorenzuola d'Arda, Italy

[9] National Research Council - Institute of Agricultural Biology and Biotechnology, Milano, Italy

[10] National Research Council - Institute of Biomedical Technologies, Segrate, Italy

[11] School of Plant Sciences and Food Security, Tel Aviv University, Tel Aviv, Israel

[12] Montana State University, Bozeman, MT, USA

[13] National Research Council - Institute of Biosciences and Bioresources, Bari, Italy

[14] Çukurova University, Faculty of Agriculture, Department of Field Crops, Adana, Turkey

[15] Global Crop Diversity Trust, Bonn, Germany

[16] Department of Animal, Plant and Soil Sciences, La Trobe University, Bundoora, VIC, Australia

[17] Department of Soil, Plant and Food Sciences, University of Bari Aldo Moro, Bari, Italy

[18] Department of Agricultural and Environmental Science, University of Bari Aldo Moro, Bari, Italy

[19] United States Department of Agriculture, Agricultural Research Service, Edward T. Schafer Agricultural Research Center, Fargo, ND, USA

[20] Department of Agriculture, Nutrition, and Food Systems, University of New Hampshire, USA

[21] Department of Crop and Soil Sciences, Washington State University, Pullman, WA, USA

[22] Crop Development Centre and Department of Plant Sciences, University of Saskatchewan, Saskatoon, Saskatchewan, Canada
[23] Global Institute for Food Security, University of Saskatchewan, Saskatoon, Saskatchewan, Canada
[24] Swift Current Research and Development Centre, Agriculture and Agri-Food Canada, Swift Current Saskatchewan, Canada
[25] CREA-Research Centre for Cereal and Industrial Crops, Bergamo, Italy
[26] School of Applied Systems Biology, La Trobe University, Bundoora, Vic 3083, Australia
[27] School of Life Sciences Weihenstephan, Technical University Munich, Freising, Germany

^ These authors contributed equally to the work.
§ These authors jointly coordinated this work and to whom correspondence should be addressed: k.mayer@helmholtz-muenchen.de; ceriotti@ibba.cnr.it; curtis.pozniak@usask.ca; luigi.cattivelli@crea.gov.it
ç Senior Project Coordinator

This file includes:

Additional Materials and Methods

Supplementary Text

References

Supplementary Figures: 1 to 39

Supplementary Tables: 1 to 24

Captions for Supplementary Data Sets 1 to 13

Other Supplementary Materials for this manuscript includes:

Supplementary Data Sets 1 to 13.

1. Additional Materials and Methods

1.1. Gene content, genome annotation and pattern of gene expression

1.1.1. RNA sequencing

An initial set of 57 durum wheat samples from plants of cultivar Svevo at different developmental stages and subjected to diverse treatments were collected and stored at -80°C. Total RNAs were extracted using Direct-zol™ RNA miniprep kit (Zymo Research) for most tissues and TRIZOL reagent (Invitrogen) with minor modifications for ovaries, anthers and developing grain. Total RNA samples were pooled in 9 batches grouping samples belonging to similar tissue/treatments. Additional samples were collected for a deeper analysis of specific organ/developmental stages: i) leaf and root seedlings, anthers + ovaries, grains at milk and dough stage from Svevo; ii) leaf and root seedlings, and grains at dough stage of twelve varieties representing different breeding steps; iii) caryopses at six different developmental stages (3, 5, 11, 16, 21 and 30 days after anthesis) from Svevo and Cappelli. All samples were collected and stored at -80°C and used for RNA extraction. All samples are listed in Supplementary Table 5 and detailed in Supplementary Data Set 7. 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 Truseq SBS kit v3.

1.1.2. Gene annotation pipeline

Both DW (Svevo) and WEW (Zavitan) where annotated with the same pipeline and parameters as described below to avoid technical differences solely based on different detection methods. The resulting WEW gene annotation version 2 (67,182 HC genes) thus differs from version 1 published previously with 65,012 HC genes[1].

The annotation pipeline combined evidence from protein reference sequences and gene expression data to predict transcript sequences on the genome assembly. Open reading frames were then predicted on the potential transcript structures and final classification yielded a set of high confidence genes.

We used the spliced alignment tool *Genomethreader* (version 1.6.6)[2] to align protein sequences from related grass species *Setaria italica*[3], *Brachypodium distachyon*[4], *Oryza sativa* L.[5], *Sorghum*

bicolor[6] as well as protein sequences from *Arabidopsis thaliana*[7] and all annotated protein sequences from the Triticeae tribe to the DW assemblies. The Triticeae protein sequences were downloaded from UniProt database on October 05th 2016 and thereby all sequences were filtered for being marked as complete protein sequences and then clustered by 100% identity. This set included validated protein sequences from Swissprot as well as predicted protein sequences from species including *Triticum aestivum*, *Aegilops tauschii* and *Hordeum vulgare*[8]. We applied *Genomethreader* (arguments: -startcodon –stopcodon -species rice –gcmincoverge 70 –prseeldength 7 prdist 4 –gffout) on each pseudomolecule sequence separately in order to reduce memory requirement per application.

Furthermore, we used *HISAT2* (version 2.0.4, parameter: --dta)[9] to align multiple sets of RNA-seq data to the assemblies. Data sets included expression data from DW, WEW as well as hexaploid bread wheat (Supplementary Table 5; SRA accession: SRP149116). Thereby, samples included a wide variety of different tissues, environmental and stress conditions. We used *Stringtie* (version 1.2.3)[9] to assemble mapped reads into transcript sequences for each dataset separately. Thereby, we configured *Stringtie* (parameter: -m 150 –t –f 0.3) to include only transcript sequences with a minimum size of 150 bp and to include only isoforms whose expression was at least 30% of main isoform. Finally, we also included full length cDNA sequences from public databases as well as publicly available IsoSeq sequences from six different bread wheat tissues (leaf, root, seedling, seed, spike and stem)[10] into the gene annotation pipeline. We used GMAP[11] (version 06/30/2016, parameter: -K 50000) to align all sequences to the assemblies and thereby we restricted maximum intron size to 50,000 bp.

Transcript predictions from all types of evidence were then combined using *Cuffcompare* from *Cufflinks* software suite[12]. Finally, we used *Stringtie* (version 1.2.3, parameter: --merge –m 150) again to merge overlapping transcript sequences and to remove redundant transcript sequences and fragments. Transcript sequences were then extracted from the gtf file using cufflinks_gtf_to_cdns_fasta.pl script from the *Transdecoder* package (version 3.0.0). We then used TransDecoder.LongOrf (parameter: -p 0) to extract the longest open reading frames for each transcript sequence and to translate them into predicted protein sequences. These potential protein sequences were then compared to a reference protein database using BLASTP (NCBI blast 2.3.0+, parameter: -max_target_seqs 1-evalue 1e-05) and checked for abundance of known protein domains using *Hmmscan* (version 3.1b2). Both output tables were used as queries into TransDecoder.Predict to select a single best open reading frame for each transcript structure. The final gene predictions were combined with protein structure prediction from *Genomethreader* to compensate for potentially differentiating open reading frame predictions by the two tools.

1.1.3. Confidence classification and functional annotation

To differentiate the predicted protein sequences into (i) canonical proteins, (ii) non-coding transcripts, (iii) incomplete genes and (iv) transposable elements, we applied a confidence classification to all potential protein/transcript sequences. Therefore, we used all potential protein sequences in BLAST against two protein reference databases. The first database contained all validated *Magnoliophyta* protein sequences from Uniprot and the second database contained all annotated *Poaceae* protein sequences from Uniprot (both downloaded on August 03rd 2016). The second database was further filtered to contain complete protein sequences only. Furthermore, to filter out transposons, we used all potential protein sequences in BLAST against the translated TREP[13] (release 16, http://botserv2.uzh.ch/kelldata/trep-db/index.html) database.

Based on the E-value distribution of best hits, those with an E-value below 10^{-10} were considered as significant. To filter out fragmented alignments due to fragmented protein annotations or local alignments of domains, we filtered the significant alignments for query and subject coverage. For comparison with the protein databases, we considered only alignments with query and subject coverage of at least 90% as representative hits and for the comparison with the TREP database we considered alignments with a query coverage of at least 75% as representative hits. Based on representative blast hits and completeness of protein sequences (annotated start and stop codon), all potential transcript sequences were then classified into two confidence classes including 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 not annotated completely or that showed only insufficient homology to reference proteins or were likely candidates for transposons.
- LC1: incomplete coding sequence but significant match to reference protein sequence
- LC2: no significant match to reference protein sequence but complete coding sequence
- REP: match to transposon elements database

High confidence genes were defined as those loci that contained at least one high confidence transcript. All low confidence transcripts that were overlapping with a high confidence transcript were then removed.

Gene functions where annotated with the AHRD tool (Automated Assignment of Human Readable Descriptions, https://github.com/groupschoof/AHRD, version 3.3.3). AHRD scores blast hits taken from searches against different databases based on the trust put into these databases and the alignment quality. The blast hit descriptions are tokenized into informative words and a lexical analysis scores the tokens according to their frequency and the quality of the blast hits they occur in. Finally, the best scoring description is assigned. Together with InterProScan Runs (version 5.23-62.0) blast hits against the following three databases were used as AHRD input: Swiss-Prot (version 02-15-10), Arabidopsis Araport 11 (version 201606) and a TrEMBL (version 02-15-10) *Viridiplantae* subset. The Gene ontology (GO) assignments were derived from the Interpro to GO mapping (interpro.xml file from www.ebi.ac.uk/interpro/ download.html).

To study the gene expression pattern, RNA-seq libraries were aligned to the DW genome using *HISAT2* (version 2.0.4). The BAM files produced were filtered for reads that aligned concordantly exactly one time based on mapping quality >40. The transcript abundance was then calculated using *Stringtie* (version 1.2.3) and *Ballgown*[9].

1.1.4. Validation of the DW genome assembly and annotation

We performed two validation steps to evaluate the completeness of the DW genome assembly and to determine quality of the annotations. Initially, we used the *BUSCO* (Benchmarking Universal Single-Copy Orthologs) tool (version 2, Embyophyta odb9)[14] to determine the abundance of strongly conserved genes in the sets of all annotated genes and of HC genes. In addition, the predicted gene models in DW and WEW were verified using 216 experimentally-validated complete gene sequences kindly provided by Jorge Dubcovsky (University of California, Davis, CA, Supplementary Data Set 8). 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) from DW and WEW.

Furthermore, to validate the predicted protein sequences, we downloaded all available Triticeae protein sequences from the Uniprot database (downloaded on April 27th 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.

1.1.5. Repeat annotation

Basic *k-mer* defined repetitivity was calculated for all 20-mers along the 14 DW chromosomes (sliding window, one bp shift) with the program *Tallymer*[15] against an index of the complete durum wheat assembly or the respective sequence set (Supplementary Fig. 19A). The location of centromers was determined by distinct 20-mer frequency peaks reflecting the highly repetitive nature of the

tandemly arranged centromere components and reported in Supplementary Fig. 20 together with transposon and gene distributions.

Transposons were detected and classified by a homology search against the REdat_9.7_Triticeae and the PGSB transposon library[16]. The program *Vmatch* (www.vmatch.de), a fast and efficient matching tool suited for large and highly repetitive genomes, was used for this computationally intensive task with the following parameters: identity ≥70%, minimal hit length 75 bp, seed length 12 bp (exact commandline: -d -p -l 75 -identity 70 -seedlength 12 -exdrop 5). To obtain an overlap free annotation the *Vmatch* output was filtered for redundant hits via a priority-based approach from a score-sorted match list. Higher scoring matches where assigned first and shortened if they overlapped with an already assigned element and if their rest length was at least 50 bp and ≥ 10% of their hit length. All other matches were discarded.

The identification of full-length LTR retrotransposons (fl-LTRs) was based on the program *LTRharvest*[17], a *de novo* finder that scans the genome sequence for structural properties like long terminal repeats, primer binding sites and target site duplications. For the DW assembly *LTRharvest* reported 366,143 non-overlapping fl-LTR candidate sequences under the following parameter settings: overlaps best -seed 30 -minlenltr 100 -maxlenltr 2000 -mindistltr 3000 -maxdistltr 25000 -similar 85 -mintsd 4 -maxtsd 20 -motif tgca -motifmis 1 -vic 60 -xdrop 5 -mat 2 -mis -2 -ins -3 -del -3. All candidates were annotated for PfamA domains with *Hmmer3* [http://hmmer.org/[18]] and stringently filtered for canonical elements by the following criteria: *i)* presence of at least one typical retrotransposon domain (RT, RH, INT, GAG); *ii)* absence of gene Pfam domains; *iii)* strand consistency between domains and primer binding site, *iv)* tandem repeat content below 25%; *v)* long terminal repeat ≤ 25 of element length; *vi)* and an N content <5%.

1.1.6. MicroRNA sequencing and annotation

Starting from the nine total RNA pools used for RNA-Seq (Supplementary Data Set 7) small RNA libraries were constructed with the TruSeq Small RNA Sample Prep Kit (Illumina, San Diego, CA) according to the manufacturer's instructions. Briefly, 1 μg of total RNA was ligated with two adapters at 3' and 5' ends and reverse transcribed with SuperScript II RT (Invitrogen), then PCR-amplified (15 cycles). The cDNA libraries were purified on a 6% TBE PAGE and quality and concentration were evaluated with the Agilent 2100 Bioanalyzer DNA1000 assay. The nine prepared indexed libraries were evaluated as described for RNA sequencing and then put together as 10 μL of the pooled library at a final concentration of 2 nM for sequencing in 8 lanes of Illumina HiSeq2000 with a 50 nt Single-Read sequencing module.

Raw sequencing data were checked for quality with *FastQC* (version 0.11.4) and no quality filter was applied. Sequencing reads were then trimmed using the program *Cutadapt*[19] version 1.10 with the settings: --trim-n -a TGGAATTCTC –discard-untrimmed –m 15 –M 40. For each sample, trimmed reads were analysed with *ShortStack* version v.3.4[20] with settings –mmap r –mincov 20. The non-redundant set of loci expressed in at least one of the nine libraries was then produced and the putative corresponding precursor sequences were extracted and used in BLAST against all plant hairpin sequences (E-value $1.0e^{-5}$) and all mature miRNA sequences (E-value $1.0e^{-3}$) present in *mirBase* version 21. Manual curation of BLAST results distinguished between hairpins belonging to known plant miRNA families and putative durum wheat-specific sequences.

1.1.7. Long non-coding RNA annotation

Long non-coding RNAs (lncRNAs) annotation was performed as previously outlined[21] with some modifications. A set of criteria was applied to transcriptome assembly through analyzing the following features to select potential lncRNAs: *i)* length of transcripts; *ii)* homology to known protein-coding sequences; *iii)* homology to contaminants; *iv)* coding potential; *v)* open reading frame (ORF) size. As a first step, transcripts shorter than 200 nucleotides were eliminated whereas transcripts that met the length criteria were then assessed for their homology to known protein-coding transcripts and protein sequences using BLAST+ 2.6.0[22]. Databases used were as follows: Uniprot/Swissprot database (http://web.expasy.org/docs/swiss-prot_guideline.html) (BLASTX: -Evalue $1e^{-10}$; TBLASTN: -Evalue $1e^{-06}$); NCBI nr database (BLASTX: -Evalue $1e^{-03}$); Svevo high confidence coding sequences (BLASTN: -Evalue $1e^{-05}$); *T. aestivum*, *H. vulgare*, *B. distachyon*, *O. sativa* and *S. bicolor* UniGene coding sequences (https://www.ncbi.nlm.nih.gov/unigene) (BLASTN: -Evalue $1e^{-05}$); *T. aestivum* IWGSC high confidence coding sequences (BLASTN: -Evalue $1e^{-05}$); *T. aestivum* Uniprot protein sequences (http://www.uniprot.org/) (BLASTX: -Evalue $1e^{-10}$; TBLASTN: -Evalue $1e^{-06}$); *H. vulgare* protein sequences (http://plants.ensembl.org/) (BLASTX: -Evalue $1e^{-10}$; TBLASTN: -Evalue $1e^{-06}$); *B. distachyon* protein sequences (http://pgsb.helmholtz-muenchen.de/plant/brachypodium/)[23] (BLASTX: -Evalue $1e^{-10}$; TBLASTN: -Evalue $1e^{-06}$); *O. sativa* protein sequences (http://rapdb.dna.affrc.go.jp/download/irgsp1.html) (BLASTX: -Evalue $1e^{-10}$; TBLASTN: -Evalue $1e^{-06}$) and *S. bicolor* protein sequences (http://pgsb.helmholtz-muenchen.de/plant/sorghum/)[23] (BLASTX: -Evalue $1e^{-10}$; TBLASTN: -Evalue $1e^{-06}$). All transcripts that had similarity with the sequences in the databases were excluded. The remaining transcripts were examined for homology to previously identified rRNA, tRNA, snoRNA, snRNA sequences and also *T. aestivum* organellar sequences deposited at NCBI and ENA (European nucleotide archive -

https://www.ebi.ac.uk/ena) databases (BLASTN: -Evalue $1e^{-05}$) and only the transcripts with no similarity were chosen.

After the homology-based eliminations, the potential to encode proteins of the transcripts were calculated by using *CNCI* software (version 2, options: -m pl)[24]. Only the sequences marked as 'noncoding' were taken and subjected to ORF size prediction. *Transdecoder* (-m 100) was utilized to distinguish the transcripts with less than 100 amino acids ORF size. Additionally, transcripts with ORF sizes between 30 and 100 nucleotides were subjected to search for conserved protein domains with *Hmmer* (version 3.12.1) against Pfam domains[25].

Finally, all transcripts left after lncRNA identification criteria were mapped to Svevo chromosome sequences with *GMAP* software (version 2018-03-25; -n 1 --nofails -f 2 -x 0)[11]. Sequences that were mapped with *GMAP* score of 40 or higher were assessed for their canonical or non-canonical splice sites by using *Gffread* software (https://github.com/gpertea/gffread). To avoid transcripts with a potential similarity to a protein-coding gene, the reverse complements of the transcripts with non-canonical splice sites were mapped to the chromosomes again to eliminate sequences having canonical splice sites after reverse complement conversion. After exclusion of transcripts meeting the five criteria above described and splice site checking, the residual transcripts were defined as lncRNAs.

1.1.8. Annotation of prolamin seed storage genes

Automatically annotated glutenins, gliadins, and avenin-like sequences were retrieved from the Svevo genome browser (http://d-gbrowse.interomics.eu). Subsequently, BLAST analysis with sequences that were specific of each prolamin family/type was performed against the Svevo genome (http://d-annotator.interomics.eu) to recover sequences that may have been undetected by the gene annotation process. All sequences were manually examined and re-annotated by using the collaborative genomic annotation editor Apollo (http://d-annotator.interomics.eu). WEW wheat prolamins were identified through BLASTN against the Triticeae and Avena database (https://wheat.pw.usda.gov/GG3/wildemmer_blast) using as query sequences representative for each prolamin family. Subsequently, the identified prolamin sequences were manually inspected at the Zavitan Genome Browser (https://wheat.pw.usda.gov/GG3/jbrowse_Zavitan). Overall, 124 and 107 sequences were identified in the genome of Svevo and Zavitan, respectively (Supplementary Table 17). The correct gene classification for each annotated sequence was verified by BLASTN and BLASTX (https://blast.ncbi.nlm.nih.gov/Blast.cgi) homology search against the NCBI non-redundant nucleotide and protein sequences. In Supplementary Table 17 is reported the number of sequences that were classified as i) full-length open reading frame, ii) partial (incomplete sequences

> 300 bp) and potentially full-length open reading frame (incomplete sequences in which it is possible to recognize a start and stop codon), and iii) full-length and partial defective open reading frame (presence of premature stop codon or frame shift mutations in sequences > 300 bp).

1.1.9. NLR gene family organization in durum and wild emmer wheat

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[26]) to annotate the loci associated with Nucleotide-Binding Leucine-Rich Repeat domains (NLRs) both in DW and WEW. The pseudomolecules were first fragmented into 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*[26]. 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*[12] to identify possible novel loci not present in the transcriptome-based annotations.

1.1.10. Annotation of CpG islands and of transcription factors binding sites

CpG islands detection was performed with *CpGCluster*[27]. The analysis was run with the 50th percentile of the genomic CpG distance distribution (median distance) as the threshold distance, and with the cutoff P-value of 1e^{-5}.

Transcription factors binding sites (TFBS) and promoter binding elements search was run jointly with *Find Individual Motif Occurrences* (*FIMO*)[28]. A total of 231 TFBS known motifs were retrieved from the Jaspar CORE 2016 database Plantae section (http://jaspar.genereg.net), and four promoter binding element motifs (TATA box, GC box, CAAT box and Initiator) were retrieved from the Eukaryotic Promoter Database (http://epd.vital-it.ch/promoter_elements.php). Overall 235 motifs were used to scan 2,000 bp regions upstream the start codon of predicted protein-coding genes and microRNA loci for the presence of putative binding sites.

1.1.11. Annotation of plant functional non-tandem duplicated gene cluster

A pipeline was developed based on a chromosome sliding window conducting GO enrichment tests over adjacent GO–BP (Biological Process)-assigned genes. GO terms were assigned via Blast2GO PRO. A scanning master window of 24 GO Biological-process equipped High Confidence genes was defined and fine-tuning of cluster length and positioning was tested by defining 13 sub-windows with different combinations in gene number (from 6 to 24 GO-BP assigned genes) and position (from first to last gene in master window). In case of detection of tandem duplicated and/or

homologous genes, only one representative member of homologous genes was kept. All sub-windows were tested for GO enrichment via a hypergeometric test as implemented in *Bioconductor* package *GOstats*[29] (version 2.42.0). For GO enrichment call, stringent *p*-values $p \leq 10^{-5}$ (and up to $p \leq 10^{-8}$) were set. A minimum of 3 genes sharing the same GO-BP was set as a pre-requisite for candidate functional non-tandem duplicated gene cluster (FNTDC) call. The sub-window displaying, if any, the lowest *p*-value for enriched GO was then considered as the best approximation of functional cluster length and positioning associated to the parent master window region. The Universe set consisted of all high confidence, GO-BP equipped genes in durum wheat.

Prior to each GO enrichment testing, tandem duplicated genes were detected by an all-against all BLASTP (*blast2*; version 2.2.26, gapped alignment, default parameters). In the case of genes with more than one isoform, the longest isoform was considered as representative for the corresponding gene in blast analyses. When tandem duplicates were detected, only the first occurring member of such a family was kept as representative for subsequent GO enrichment testing. Four different settings for homology detection were conducted, the most tolerant settings (leading to a large fraction of detected tandem duplicates and thus to fewer called FNTDC) were 40% identity over at least a ratio of 0.4 (both alignment length to query length and alignment length to subject length). The 40% settings are expected to provide high rates of true positive FNTDC while leading to a significant number of false negatives. In fact, distinct, independent genes may be misclassified as tandem duplicates genes because of sharing some sequence homology in discrete parts, as expected e.g. for genes sharing common cofactor binding modules or even minimally differing genes as e.g. neo-functionalized ones which may synthesize different products following just one amino-acid substitution[30,31]. Three additional settings were tested, namely 70% identity and ratio 0.7; 90% identity and ratio 0.9 and 98% identity and ratio 0.98. The last setting led to virtually no exclusion of tandem genes and the called clusters can be considered as a large reference superset of all co-localized genes sharing a GO-BP (including nearly all genuine tandem duplicates).

1.2. Comparison between durum wheat and wild emmer genomes

1.2.1. Identification of durum and wild emmer specific genes and intact-gene-number variation

WEW (accession Zavitan) and DW (cultivar Svevo) HC genes were aligned and clustered using strict sequence homology criteria (alignment e value < 10^{-10}, overlap >75% and identity >75%). The resulting all *vs*. all blast matrix was clustered with the *R igraph* package (v1.0, igraph.org) using connected components (*netcluster* option) leading to 36,434 unigene gene clusters which were used for the analyses shown in Fig. 2a and 2b. Metrics describing the main scenarios of intact gene number variation, either due to large structural variations or to modifications of the gene integrity, are

described in Supplementary Table 10. Since the unigene classification was run with HC genes only, any mutation that rules out a gene from the HC class in Svevo or in Zavitan, lead to an asymmetric unigene distribution. Gene ontology terms (GO) enriched for Svevo genes belonging to balanced, more in Svevo, more in Zavitan and Svevo unique gene clusters are shown in Supplementary Fig. 3. The complete list of unigene clusters and their composition is reported in Supplementary Data Set 9.

To check if the lineage specific genes missing a close homoeolog in the HC gene set of the other genome are absent altogether from the sequence or only missing in the annotation, we mapped all genes occurring only in DW (4,811) or WEW (4,809) with high stringency (≥95 identity, ≥95 coverage) to the sequences of the other genome using BLAT (spliced alignments). All in all, we found hits for about 70 % of all DW specific genes on the WEW genome. Of them, 965 (1.4% of all DW genes) do not overlap with existing annotations and could be classified as missed genes in the WEW annotation if they display correct and long enough open reading frames. Another 1,225 genes map to an HC gene of the other genome, even though they were not clustered to their respective partners. They are probably too short to pass the 75% overlap limit used for clustering. The rest (1,095) correspond to either LC genes or pseudogenes. In summary most of the presence absence variations defined on the HC gene sets are not completely absent from the partner genome sequence. Around 70% of them still exist as structurally altered or degenerated counterparts on the other genome.

1.2.2. Identification of pseudogenes and gene fragments

The computational identification of pseudogenes was carried out by exploiting their sequence homology to functional genes. The pseudomolecules were split into batches for parallel processing. The CDS nucleotide sequences of all high-confidence gene isoforms that had no indication of being transposable element-related (397,624 isoforms from WEW and DW) were then mapped onto the genome sequences using BLAT[32] (minimal identity 70%, max. intron length 2,500 bp), which creates spliced alignments and thus recovers the exon-intron structures. Each potential hit has a parent gene, which was used to detect it. This parent gene does not have to be from the same species and is rather the most similar functional gene from either DW or WEW. All hits were filtered to have a length of at least 100 bp and at least one fragment with 50 bp. Gaps up to a size of 9 bp were closed and considered in the calculation of the sequence identity. Premature termination codons were then determined independently for each pseudogene 'exon', always starting in the correct frame compared to the parent gene.

The pseudogene candidates were then extensively filtered to remove transposable element genes, sequences with low information content or hits overlapping functional genes. All pseudogene candidates were first clustered using *CD-HIT*[33,34] (version 4.6.5 with 95% identity, 95% coverage) to

identify domains with high copy number. Sequences occurring more than 1,000 times in the genome were filtered due to the assumption of them being transposable element related. Gene self-hits as well as hits overlapping other genes were filtered. Nonspecific hits, as well as hits with low information content, were filtered using the WU-BLAST dust filtering[35] (default settings) and the *Tandem Repeats Finder*[36] (max. 65% masked, ≥ 50 base pairs remaining). Transposable element genes were identified by mapping the pseudogene sequence to the TREP database[37] and filtering sequences with a TREP hit covering at least 75% of their CDS with a minimal sequence identity of 90%. Finally, pseudogenes overlapping the transposable element annotation (≥ 75% overlap) were filtered. All combined filtering steps reduced the number of pseudogene candidates by up to 90%.

In case hits were overlapping, the longest hit was chosen as a representative for the locus. If multiple hits with the same length occurred, then the one with the highest sequence identity to its parent was chosen. Parent genes from the same species gene set were favored. If the representative covered less than 60% of the locus, then all hits shorter than half of the representative and overlapping with it were removed, as well as the hit with the shortest exon length but also the longest total length. This allowed the hit cluster to split up into multiple loci and newly determined representatives to be of good quality. Results of the pseudogene annotation are given in Supplementary Table 18.

1.2.3. Pseudogene classification and GO analysis

Intron sequences were used to classify pseudogenes as *duplicated* or *retroposed*. Fragmented pseudogenes often do not span over splice sites, rendering this type of classification impossible for them. For the intron loss/retention criterion, we defined five pseudogene classes: *i)* duplicated pseudogenes still containing introns at each covered splice site; *ii)* retroposed or processed pseudogenes that have lost all introns; *iii)* chimeric pseudogenes that have both retained and lost introns; *iv)* single-exon gene pseudogenes from genes or isoforms with only one exon; *v)* fragmented pseudogenes which do not sufficiently cover a splice site. A splice site is only covered, if at least 10 bp of the exons on either side are present in the duplicate. The gap has to be at least 30 bp long to be considered a duplicated intron.

Pseudogenes were associated with the Gene Ontology (GO) terms of their parent genes. Under- or over-represented GO terms in the pseudogene set of each wheat species compared to the combined pseudogene set (universe) were determined using the free open-source GOstats R package[29] with a *p*-value cutoff of 0.05. For this, only GO terms occurring at least ten times in the universe were used.

1.2.4. Orthologous gene family analysis in the durum wheat genome

Gene families where analysed with *OrthoMCL* (version 2.0, www.orthomcl.org) in three different settings: (i) DW A, B subgenome-located and unassigned genes, (ii) DW subgenomes and three other grass genomes (barley, *Brachypodim* and rice) and (iii) DW and WEW A and B subgenomes. Prior to the analyses splice variants were removed from all data sets, keeping the representative/longest protein sequence prediction, and data sets were filtered for internal stop codons and incompatible reading frames. The first step for each setting was the calculation of pairwise sequence similarities between all input protein sequences using BLASTP with an e-value cut-off of $1e^{-05}$. Markov clustering of the resulting similarity matrix was subsequently used to define the ortholog cluster structure, using an inflation value (-I) of 1.5 (*OrthoMCL* default).The input datasets for the first setting consisted of the following high confidence (HC) genes DW A subgenome HC (31,718 HC genes), DW B subgenome (32,275 HC genes) and DW of unknown origin (2,566 HC genes). A total of 53,207 coding sequences from these three datasets were clustered into 20,366 gene families. An overview of the cluster structure is shown in Supplementary Fig. 21.

The second setting comprised DW HC genes and the annotated gene sets of four grasses from diverse grass sub-families. The six input datasets were: DW A genome (31,718 HC genes), DW B genome (32,275 HC genes) DW genes of unknown origin (2,566 HC genes), *Hordeum vulgare* HC IBSCv1.0 (39,734 genes)*, Brachypodium distachyon* v2.1 (31,694 genes) *and* rice MSU7.0 (39,049 genes). The coding sequences from these species were clustered into 26,849 gene families. An overview of the cluster structure is shown in Supplementary Fig. 22, where the genes from the durum unknown subgenome origin are not shown together with the other entities.

The third *OrthoMCL* setting contained DW and WEW HC genes separated by subgenome origin. The six input datasets were: DW A genome (hig: (31,718 HC genes), DW B genome (32,275 HC genes) DW genes of unknown origin (2,566 HC genes); WEW A genome (32,372 HC genes), WEW B genome (32,661 HC genes) *and* WEW genes of unknown origin (2,149 HC genes). The *OrthoMCL* analyses identified 29,875 gene families with at least two members. An overview of the cluster structure is shown in Supplementary Fig. 23, genes with unknown subgenome origin are not shown.

This last setting is similar to the unigene analyses of Svevo and Zavitan described in section *1.2.1.* The unigene clusters are also based on a BLASTP gene similarity matrix, although with more stringent parameters: only e-value $1e^{-5}$ for *OrthoMCL* versus e-value $1e^{-10}$ plus an additional length and identity filter (overlap>75% and identity>75%) for the unigene clusters. In addition, the unigene clusters where not calculated with *OrthoMCL* (Markov Cluster algorithm), but with a graph-based method (*R igraph*) method. In spite of the method differences, the outcome is surprisingly robust and roughly comparable between the two approaches. Similar values are for instance the number of

clusters (*orthoMCL* 29,875 vs unigenes 28,794 vs) and the unigene groups with identical copy numbers (23,427 vs 21,774). However, *orthoMCL* results in more singletons (29% vs 21%) and has on average smaller clusters (4.1 vs 4.4 members).

1.2.5. Genome-wide identification of sequence variants with likely high impact on wild emmer gene function

Leveraging the high-quality references for both wild emmer (accession Zavitan) and cultivated durum (cv. Svevo) wheat, we developed a genome-wide atlas of putative functional variants between the two species as a general resource to guide trait dissection. To create such an atlas, we first aligned all Svevo CDS transcripts to the annotated Zavitan genome using *BWA*[38] and called variants (SNPs and indels) using the *SAMTools/BCFTools* pipeline[39]. All identified variants were then functionally annotated with Ensembl's *Variant Effect Predictor* (*VEP*) pipeline[40], using the offline mode with default parameters and only those variants classified as likely HIGH IMPACT by *VEP* were retained (e.g. start lost, stop gained, stop lost, and frameshift variants). Because our specific interest here is in those wild emmer genes whose functional modification underpins the domesticated/cultivated phenotype, we relied on the Zavitan annotation for functional inference. To precisely locate the corresponding positions (bp) of identified high impact functional variants in the Svevo genome coordinate system, we developed a custom Perl script which sampled 10 kb of genomic sequence adjacent to each variant and aligned those sequences to the Svevo genome using *BWA*. Very conservative parameters were applied in this alignment, and only those alignments exhibiting the highest possible alignment score (250) were retained. The detailed results of this analysis, in which 597 putative high impact variants were identified and manually inspected (~43/chromosome), are presented in Supplementary Data Set 10 and a graphical depiction of the distribution of those variants across the Svevo genome is presented in Supplementary Fig. 1. It is possible that the high impact variants identified by this analysis are simply genotype-specific (variation between two individuals) rather than species-specific (variation attributable to speciation), so this atlas of functional variants should be considered as a resource to guide candidate gene hypotheses, as illustrated below in the context of the *TdHMA3-B1* region (section *2.3.2.*).

1.3. Genetic mapping and analysis of genetic diversity in the Global Tetraploid wheat Collection

1.3.1. Genetic mapping, marker projection on the durum wheat genome and genome-wide investigation of recombination rate

The availability of both a high-density Svevo × Zavitan genetic map and Svevo linkage disequilibrium (LD) pattern based on the physical assembly allowed investigating the variation in recombination rate, gene density and gene diversity in tetraploid wheat germplasm. The high-density Svevo × Zavitan reference genetic map included 939,536 Genotyping by Sequencing (GBS) marker tags and 14,088 single nucleotide polymorphisms (SNPs) from the wheat Illumina iSelect 90K SNP array[1,41]. An additional resource consisting of 17 tetraploid wheat genetic maps (Supplementary Table 19) was considered for genetic mapping and anchoring to the Svevo assembly of genetically mapped SNPs, Diversity Array Technology (DArTs®), genomic- and genic- simple sequence repeats (SSRs), and expressed sequence tags (ESTs) or sequence tagged sites (STS). The genetic maps were produced following a pipeline including: *i)* scripts (https://github.com/plantinformatics/Durum_iSelect_90kSNP_GenotypeCalling) for genotype calling in unrelated samples, sample cluster assignment, confidence score estimates, and final genotype call from Illumina raw data project files; *ii)* quality check and filtering of genotype calls; *iii)* marker grouping and ordering in *MST-map*[42]. The Script parameters used for genotype calling were as follows:

-d 3, sample must be within 3 standard deviations of a known cluster position.

-r 0.8, minimum confidence score that sample belongs to the cluster to which it was assigned versus the next closest cluster; a value of 1 indicates highest confidence.

-g 0.4, minimum sample genotype call rate before reporting a SNP.

The genotype call outputs for the mapping population data-sets were filtered for SNP call rate (minimum 90%), cross-over rate, presence of identical samples. Marker grouping and linkage mapping were performed in *MST-map*[42] (http://www.mstmap.org/), which can efficiently determine the correct order of markers by computing the Minimum Spanning Tree of an associated graph. For each map, the linkage groups were aligned and oriented based on the tetraploid SNP consensus map[43].

The 17 genetic maps, including a revisited Svevo × Zavitan SNP map with 16,372 mapped SNPs, provided genetic and physical map positions for 38,340 Illumina iSelect SNP, 1,341 DArT, 835 SSR, and 109 STS markers as reported in Supplementary Data Set 11 and in the websites hosting the Svevo genome sequence browser (Interomics website, https://www.interomics.eu/durum-wheat-genome and GrainGenes, https://wheat.pw.usda.gov/GG3/jbrowse_Durum_Svevo).

After initial assembling, ordering and orienting of scaffolds by Hi-C, the sequences of publicly

available Illumina and Affymetrix SNP, SSR, EST-SSR and EST-STS markers were positioned on the DW physical assembly by BLASTN. *NCBI-BLAST* version 2.3.0 was used for BLAST with the following thresholds/parameters: E value = 10^{-10} threshold, coverage ≥ 75%, filtering for the five best hits based on the highest score. To address issues of homoeologs/paralogs, the BLAST results were cross-checked with the genetic mapping results from the 17 tetraploid wheat genetic maps.

The Svevo × Zavitan reference map (filtered for GBS and SNP markers with cross-matched physical and genetic positions) was plotted onto the physical map. Subsequently, chromosome segmentation/change point analysis was carried out using *R* package *changepoint* v1.0.6[44] as previously reported[45] for the Chinese Spring bread wheat chromosome 3B. Analysis features were: *trait*, recombination rate; *sliding window size*, 10 Mb; *step*, 1 Mb; thresholds to define high versus low-recombination segments, 0.40 cM/Mb and 0.05 cM/Mb, respectively.

1.3.2. MetaQTL analysis and projection of QTLs to DW assembly

Quantitative Trait Loci (QTLs) for major trait categories reported in tetraploid wheat, including domestication, fertility (grain-yield related traits, GY), heading date (HD), plant height (PH), biomass, grain yellow pigment content (GYPC), grain quality in general, disease resistance and root architecture-related traits, were considered for projection onto the DW assembly Supplementary Data Set 1. Dataset of a total 2,105 QTL signals was compiled based on published literature of QTLs (n=1,162) and genome wide association studies (n=943) in T. *turgidum* biparental and tetraploid wheat collections (as on June 30th, 2018). They were projected to DW genome assembly using a high-density reference binned Svevo × Zavitan linkage map (Supplementary Data Set 1) as genetic framework anchored to the physical assembly, enriched of additional DArT, SSR and EST-SSR/STS markers with matching genetic and physical locations (see marker projection results in section *2.1.10.*).

QTL projection was carried out with two different approaches. When several coincident QTLs for a trait (e.g. from different year/location combinations) were available from the same study, only the most consistent QTL was retained for the projection. On the other hand, when a cluster of QTLs was reported, being in the same genomic regions (overlapping intervals) but not coincident, they were considered for meta-analysis in order to refine the locus position and interval. These QTLs were calculated for each chromosome separately using the *Biomercator* version 4.2 software[46] which tested the most likely assumption between 1, 2, 3, 4 to *n* MetaQTLs (MQTLs). The Akaike Information Criterion (AIC) was considered to select the best MQTL model. The model with the lowest AIC value was considered as the best fit[46]. Moreover, for each QTL to project, a confidence interval (CI) was estimated[47] for QTLs identified in biparental populations and for association mapping studies we used

the reported LD extent for GWAS-QTLs (considering the average threshold of $r^2 \leq 0.3$). Flanking markers of the CI in the original map and present in the Svevo × Zavitan -Avni-binned map (Supplementary Data Set 1) were directly located on the Svevo genome assembly. This allowed to calculate the genetic distance ratio between the original and the reference maps. Based on this ratio, the original CI was projected onto the Svevo × Zavitan-Avni 2014-binned map anchored to the physical assembly. As a consequence, each CI was physically defined. When CIs from original maps were defined by markers not available in the reference map, the tetraploid wheat consensus map[43] was used as intermediate map. After locating the markers defining the original CI to the Svevo × Zavitan-Avni-binned map anchored to the physical assembly, the average bp of the Svevo × Zavitan-Avni-binned where the QTL peak was projected was considered as the most probable QTL position and used to locate QTLs in Fig. 1e and to obtain QTL distribution summary statistics.

Out of 2,105 recorded MQTLs, the summary included 47 for domestication, 775 for grain yield grain-yield related traits, GY, 200 for phenology (mainly heading date, HD), 104 for plant height (PH), 133 for biomass, 407 for disease response, 185 for grain quality in general, 233 for root architecture-related traits and 21 for other traits.

1.3.3. Genome-wide investigation of genetic diversity and linkage disequilibrium decay rate in the Global Tetraploid wheat Collection (GTC)

A survey of genetic diversity referred to the Svevo genome assembly was carried out in a Global Tetraploid wheat Collection (GTC) based on the Illumina iSelect 90K SNP genotyping platform[48]. The Tetraploid wheat germplasm includes up to 11 different taxa (Supplementary Table 11). Past human migrations and trade and modern industrial agriculture contributed to widespread the main taxa of *T. turgidum* ssp. *dicoccum* (domesticated emmer) and *T. turgidum* ssp. *durum* over wide areas, from the Fertile Crescent to North Africa, Europe, Transcaucasia, India and Ethiopia. In the last two centuries, durum wheat also expanded to Northern America (Canada, USA, Mexico), Southern America (Argentina, Chile) and Central Asia (Kazakhstan). This, coupled to the human-driven selection for traits related to domestication, plant architecture, adaptation and grain quality, resulted in a germplasm characterized by a wide range of biodiversity heritage, which is relevant for present and future targets of wheat (both tetraploid and hexaploid) improvement. Numerous studies targeted the genetic diversity present in tetraploid wheat, nevertheless, no previous study provided a complete diversity survey for all major tetraploid germplasm groups. A large panel of wheat accessions comprising different tetraploid taxa and germplasm pools were surveyed using high-density SNP arrays to explore genetic diversity in tetraploid germplasm. The survey was carried out by combining information from accessions/panels previously genotyped by the authors as well as by selecting

further sets of accessions to improve the representativeness of the collection. These accessions were purified by single seed descent (SSD) in greenhouse and then genotyped as detailed below.

Overall, we produced the raw genotyping data for a total of 2,558 tetraploid wheat accessions[43,48,49]. Raw Illumina iSelect genotyping data from previously genotyped durum wheat elite, landrace, emmer and wild emmer panels available from AgriBio, CREA, University of Bologna, University of Saskatchewan, and USDA-ARS were provided for joined analyses. To complete the panel representativeness, additional 490 tetraploid wheat accessions of world-wide origin and specifically selected from the main domestication and cultivation areas (Fertile Crescent, the Mediterranean Basin, Western Asia and Eastern Africa) were included. The additional accessions were chosen from the collection established by Dr. Benjamin Kilian and Dr. Hakan Ozkan and from the U.S. National Plant Germplasm System; single-plants were grown in greenhouse in 2016/2017 and genotyped with the wheat iSelect 90K SNP assay at the USDA-ARS Genotyping Lab, Fargo, North Dakota. In total, 90K SNP raw genotype data were obtained for 2,558 tetraploid wheat accessions.

The raw data (Theta/R) from single Illumina genotyping experiments were jointly analyzed for cluster assignment and genotype calling using a custom script for genotype calling in unrelated samples (https://github.com/plantinformatics/Durum_iSelect_90kSNP_GenotypeCalling), as described for the mapping population analysis. In brief, the script assigns samples to clusters previously identified from the genetic mapping analysis. A sample was assigned to a cluster if its probability to belong to that cluster (*vs.* the next closest cluster) exceeded 0.8. Samples assigned to each cluster were assigned a genotype call if the segregating allele tagged by the cluster could be unambiguously tracked; i.e. the allele it tracks was previously genetically mapped. Based on the complexity of the signal, cluster could be two (best situation corresponding to one single Mendelian locus) or multiples. Thus, the assigned genotype was an arbitrary allelic state, i.e. AA, BB or NC (not called). The cluster file underpinning the script used for genotype calling was based on 38,340 genetically mapped SNP loci mapped across 17 mapping populations.

The script parameters used for genotype calling were: -d 3, called sample were within 3 standard deviations of a known cluster position; -r 0.8, minimum confidence score that sample belongs to the cluster to which it was assigned versus the next closest cluster; a value of 1 indicates highest confidence. The genotype call pipeline allowed us to retrieve 34,543 SNP polymorphic on the complete dataset of 2,558 accessions. This complete dataset was subjected to two consecutive rounds of filtering for redundancy, initially based on passport information (accession name/international code) and then genetic similarity matrix (simple matching genetic similarity) among accessions based on SNP data. Accessions were filtered for genetic similarity ≥0.95, allowing for one representative

only for each highly similar accession group. After filtering for redundant genotypes, the final composition of the tetraploid wheat diversity panel consisted of 1,861 non-redundant accessions: three *T. aestivum,* and two *T. petropavlovskyi* Udacz. et Migush, included as hexaploid wheat (AABBDD genome) and 1,856 *T. turgidum* of 11 taxa (GTC, Supplementary Table 11 and Supplementary Data Set 2). The SNP genotype and passport data file of the GTC have been made available for download in Interomics Durum Wheat Genome (https://www.interomics.eu/durum-wheat-genome) and GrainGenes (https://wheat.pw.usda.gov/GG3/jbrowse_Durum_Svevo) databases. The seeds of the GTC collection are available upon request.

The tetraploid diversity panel of 1,856 accessions showed polymorphism for 34,543 SNPs. We selected a SNP dataset of 23,862 SNPs based on the following three criteria: *i)* only SNPs uniquely mapping as single Mendelian loci were retained when both their physical position (by BLAST) and genetic position (based on the reference mapping populations) were cross-checked, SNPs genetically mapping to multiple locations in diverse mapping populations were filtered out; *ii)* to limit the interference effects caused by ascertainment bias (particularly relevant for wild emmer and domesticated emmer accessions), the SNPs were further selected for overall null allele frequency ≤0.25 (failure rate); *iii)* singletons and double singletons were filtered out. After filtering for uniqueness based on LD (r^2 = 0.99) the dataset considered for further analysis reduced to 17,340 unique, non-redundant, single Mendelian SNP markers that were both genetically and physically mapped. This dataset was considered for all analysis related to diversity survey and detection of selection signals. For population structure and genetic relationship analysis, the dataset was further pruned to r^2 = 0.50 and for allele frequency <0.02 (MAF 0.02), yielding a core set of 5,787 SNP.

For the analysis of genetic diversity, the gene diversity[50] expressed as Nei's genetic diversity (D) was calculated based on the Nei's formula:

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

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

D was calculated for each of the four main germplasm pools. F_{st} differentiation index was calculated in *Arlequin* version 3.5.2[51] and in the *R* package *hierfstat* (Weir and Cockerham F_{st}) version 0.04-22[52].

Four main tetraploid germplasm groups were considered for the genome-wide analysis of genetic diversity: wild emmer wheat, WEW; domesticated emmer wheat, DEW; durum wheat landraces, DWL; durum wheat cultivars, DWC.

To obtain an initial picture of genetic diversity depletion distribution, the genetic diversity for the four germplasm groups was genome-wide scanned based on the filtered wheat iSelect 90K SNP array and averaged by non-overlapping windows of 10Mb steps (D_{10Mb}). Chromosome regions of

strong and extended genetic diversity divergences in at least one of the four groups (mainly genetic diversity depletions associated to domestication/breeder selection) were highlighted based upon the concomitant occurrence of the following conditions: *i)* diversity depletion $D_{10Mb} \leq 0.1$ in at least one of the four groups; *ii)* cross-population $\Delta D \geq 0.2$; *iii)* diversity-depleted region extending ≥20 Mb (at least two non-overlapping 10Mb windows). These regions were summarized in Supplementary Table 20.

Pairwise LD values were estimated by means of the *snpgdsLDMat* function in the *R* package *SNPRelate*[53], using an LD composite measure without sliding window. The LD estimates (allele frequency correlation, r^2) were calculated separately for each germplasm group (i.e. WEW, DEW, DWL, and DWC). For LD analysis we considered only SNPs with a minor allele frequency of at least 0.10 in the respective subsets. The intra-chromosomal smoothed r^2 values were plotted as a function of the physical distance between markers (in Mb), considering a maximum distance up to 6 Mb which corresponded to the distance at which the LD decay (r^2) reached the background baseline for DWC (the group with the slowest decay rate among the four considered). Smoothing was performed by calculating rolling mean[54] (R package *zoo*).

To highlight differences in local LD decay between proximal and distal chromosomal regions, we calculated the average LD values over the 100 markers nearest to a focal SNP and let the focal marker slide along the chromosomes (focal-LD). This procedure was performed separately for each germplasm group and the averaged LD values were smoothed by means of a rolling mean (window of 200 focal SNPs).

1.3.4. Genome-wide investigation of population genetic structure in a Global Tetraploid wheat Collection

We assessed the overall population genetic structure of 1,856 accessions representing the Global Tetraploid wheat Collection (GTC) using *i*) pairwise dissimilarity-based neighbor joining (NJ) phylogenetic tree, *ii*) standard principal component analysis (*PCA*), and *iii*) four alternative model- and non-model-based non-hierarchical, quantitative clustering analysis. For all analyses we used a non-redundant SNP dataset obtained after removing rare alleles with minor allele frequency <0.02 (MAF 0.02) and by pruning out SNPs with intra-chromosomal LD r^2 >0.5 to remove the bias caused by LD[55]. The non-redundant, LD-pruned, SNP dataset used to estimate population structure included 5,787 SNPs.

The *NJ* phylogenetic tree was obtained by calculating the pairwise genetic distances, performing 1,000 bootstrap resampling, and obtaining the tree in *R*, using the *dist.gene*, *boot.phylo*, *write.tree* and *write.nexus* functions (*poppr*, *pegas*, *ape*, *adegenet*, *ade4* libraries). *PCA* was performed using

EIGENSTRAT[56]. PC's 1, 2 and 3 are visualized in a 3D scatter plot with each PC explaining 29.65, 15.52 and 7.5 percent of variance, respectively.

As to the non-hierarchical, quantitative clustering analyses, two model-based likelihood methods *ADMIXTURE*[57], *fineSTRUCTURE*[58] and two non-model-based, geometric methods Discriminant Analysis of Principal Components, *DAPC*[59,60] and *sNMF*[61] were used. Clustering was explored for *K* groups ranging from 2 to 20. *ADMIXTURE* analysis was run based on 100 replications with different random seeds and with 10 cross-validations for *K* ranging from 2 to 20. The replicate with the highest log-likelihood for each *K* was considered. In *ADMIXTURE* analysis, cross validation (CV) values had a trend of continuous decrease with *K*, indicating the presence of a complex population structure (Supplementary Fig. 24). For *fineSTRUCTURE* analysis, we used the linked model with recommended settings by developers with the option "*–go"* to run the entire pipeline properly and to ensure that the Markov Chain Monte-Carlo has been run for enough number of iterations. *fineSTRUCTURE* does not accept missing data. SNP imputation was done using *FImpute* software with its default parameters[62]. To evaluate the imputation accuracy, we ran 1,000 replicates of randomly masked 1% of the called genotypes, imputed them with *FImpute* and calculated the concordance rate as the proportion of truly imputed genotypes[63,64]. An average imputation accuracy of 98.6% was observed across all replicates.

For *DAPC* analysis, *find.clusters* function from the R package *adegenet*[60] was first used to identify the optimum number of clusters (*K*) useful to describe the data. We run the *K-means* procedure (*kmeans* function) sequentially at increasing values of *K* from 2 to 20 and computing the BIC (Bayesian Information Criterion) statistics to measure the goodness of fit at each *K* with the following parameters: *n.pca* (n. of retained PCs) = 1500, *stat* = "BIC", *n.iter* (n. of iterations) = 1000. The BIC plot was retained for optimal cluster number evaluation. The *DAPC* function was implemented to describe the genetic diversity between these clusters using 50 principal components and 7 discriminant functions or synthetic variables saved (*n.pca* = 50, *n.da* =7).

The cross-validation procedure and the maximization 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) was used to choose the number of retained PCs and discrimination functions in the final analysis during the dimension-reduction step. Also, *DAPC* function provided the membership probabilities of each individuals to belong to different clusters. Further, *Ward clustering* was used as a valuable alternative to *K-means* analysis to explore the grouping of accessions based on the same number of principal components and discriminant functions analysis.

sNMF implements a non-Hardy-Weinberg model based population structure analysis based on non-negative matrix factorization (NMF) algorithms and compute least-squares estimates of ancestry coefficients[61]. As in the likelihood model implemented in *ADMIXTURE*, *sNMF* supposes that the genetic data originate from the admixture of *K* parental populations, where *K* is unknown, and it returns estimates of ancestry proportions for each multilocus genotype in the sample. Runs of *sNMF* were performed with K = 2-20 using parameters and procedures previously described[61]. Like *ADMIXTURE*, the cross-entropy values from *sNMF* method declined continually indicating presence of complex population structure. Nevertheless, both methods showed very similar clustering of accessions. The cross-comparison of clustering results generated from the four non-hierarchical quantitative clustering methods was carried out by inspecting the results and by calculating Pearson's *r* correlations and root mean square error (RMSE) values among the matrices obtained from each method. Based on this comparison, the *ADMIXTURE* output was retained for further investigation of genetic relationships among the taxa and populations within taxa.

After obtaining the global population structure representations, a more detailed clustering was carried out in *ADMIXTURE* separately for each of the four-main germplasm groups herein considered: WEW, DEW, DWL, DWC, respectively, based again on the investigation of *K* groups ranging from 2 to 20. *NJ* trees were obtained with the same procedure as above. For each of the four-main taxon groups, main *K* groups (populations) and further populations divisions were defined based on the decline of non-admixtured (Q >0.5) accessions' assignment ratio (fraction of accessions assigned to individual clusters with Q >0.5) at increasing *K* values, cross validation results and overall matching with the known accession's passport information, known genetic relationship patterns and dispersal routes. For each main taxon we defined the population structure based on two *K* values, a low *K* value that captured the main populations in the germplasm, and a higher *K* value capturing the subpopulations present within population.

Phylogenetic networks were computed using SplitsTree4 version 4.14.6[65] to better visualize complex evolutionary relationships between taxa (hybridization, horizontal gene transfer, recombination, or gene duplication and loss). Phylogenetic networks enable richer visualization than phylogenetic trees, which has been beneficial also for inferring crop domestication history[66,67]. More specifically, NeighborNet planar graphs of Hamming distances based on 90K-derived SNP data were computed for major germplasm groups and for the three major transition scenarios. Taxa are represented by nodes and evolutionary relationships by edges. The reticulated networks represent the same split by parallel lines/branches. Splits are branches with weights (lengths). As compared to the bifurcating phylogenetic trees, the reticulated trees appearing like boxes indicate competing patterns

of relationships while parallel lines indicate consistent splits and highlights predominant phylogenetic signals.

To gain more detailed insights into the genetic relationships among main taxa, we performed a hierarchical AMOVA and calculated *Fst* and *Nei's genetic distances* among and within main tetraploid taxa and populations. The main populations identified in the detailed ADMIXTURE analysis were considered as the base of this analysis, including 2 main WEW populations (North-eastern fertile crescent and Southern Levant), 6 main DEW and DWL populations (further subdivided into Northern and Southern populations) and 5 main DWC populations reflecting the most important currently cultivated breeding pools. Then, to minimize the potentially confounding effect of recent admixture on this detailed genetic analysis, accessions showing appreciable admixture were removed from this analysis[68] (only accessions with Q >0.5 for WEW and DEW and Q > 0.4 for DWL/DWC were retained, the list is reported in Supplementary Data Set 2.

We tested the hierarchical level of *Fst* differentiation, using the *R* package *hierfstat*[52] at three levels:

- level 1, among taxa, including WEW, DEW, DWL and DWC;
- level 2, among domestication origins within taxa, including North East fertile crescent (NE), southern Levant (SL), Ethiopia (ETH);
- level 3, among 19 populations within origins and taxa.

We then computed pairwise *Fst* values and *Nei's genetics distances* among populations and used the *boot.ppfst* function (1,000 bootstraps) to calculate their upper and lower confidence limits. We also calculated the expected heterozygosity within population as a reference.

1.3.5. Reduction of diversity, differentiation and selection signals associated to main domestication and improvement factors in tetraploid wheat

To assess the level and extent of genetic diversity reduction associated with emmer domestication and durum wheat evolution and breeding as compared to their most proximal ancestors, we calculated genome-wide genetic Diversity (*D*) and Diversity Reduction Indexes (*DRI*). Then, we assessed the presence of divergence and selection signatures and their co-occurrence with diversity reduction using four complementary selection metrics based on: *i)* F_{st}[69], *ii)* divergence of site frequency spectrum measured through cross-population composite likelihood score, *XP-CLR*[70] *iii)* haplotype-based metrics such as cross-population extended homozygosity, *XP-EHH*[71] and the haplotype-based FLK test, *hapFLK*[72].

The genome-wide scan was carried out using dataset of 17,340 non-redundant, genetically and physically mapped SNP as already reported (section *1.1.14.*). To reduce erraticism associated to

single-SNP based signals while considering for the average LD-decay rate, we compared three smoothing methods: *i*) averaged metrics on overlapping windows of fixed physical interval (25 kb, 50 kb, 1Mb, 2Mb, 10 Mb) with a smaller-size sliding step, *ii*) averaged metrics on overlapping windows of fixed genetic intervals (1-5 cM), *iii*) averaged metrics on overlapping windows of fixed SNP number (11, 15, 21, 25). The last method has two combined advantages, namely the constant number of markers in the sliding window and the capability to cope with the irregular marker density and recombination rate, a feature particularly useful for the pericentromeric regions. We therefore used a constant 25-SNP average sliding window with 1 SNP step for all metrics except *XP-CLR* and *XP-EHH*. The *D* was calculated for each of the four-main taxa groups as indicated in section *1.1.14.* using the average 25-SNP sliding window.

Accessions that were highly admixed (based on *ADMIXTURE* results), accessions of Ethiopian origin and those that grouped with the Ethiopian accessions were excluded from further analysis. After filtering, the lines were grouped into four broader groups based on their taxa or domestication status: WEW (n=104), DEW (n=248), DWL (n=591) and DWC (n=394). The cross-population metrics *DRI, Fst, XP-CLR, XP-EHH, hapFLK* associated to each of the three transitions were, then, calculated. For *D* and *DRI* 2.5 percentile of the top and bottom distribution were considered as outliers. Similarly, for *Fst, XP-CLR, XP-EHH* and *hapFLK* top five percentile of the distribution is considered as outliers. Adjacent outlier windows interrupted by one or few SNP in less than 10 Mb distance were merged to unique features. Based on the evidence of extended and strong signals detected in the centromeric regions, peri-centromeric regions were subsequently masked from the distributions and 2.5 and 5 percentile distributions were re-calculated[73]. To prioritize selection signals, we further calculated the 1 top-percentiles of the same distributions and projected on the selection signal map a dataset of 41 cloned loci relevant for domestication or selection and all the tetraploid QTL projected on the Svevo genome.

The cross-population *DRI* was calculated as: $DRI = (DI_{wild} + 0.1) / (DI_{derived} + 0.1)$. Constant value was added to the formula to consider for regions extremely low in diversity. A *DRI* of 2 means that the diversity in derived germplasm is half of that in the wild.

We computed F_{st} statistics at each locus (*R* package *pegas*[74]) as previously described[69]. Thresholds for both metrics were computed by running the *DRI* and F_{st} with 100,000 permutations carried out in R with a custom script (https://github.com/cnr-ibba/svevo-permutations).

XP-CLR test is based on modelling the likelihood of multi-locus allele frequency distribution between two populations. *XP-CLR* is shown to be less sensitive to ascertainment bias and have high power of detecting ancient selection signals[70]. For the analysis we used 0.5 cM sliding window with 50kb bp steps across the whole genome, smoothed over 1 Mb size intervals.

The *XP-EHH* haplotype-based metric was estimated using software *Selscan*[75]. The method is based on extended haplotype homozygosity, which measures the reduction in haplotype diversity in cross-population comparisons. *XP-EHH* values were normalised[71,75] and normalized values were transformed into absolute values. Normalized *XP-EHH* scores were averaged across 50kb windows and smoothed over 1 Mb size intervals.

We used *fastPHASE* v1.4.8[76] and *R* package *imputeq*[77] to reconstruct the haplotypes from SNP data and thus identify the optimum number of haplotype clusters. We run 5 tests using *imputeq* and imputing the genotypes with *fastPHASE* using the following parameters: fastPHASE -T10 -C25 -K{5:25} -H-1 -n -Z. Thus, we imputed the genotypes at ranges of clusters K from 5, 10, 15, 20 and 25. We estimated error using Estimate-Errors function in *imputeq*. The *K* that minimizes the error was selected as the optimum. *HapFLK* metric[72] was computed individually on each chromosome on 3 sets of data using the following parameters: *K* number of haplotypes = 10, and *nfit* = 50. *HapFLK* was run with a custom script (https://github.com/cnr-ibba/svevo-hapflk).

A set of 41 wheat cloned loci which are either targets for domestication or crop improvement were selected from the literature (Supplementary Table 12) and compared with evidences of putative selection signals. QTLs for domestication, phenology, disease resistance and quality were also considered for evidences of overlap with the selection signals. Multiple overlapping selection signals with peaks within 10 MB region from both within each transition (WEW-to-DEW, DEW-to-DWL, DWL-to-DWC) and across all transitions were considered as a unique selection signal cluster.

1.4 Identification of a locus controlling cadmium accumulation in durum wheat grain

1.4.1. Localization of Cdu-B1

Construction of the refined interval for *Cdu-B1* involved three independent mapping populations: Svevo × Zavitan, Kofa × W9262-260D3, and 8982-TL-L × 8982-TL-H. The first two populations were genotyped using the wheat iSelect 90K SNP array[48]. Markers from the array were mapped to the genome of Svevo by *GMAP*[11]. Markers uniquely mapping to chromosome 5B were used for Single Marker Regression analysis using Windows *QTL Cartographer* (https://brcwebportal.cos.ncsu.edu/qtlcart/index.php) and plotted according to their position and ability to discriminate between high and low Cd accumulators. To further refine *Cdu-B1*, markers determined to be closely linked to *Cdu-B1* in the Kofa x W9262-260D3 population[78] were used to screen the third population, a large F_2 mapping population from intercrossing of isogenic lines 8982-TL-L and 8982-TL-H derived from the cross Kyle × Nile[79]. We screened 5,081 F_2 plants from this population with *ScOpc20, Xusw47* (flanking markers) and *Xusw14* (*Cdu-B1* co-segregating marker)

and identified 20 F_2 plants that showed recombination. The seven remaining markers that co-segregated with *Cdu-B1* in the DH population were assayed in the recombinants and no recombination events were identified between the seven *usw* markers (*Xusw59 HMA3-B1*, *Xusw50*, *Xusw51*, *Xusw52*, *Xusw15b*, *Xusw17*, and *Xusw47*). The 20 recombinant F_2 plants from intercrossing of isogenic lines 8982-TL-L and 8982-TL-H were classified into one of seven haplotype groups (Supplementary Table 21). To further support the refined *Cdu-B1* interval, we performed high-throughput sequencing of exomes for each of the haplotypes. DNAs from each haplotype were combined and then exome sequenced[80]. Sequence reads were processed by *Trimmomatic* version 0.32[81] and processed reads were aligned to the genome of Svevo using *Novoalign* version 3.02.05 (www.novocraft.com/products/novoalign). Duplicate read mappings and improper read pairs were removed using *Picard-Tools* (https://broadinstitute.github.io/picard). Variants were called using the *SamTools* version 1.2.1[39] *mpileup* command. Filters were applied requiring each bulk to be homozygous and carry an allele that is segregating in the parental lines. For comparative sequence analysis, exome sequencing was also performed for the hexaploid wheat cultivar Sumai 3 using the same method as above, but with the Chinese Spring reference genome TGAC v1.0[82]. In addition, *Cdu-B1* was identified in the genomes of DW and WEW using the seven marker sequences from the mapping experiments. A detailed alignment of *Cdu-B1* between DW and other assemblies (WEW, Langdon bacterial artificial chromosomes, TGAC v1.0, and *Triticum* 3.1[1,10,82]), were performed by *MUMmer* 3.23 using the 1-to-1 alignment from the *dnadiff* command.

1.4.2. Functional characterization of TdHMA3

A global collection of durum wheat cultivars and breeding lines available at University of Saskatchewan was used to validate the association of allelic variation in *TdHMA3-B1* with phenotypic expression of Cd accumulation in grain. Phenotypic data for Cd accumulation in grain were collected from field trials conducted previously at Saskatoon, Saskatchewan, over two years. Cd concentrations in grain were determined using procedures described previously[80].

To functionally characterize *TdHMA3*, full-length DW *HMA3* cDNAs (*TdHMA3-A1* and *TdHMA3-B1* homoeologs) were cloned and sequenced. Total RNA was extracted using the RNeasy Midi Kit (Qiagen, Hilden, Germany) from roots of 21-d-old plants of 8982-TL-L (low Cd) and 8982-TL-H (high Cd) isogenic lines grown in hydroponic culture, as described previously[83]. cDNA was then synthesized from DNase-treated RNA using oligo(dT) and Superscript III (Invitrogen, Carlsbad, CA). Full-length CDS of *TdHMA3-A1* and *TdHMA3-B1* were amplified using flanking primers (*TtHMA3-F3* and *TtHMA3-R3*, Supplementary Table 22) and cloned into pJET1.2/blunt

(Thermofisher Scientific, Waltham, MA) for sequencing (Genbank accessions KF683290-KF683295).

For comparative sequence analysis, protein sequences or translated CDS sequences for P_{1B}-ATPases (HMAs) from *Arabidopsis*, *Brachypodium distachyon*, and rice were compiled. DW *HMA* genes were identified by TBLASTN of the Svevo genome using *Brachypodium* and rice HMA proteins as queries (E-value < 10^{-3}), and the DW *HMA* gene models were predicted with *Fgenesh+*[84] using relevant wheat or barley HMAs as homologs. The sequences and locus identifiers of the proteins included in the phylogenetic analysis are shown in Supplementary Data Set 12. Sequences were aligned with *MAFFT L-INS-i* (version 7.311, https://mafft.cbrc.jp/alignment/server) using the default settings[85]. Gaps and poorly aligned regions were removed from the multiple sequence alignment (MSA) by *Gblocks* version 0.91b[86] using less stringent selection criteria[87]; http://molevol.cmima.csic.es/castresana/Gblocks_server.html). The trimmed MSA consisted of 570 positions (31% of the untrimmed MSA), including 58 invariant sites. A phylogenetic tree was reconstructed using the maximum-likelihood method with *PhyML* version 3.1[88] using the best fit model (LG+I+G+F; LG amino acid substitution matrix[89], the proportion of invariant sites estimated from the data, 4 gamma-distributed substitution rate categories, and empirically determined amino-acid frequencies) as determined by *SMS*[90] (http://www.atgc-montpellier.fr/phyml-sms/). The reliability of internal branches was tested by bootstrap analysis with 1000 replicates. Bootstrapped trees were summarized as a majority-rule consensus tree with *Phyutility*[91]. The phylogenetic tree was displayed and annotated using *FigTree* version 1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/).

Yeast (*Saccharomyces cerevisiae*) strains used for heterologous expression of *TdHMA3-A1*, *TdHMA3-B1a*, and *TdHMA3-B1b* included the BY4741 parental strain (Euroscarf Y00000: *MATa*; *his3Δ1*; *leu2Δ0*; *met15Δ0*; *ura3Δ0*), the Cd-sensitive *ycf1* mutant (Euroscarf Y04069: *MATa*; *his3Δ1*; *leu2Δ0*; *met15Δ0*; *ura3Δ0*; *ycf1::kanMX4*), and the Zn-sensitive *zrc1cot1* mutant (MATa, *zrc1::natMX3*, *cot1::kanMX4*, *his3Δ1*, *leu2Δ0*, *met15Δ0*, *ura3Δ0*; kindly provided by Ute Krämer, Ruhr University Bochum, Germany)[92]. Yeast codon-optimized open reading frames (ORF) of *TdHMA3-A1* (KF683291:87-2537), *TdHMA3-B1a* (KF683294:85-2574), and *TdHMA3-B1b* (KF683295:85-267) were synthesized by IDT (Coralville, IA). The longest alternative (5'-truncated) *TdHMA3-B1b* reading frame, ORF2 (KF683295:534-2591, 2058 bp), was PCR-amplified from codon-optimized *TdHMA3-B1a* (native *TdHMA3-B1a* and *TdHMA3-B1b* are identical in this region) using primers yTtHMA3-ORF2-BamHI and Linker-EcoRI (Supplementary Table 22). The *YCF1* ORF (YDR135C) was PCR-amplified from BY4741 genomic DNA using primers YCF1-BamHI-S and YCF1-XhoI-AS. The *ZRC1* ORF (YMR243C) was PCR-amplified from BY4741 genomic DNA using primers ZRC1-SpeI-F1 and ZRC1-EcoRI-R1. Transport-activity deficient *TdHMA3-A1* and

TdHMA3-B1a constructs were created by overlap extension site-directed mutagenesis[93] of the conserved P-ATPase phosphorylation site, Asp-411 (mutated to Ala: D411A), which is necessary for P-ATPase transport activity[94]; mutagenesis primers: Supplementary Table 22. The constructs were ligated into *Bam*HI (*Spe*I for *ZRC1*) and *Eco*RI (*Xho*I for *YCF1*) restriction sites of a single-copy (centromeric) yeast expression vector for TEF (p413TEF::*HIS3*) promoter-mediated expression[95]. All constructs were verified by sequencing and transformed into yeast strains using the lithium acetate procedure[96]. Transformants were selected on synthetic complete plates lacking histidine [SC-His: 2% (w/v) agarose; 2% (w/v) glucose; 0.67% (w/v) yeast nitrogen base w/o amino acids (Difco 291940); amino acid concentrations according to[97], but with elevated concentrations for the other auxotrophic amino acids (500 mg L^{-1} Leu, 100 mg L^{-1} Met, 150 mg L^{-1} Ura)[98]] after 3 d incubation at 30°C.

Complementation of *ycf1* and *zrc1cot1* by *TdHMA3* was determined using growth assays conducted in 96-well microtiter plates (modified from[99]). Yeast cultures were grown overnight from single colony inoculations (SC-His liquid medium) to mid log-phase at 30°C with continuous shaking. The overnight cultures were washed in sterile deionized water and resuspended in SC-His to 1.25 OD_{600}. SC-His media (115 µL) containing $CdCl_2$ or $ZnSO_4$ at the concentrations indicated was aliquoted to wells of U-bottom 96-well culture plates (Greiner Bio-One 650-185). Individual wells were inoculated in triplicate with 10 µL of SC-His (blank wells), or with 10 µL of 1.25 OD_{600} yeast culture to achieve an initial OD_{600} (10 mm path length) of 0.1. The microtiter plate lid was sealed with Parafilm and the plate was incubated for 48 h in an Eon microplate spectrophotometer (Biotek, Winooski, VT) at 30°C with variable shaking (alternating between orbital (180 s, 559 cpm, 1 mm) and linear (45 s, 1096 cpm, 1 mm) shaking). OD_{600} measurements were taken at 10 minute intervals. Growth curves for each experiment were calculated as the average of the triplicate wells following subtraction of the average OD_{600} for the blank wells. Multiple independent experiments were used to assess reproducibility of the growth assays. Although not all combinations of genotypes and treatments were included in each growth assay experiment, all experiments included negative controls (empty vector) and positive controls (*ycf1* and *zrc1cot1* expressing p413TEF-YCF1 and p413TEF-ZRC1, respectively) to permit comparisons across experiments. Plotted growth curves show the mean response ± 95% confidence intervals from at least 3 experiments for each genotype. Growth of *ycf1* and *zrc1cot1* expressing the positive controls (p413TEF-YCF1 and p413TEF-ZRC1, respectively) was equivalent to the parental strain, BY4741, expressing p413TEF in the presence of added Cd or Zn (BY4741 data not shown).

Cd and Zn accumulation in yeast was determined after 4 h exposure to 5 µM Cd or 50 µM Zn. Yeast cultures were grown overnight in SC-His (30°C with continuous shaking) to mid log-phase, washed in sterile deionized water, and resuspended in SC-His to 0.3 OD_{600}. After incubation for 1 h,

5 µM $CdCl_2$ or 50 µM $ZnSO_4$ were added to the media and the cells were cultured for 4 h. The cells were collected on 0.2 µm nylon filter spin columns (Norgen Biotek, Thorold, ON), and washed twice with cold (4°C) 50 µM EDTA (pH 5.0) and twice with cold (4°C) deionized water. The cells were digested overnight at room temperature by addition of 100 µL HNO_3 to the column. The digests were eluted from the columns by centrifugation, and the columns were washed twice with 200 µL deionized water. The Cd and Zn concentrations of the pooled eluate were determined by atomic absorption spectroscopy, as previously described[83]. The Cd and Zn accumulation experiments were repeated with similar results.

Subcellular localization of TdHMA3-A1 and TdHMA3-B1 proteins in yeast was determined using in-frame C-terminal fusions with yeast codon-optimized GFP (yEGFP) derived from pKT128[100]; Euroscarf P30174). The *TdHMA3-GFP* fusions, with flanking 5'-*Bam*HI-AAAA and 3'-*Eco*RI restriction sites, were generated by overlap extension[93] by utilizing the overlapping linker sequence 3' of the *TdHMA3-A1* and *TdHMA3-B1* stop codons (mutated to TTA (Leu) during overlap extension) and 5' of the *yEGFP* start codon (Supplementary Table 22). The *TdHMA3-GFP* constructs were ligated into the *Bam*HI and *Eco*RI restriction sites of p413TEF, confirmed by sequencing, and transformed into *ycf1*. Yeast cultures were grown in SC-His in microtiter plates for 16 h prior to GFP imaging. The GFP fluorescence of yeast was observed with an Axio Imager.M1/LSM 510 META confocal microscope (Carl Zeiss Microscopy GmbH, Oberkochen, Germany) with GFP excitation, 488 nm, and detection, 505–530 nm. GFP images were cropped and subject to linear threshold adjustment using ZEN 2012 (Carl Zeiss Microscopy GmbH, Oberkochen, Germany).

The phenotypic effect of the non-functional allele, *TdHMA3-B1b*, on Cd accumulation in roots, shoots, and grain was determined by comparing Cd accumulation during grain filling of near-isogenic lines (NILs) 8982-TL-L (low Cd) and 8982-TL-H (high Cd), which are homozygous for alleles *TdHMA3-B1a* and *TdHMA3-B1b* respectively. The high Cd NIL consistently accumulates 2 to 4-fold greater Cd concentrations in mature grain than the low Cd NIL[101]. The NILs were grown in chelator-buffered nutrient solution until physiological grain maturity. The chelator, HEDTA, was added to the solution at a 25 µM excess over the total concentration of transition metal cations (including 0.5 µM Cd), thereby buffering free metal activities at levels similar to those found in uncontaminated agricultural soils[102]. Whole-plants were harvested at 0, 7, 14, 28, and 42 days post-anthesis (DPA) and separated into roots, shoots, and grain prior to Cd determination. The complete protocol used for the nutrient solution experiment was described previously[83].

Three DW cultivars, the low and high Cd NILs (8992-TL-L/H) and AC Avonlea (high Cd), were grown in cooperation with Western Cooperative Fertilizers (Calgary, AB) at a field site in Alberta, Canada (+51° 2' 46.43" N, -112° 45' 1.06" W) in 2003. The soil at the site was an Orthic Dark Brown

Chernozem (pH 7.8) with 0.185 mg kg^{-1} extractable Cd (0.005 M DTPA-extractable). Each cultivar was grown in four replicates in a randomized block design. Blocks consisted of 6 rows of plants in plots of 7.6 m length. Plots were seeded, fertilized, and managed according to regional, dryland durum wheat cropping practices. Approximately 100 culms from the center four rows of each plot were labeled at anthesis, and five labeled culms per plot were randomly selected and harvested at 0, 7, 14, 21, and 28 DPA. Grain was manually separated from the pooled culms, and the leaf and stem tissues were washed successively under running tap water and deionized water. The plots were combine-harvested at maturity (43 DPA) and the cleaned grain was subsampled for analysis. Grain and shoot tissues were oven-dried at 65°C for 3 d and analysed for Cd accumulation as previously described[83]. The elemental composition of the mature grain (43 DPA) was determined by a commercial laboratory (ALS Global, Edmonton, AB). Accumulation of Cd in shoots and grain of AC Avonlea was similar to that observed in the high Cd NIL (AC Avonlea data not shown).

1.4.3. Chromium sequencing of Svevo

Whole genome sequencing was performed for Svevo using the Chromium 10x Genomics platform. Nuclei were isolated from 30 seedlings[103], and high molecular weight genomic DNA was extracted from nuclei using CTAB. Genomic DNA was quantified by fluorometry using Qubit 2.0 Broad Range (Thermofisher) and size selection was performed to remove fragments <40 kb using pulsed field electrophoresis on a Blue Pippin (Sage Science) according to the manufacturer's specifications. Final DNA integrity and size were determined using a Tapestation 2200 (Agilent), and Qubit 2.0 Broad Range (Thermofisher), respectively. Library preparation was performed as per the 10x Genome Library protocol (10x Genomics). Uniquely barcoded libraries were prepared and multi-plexed on Illumina HiSeq. De-multiplexing was performed by *Supernova* (10x Genomics) and fastq files were generated using *LongRanger* (10x Genomics). Reads were aligned to the Svevo reference sequence using *LongRanger WGS* (10x Genomics) and structural variants were visualized using *Loupe* software (10x Genomics).

2 Supplementary Text

2.1. Gene content, genome annotation and pattern of gene expression

2.1.1. Validation of the DW genome assembly and annotation

To validate the Hi-C based pseudomolecules of *T. durum* cv. Svevo we aligned the Svevo assembly to an independently constructed grass genome assembly of high quality, namely the genome sequence of *Brachypodium distachyon*[4]. The whole genome alignment between *T. durum* and *B. distachyon* is shown in Supplementary Fig. 25. As expected from previous studies in the Triticeae that employed the GenomeZipper approach[104], we observed large syntenic blocks that are shared between both *T. durum* subgenomes and *B. distachyon* (see for comparison Figure 2 of extended data in Mascher et al.[105]). For example, the non-recombining region of group 1 chromosomes shows high collinearity to a proximal region of *B. distachyon* chromosome 3.

The highest proportion of BUSCO genes were found in the set of all DW genes (n = 1,413, 98.1%), similar values were found for WEW[1] (n = 1,432, 99.5 %). These high values indicate that both assemblies represent an almost complete fraction of the gene space. Furthermore, 96.1% (DW) and 99.2% (WEW) of the BUSCOs were fully represented by the HC gene sets (Supplementary Fig. 26).

In addition to the BUSCO analysis, the predicted proteins in durum and WEW were validated using 216 experimentally-determined genes (Supplementary Data Set 8). As much as 97.7% (n = 211) of these genes were represented by at least one annotated gene with at least 75% protein coverage and Evalue < 10^{-05}, in both DW and WEW. Most of them were represented in the HC gene sets (95.4%).

When the set of 204,773 unique reference protein sequences was used in a BLAST search against the HC gene set, 194,131 proteins had a significant hit to DW and 194,523 to WEW gene sets (e value < 10^{-05}). From these genes, 92.3% (DW) and 92.4% (WEW) were represented by an annotated gene with at least 75% query coverage (Supplementary Fig. 27). These results indicate that both 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.

The presented DW assembly (like the WEW) captures, in contrast to previous wheat contig assemblies[106], almost all of the maximally expected k-mer defined repetitive space (target line) which was derived from randomly collected Illumina reads (Supplementary Fig. 19a). At the same time the assembly resolves the correct structure and expected amount of still intact full length LTR-retrotransposons (fl-LTRs). These 8 to 20 kb long elements have been notoriously difficult to

reconstruct by conventional read assemblers because of their (nearly) identical 1-2 kb long terminal repeats, which had a strong tendency to collapse during the assembly process (Supplementary Fig. 19b). The number of retrievable fl-LTRs in high quality assemblies shows a direct linear relationship to the genome size and can thus serve as a valuable additional metric of the ability of different assemblers to correctly resolve difficult, repetitive structures.

Based on the expression analysis, 96.18% of genes were expressed in at least one samples and 3.81% of genes were not expressed at all (log-normalised read counts >4). We found that the mean expression density was higher for genes in the centromere-proximal regions of the chromosomes (Fig. 1i).

2.1.2. Repeat annotation

The detection of full length LTR-retrotransposons resulted in a final set of 51,077 high quality full-length LTR retrotransposons for DW, which is very similar to the 53,295 fl-LTRs previously identified in WEW[1] with the same approach. About 8,000 fl-LTRs are still found in syntenic positions between DW and WEW (diagonal in Supplementary Fig. 2), because the estimated divergence time of ~ 10,000 years was too short to lead to the usually almost complete turnover of the intergenic space observed between different *Triticeae* species[107]. Dotplots of fl-LTRs subfamilies within DW for example have no diagonal between the A and B subgenomes. Contrary to genes, all formerly syntenic transposons from the common ancestor of A and B have been removed or reshuffled in the ~ 6 millon years of separate A and B evolution.

2.1.3. Identification of pseudogenes

The pseudogene distribution on the chromosomes is similar to the distribution of genes and mirrors the distribution of TEs. The B subgenome of DW harbors 6% more full-length pseudogenes than the A subgenome, even though it is only 2.8% larger (binomial test; p-value = 7.5e-07). One quarter of the pseudogenes (DW 24.4%; WEW 24.1%) could be unambiguously assigned to the duplicated subtype because they have retained the exon/intron-structure of their parental genes. In contrast, only 2.7% (WEW) and 2.6% (DW) are retroposed pseudogenes originating from an intermediate mRNA. This low amount of retroposed pseudogenes in spite of the high retrotransposon content is consistent with findings in barley, rice and Arabidopsis[108,109]. Compared to the proportion of duplicated pseudogenes in barley (14%), the higher pseudogene count in DW and WEW (24%) is likely to be a consequence of the tetraploid genome and a fingerprint of an ongoing gene loss process facilitated by the functional complementation through homoeologous gene copies (Supplementary Table 18).

2.1.4. MicroRNA sequencing and annotation

The *in silico* analysis with the *Shortstack* pipeline identified 774 unique loci displayed on the Genome Browser. The identified miRNA loci were grouped into 67 families equally distributed on the 14 chromosomes. Out of the 774 miRNA loci, 363 belong to known plant miRNA families showing sequence homology with previously annotated plant miRNA precursors and/or with deposited miRNA mature sequences; 75 are ambiguously annotated, being similar to a minor extent to some MIR families (annotated as unknown) and 336 are putatively durum wheat specific newly identified miRNAs. More than half (56%) of the predicted miRNAs have been retrieved in only one of the nine sequenced libraries, suggesting a specific expression, while 27 miRNAs have been retrieved in every sequenced library, supporting a more general expression (Supplementary Data Set 13).

2.1.5. Long non-coding RNAs

Annotation of transcriptome revealed a total of 115,437 lncRNAs identified according to[21]. All lncRNA are displayed on the Genome Browser. The length of lncRNAs ranged from 200 to 4,407 nucleotides with the average length of 279 nt where the standard deviation was 101.83 and the median was 245 (Supplementary Fig. 28). The distribution of GC content varied from 13.46% and 80.09% with the average content of 45.6%. Among all chromosomes, chromosome 3B contained the highest number of lncRNAs whereas chromosome 6A has the lowest one. Comparison of the A and B genomes revealed that both genomes contained a similar lncRNA number where A genome had 105,885 and B genome had 106,578 lncRNAs. Of these lncRNAs, 8,859 were specific to A genome whereas 9,552 were specific to B genome. Furthermore, 39.57% of the total lncRNAs were common to all chromosomes in both A and B subgenomes. Of these common lncRNAs, 65,471 were associated with transposable elements. Most of these repeats showed similarity with Type I-CACTA and Type II-Gypsy family repeats (Supplementary Fig. 29).

2.1.6. Annotation of prolamin seed storage genes

Prolamins are the main seed storage proteins of wheat and their characteristics have a strong influence on the technological properties of wheat flour. Wheat prolamins are classified as high-molecular weight (HMW) glutenin subunits, low-molecular weight (LMW) glutenin subunits, and α, γ, ω and δ-gliadins. In addition, a new family of avenin-like proteins has been more recently identified[110]. Prolamine-encoding genes belong to multigene families, in which individual members are organized in tandem and interspersed by repetitive elements. Consequently, accurate assembly of these regions can be difficult to obtain starting from short sequence reads. Notwithstanding this, 102

out of 124 sequences identified in the Svevo genome were correctly mapped to four expected chromosome regions[111-113]: the short and long arms of chromosome 1A/1B (*Glu-3, Gli-1* and *Glu-1* loci), the short arm of chromosome 6A/6B (*Gli-2* loci), the long arms of chromosome 4A and the short arm of chromosome 7A (*avenin-like* loci) (Supplementary Table 17). In agreement with previous studies, our annotation confirmed that the *Glu-3* and *Gli-1* loci are tightly linked[114]. Comparison with the Zavitan genome, which has been assembled with the same experimental approach, revealed that the total number of prolamin genes is slightly higher in Svevo (Supplementary Table 17).Moreover, our assembly clearly captured genomic features that are conserved between gliadin loci. For instance, in agreement with recent studies, we found that, like in bread wheat and *Aegilops tauschii*, all α-gliadin sequences (*Gli-2* loci) were located between glutamate-like receptor genes that were duplicated from an ancestral copy[115,116].

2.1.7. NLR gene family organization in durum and wild emmer wheat

NLRs 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[117-119]. When *NLR-annotator* was used to search for NLR motifs, 2,442 and 2,420 NLR loci were predicted for DW and WEW genomes, respectively. Since this annotation pipeline is different from the one used for the prediction of HC and LC genes, the results of *NLR-annotator* also included 390 and 417 loci in DW and WEW, respectively, not reported in the corresponding gene models. In DW, out of the predicted 2,442 NLR loci 1,487 were complete, 814 were pseudogenes/partial genes and 141 were complete genes on chrUn. In WEW, out of the predicted 2,420 NLR loci 1,462 were complete gene models, 857 were pseudogenes/partial genes and 101 were complete genes located on chrUn. 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 (Supplementary Fig. 30a). Interestingly, 172 DW-specific genes and 136 WEW-specific genes were identified. These NLRs genes were mainly localized on subgenomes B of the chromosomes (Supplementary Fig. 30b).

Three main NLR gene clusters were identified based on multiple-alignment analysis (the heatmap similarity matrix is reported in Supplementary Fig. 31a). These main clusters were differentiated at the level of domain composition[120], position of domains inside the putative genes as well as the amino acidic difference at the domain level (Supplementary Fig. 31b and Supplementary Fig. 32). While in clusters 1 and 3 the ratio of DW and WEW NLR loci was similar, cluster 2 was enriched in DW genes (8% more compared to WEW).

Synteny dot plots of the complete set of DW and WEW NLR genes (minScore = 800), computed with R package *DECIPHER*[121], showed that the NLR genes are more conserved as collinearity and number of copies between Svevo and Zavitan in genome A compared to genome B.

2.1.8. Annotation of plant functional non-tandem duplicated gene cluster

Plant functional non-tandem duplicated gene clusters (FNTDCs) are groups of co-localized, non-tandem duplicated genes sharing the same biological function. FNTDCs are thought to be evolved and maintained to facilitate coinheritance of embedded genes and co-expression, and to avoid accumulation of toxic intermediates[122-124]. To date, some 30 FNTDCs have been identified, mainly referring to secondary metabolism (e.g. terpene biosynthesis). Here we provide a rich dataset based on a genome-wide, *ab initio* heuristic approach for identifying candidate FNTDCs in durum wheat.

As shown in Supplementary Fig. 33, tandem duplicated genes strongly interfered with the detection of genuine FNTDCs, and, at a GO enrichment call threshold *p*- value of 10^{-6}, between 76 and 361 candidate FNTDCs were called depending on stringency of homology detection criteria (respectively, from 40% to 98% minimum identity and 0.4 to 0.98 alignment overlap). A genome-wide search for plant functional non-tandem duplicated gene cluster is available in figshare (https://figshare.com/; accession: 10.6084/m9.figshare.7038389).

At an intermediate representative setting (namely 70% identity, 0.7 alignment overlap and *p*-value 10^{-6}; hereafter referred to as average setting) 197 candidate FNTDCs were called with an average cluster size of 900 Kb. In a few cases, very large clusters longer than 24 GO-BP-equipped genes, the maximum window length tested, produced contiguous FNTDC. Candidate FNTDC coordinates for the average settings have been integrated into the Genome Browser.

Using the average setting parameters, 84 FNTDCs were found to have a homoeologous counterpart. This is likely to be an underestimate; nonetheless, some diversity between the two genomes with respect to FNTDC is very likely because some *bona fide* FNTDC (low *p*-values; $< 10^{-6}$ and 40% identity) lack a homoeologous counterpart (e.g. base excision repair, golgi organization, fucosylation, histone H3 acetylation and many organellar–targeted FNTDCs).

Using as reference the average setting, six secondary metabolism terpene-related candidate FNTDCs could be identified, including FNTDCs related to mono-, di- and tetra- terpenoids. Twenty-one candidate FNTDCs referred to organelle-targeted non-secondary metabolic processes as multi-enzyme complex for proton and respiratory electron transport chain and were called even at most stringent settings. This suggests that such a class of candidate FNTDCs may originate from migration of endosymbiont gene chunks towards nuclei[125]. Further, previously unreported candidate FNTDCs

include an ethylene response, fucosylation, base-excision repair, Golgi organization, amino acid transmembrane transport and several other primary and secondary metabolic processes.

At average setting, more than one thousand of yet functionally unknown (no GO BP process assigned) genes are entangled or adjacent to called FNTDCs (Supplementary Fig. 34), and, as previously found for some FNTDCs[126] (e.g. DIMBO maize FNTDC), entanglement in a FNTDC may provide hints for functional annotation of yet unannotated genes. A rich dataset is provided (https://figshare.com/; accession: 10.6084/m9.figshare.7038389) including, in addition to a searchable summary spreadsheet file and explanatory notes on files, raw files with detailed gene annotation, GO hypergeometric test results, list of homology-excluded genes, multiple alignments and alignment trees of GO BP-equipped genes for each called FNTDC at intermediate settings.

2.1.9. Genome browser

While the genomic sequences and annotation are accessible on scientific open access repositories, researchers who intend do further analysis can also explore them using the web-based Durum Wheat Genome Browser accessible after registration at the following address: http://www.interomics.eu/durum-wheat-genome. The Durum Wheat Genome Browser, which is an instance of the GBrowse genome viewer[127,128], can be used to display the following tracks:

- *Gene span* shows the total extent of the transcribed region of the annotated genes, with direction of transcription indicated. For each gene a detailed page reports a summary of the annotation results and allows analysis of the annotated mRNAs and the related CDSs. The sequence of each mRNA is also reported and the related CDS sequences are highlighted. Cross-reference data in the annotation (if any) are automatically converted into hyperlinks.
- *Protein-coding genes* shows the genes with exons and introns- Confidence class and functional annotation are also automatically shown.
- *Protein-coding genes (F/R)* is similar to the *Protein-coding genes* track. It uses two sub-tracks to represent genes on different strands. The genome browser provides visualization of one or both sub-tracks.
- *Protein-coding genes (HC/LC)* is similar to the *Protein-coding genes* track. It uses two sub-tracks to represent genes with different confidence classification. The genome browser provides visualization of one or both sub-tracks.
- *CDS* shows the extent of sequence encoding each specific polypeptide: annotation detail pages are available for each feature.
- *Frame usage* uses a *musical staff*, this representation shows the reading frame of each CDS and

presents how alternative splicing changes the reading frame.

- *3-frame translation forward*, if zoomed out (>400 bp), shows ticks at sites of stop codons for the three frames on the forward strand, whereas if zoomed in (≤ 400 bp), it shows predicted translations for each frame, using single-letter amino acid code.
- *3-frame translation revers,* if zoomed out (>400 bp), shows ticks at sites of stop codons for the three frames on the reverse strand, whereas if zoomed in (≤ 400 bp), it shows predicted translations for each frame, using single-letter amino acid code.
- *DNA/GC Content* if zoomed out (>100 bp) plots the GC content calculated over 10bp windows, whereas if zoomed in (≤100 bp) shows the double stranded DNA sequence.
- *CpG islands* – identifiers of the features are automatically displayed in the track. The p-values and annotation detail pages are displayed.
- *QTL – it* shows the QTL annotations: a details page is available for each annotated QTL.
- *Promoter Elements* – for each feature the related promoter element is displayed in the track. Information such as the p-value, q-value and sequence are shown, whereas the related web-page of the Eukaryotic Promoter Database is linked.
- *Transcription Factor Binding Sites* (*TFBS)* identifiers of the feature are reported on the track as well as a summary of the *p-value*, *q-value* and the related sequence. For each feature, the related web-page of Jaspar CORE 2016 database is linked.
- *RNA-Seq alignments* and *RNA-Seq coverage* have been added to explore both alignments and coverage. By clicking the alignment a details page with a *sam* description of the alignment is displayed, whereas by clicking on the coverage glyph a details page with some statistical information is opened.
- *lncRNA* and *miRNA* annotations. For both tracks, annotation identifiers and annotation details page are displayed.
- *NLR gene* track shows the NLR gene family organization predicted with *NLR-annotator*.
- *Clusters* track represents the functional non-tandem duplicated gene clusters (*FNTDC)* annotations. By clicking on a feature, a details page is displayed.
- *Markers* track allows to explore all markers. For each marker the type and features are displayed.

The Durum Wheat Genome Browser has also been configured to use the 'Landmark or Region' box to perform searches based on feature identifiers and functional annotation. It is possible to enter a complete description of a functional annotation as well as some keywords. For each track the related annotation data and FASTA sequence can be downloaded. The Durum Wheat Genome Browser allows to export images of the genomic region visualized in png, PDF or SVG format.

2.1.10. Genetic maps, marker projection on the durum wheat genome and genome-wide investigation of recombination rate

The 17 tetraploid wheat mapping populations considered in this study provided genetic map locations for over 939,536 GBS tags, 38,340 Illumina iSelect SNP, 1,341 DArT®, 835 SSR, 109 EST-SSR and EST/STS markers (Supplementary Table 19). An inventory of the publicly available sequences was carried out for the following marker sets: iSelect wheat 90K Infinium SNP[48], 820K wheat Axiom® Roche SNP[129] (for Axiom® SNP, refer to http://www.cerealsdb.uk.net/), TaBW280K[130], 35K wheat Axiom® Roche SNP[131], DArT® markers, SSR markers, EST markers (for DArT, SSR and EST-SSR, refer to GrainGenes database, https://wheat.pw.usda.gov/GG3/). The corresponding FASTA files were used in a BLAST search against the Svevo and Zavitan genome assemblies with stringent BLASTN parameters. Hits were retrieved for the majority of markers as reported in Supplementary Table 6A.

All BLASTN and genetic mapping results from the marker' data sets were projected onto the binned Svevo × Zavitan iSelect 90K SNP reference map[41], whose SNP markers were, in turn, anchored to the Svevo physical assembly. Based on these projections, all markers for which BLASTN hits and/or genetic mapping information were available could be successfully inserted onto the unique physically- and genetically-defined framework provided by the Svevo × Zavitan binned map (Supplementary Fig. 1). Based on these projection data, whole-genome average recombination rate was 0.212 cM/Mb (very similar to the value of 0.250 cM/Mb previously observed for the 3B Chinese Spring chromosome[45]).

Highly-recombinogenic, distal chromosome regions showing a linear relationship between genetic and physical distances were clearly identified and separated from the interstitial and peri-centromeric regions for all chromosomes (Supplementary Fig. 1). Highly-recombinogenic distal regions accounted for 2,207.1 Mb (1,481.3 cM in total), corresponding to 22.1% of total genome size (70.2% of the Svevo × Zavitan genetic map length[41]). These regions contain 27,109 HC genes (42.4% of all HC genes). The distal chromosome regions are thus characterized by an average recombination rate of 0.557 cM/Mb (=1.79 Mb/cM) and a HC gene density of 13.0 genes/Mb (18.1 HC genes per cM). About 60% of all MQTLs (1,272 over 2,105) were located in highly recombinogenic distal regions, with a MQTL density of 6.2 MQTLs/10Mb.

Recombination-depleted, pericentromeric proximal regions were also clearly detected (based on a segmentation threshold set at 0.05 cM/Mb), covering 4,430 Mb and 47.9 cM in total. These corresponded to a 44.4% of genome size or 2.3% of the genetic map. The HC gene content was equal to 14,805 (23.1% of all HC genes). In pericentromeric regions the average recombination rate was 0.011 cM/Mb, equivalent to 107.1 Mb/cM, with a gene density of 3.38 HC genes/Mb (354.36 HC

genes/cM). These regions carry 177 MQTLs (8.4% of all MQTLs), which reflects the well-known overall diversity-depletion. Chromosome by chromosome detailed statistics are reported in Supplementary Tables 4, 7, 23.

2.1.11. Global Tetraploid wheat Collection (GTC), genetic population structure

GTC is categorized into four main germplasm pools following the passport information Supplementary Data Set 2: wild emmer wheat, WEW; domesticated emmer wheat, DEW; durum wheat landraces, DWL; durum wheat modern cultivars, DWC. Genetic diversity was assessed using the Illumina iSelect 90K wheat SNP array, a high-density, high quality genotyping platform widely used in wheat genomics[43,48]. The use of a fixed SNP-platform implies the presence of some SNP-ascertained bias[132133]. The wheat 90K assay was developed by SNPs ascertained in a relatively wide discovery panel of both hexaploid and tetraploid wheats[48]. Interestingly, the SNPs ascertained in the A and B genomes of hexaploid wheat provided a relatively non-biased representation of the allele frequencies in the ancestral tetraploid A and B genomes probably due to gene-flow between wild and cultivated wheats. This feature was observed for hexaploid wheat, except for the D genome[134,135], suggesting that limited diversity loss took place in the A and B genomes during evolution of *T. aestivum* from domesticated tetraploid wheat and diploid *Aegilops tauschii*. As a consequence, when the wheat 90K assay was used to build a consensus map for tetraploid wheat[43], the composition of the SNPs ascertained in the wheat 90K assay allowed to genetically map a high number of technically functional, evenly distributed SNPs in the mapping populations obtained from the ancestral tetraploids (WEW, DEW) × modern durum crosses (e.g. the Svevo × Zavitan DWC × WEW mapping population and three additional DEW × modern durum mapping populations). This provided the tetraploid consensus map with an even marker density and genome coverage that was otherwise not possible to be reached by relying solely on the mapping populations obtained from intercrossing the modern durum wheat[43]. Based on these mapping results, the wheat 90K array was considered as a valuable SNP array for mapping and genetic diversity studies in the whole tetraploid wheat genome. For assessment, the folded site frequency spectrum (SFS) observed for WEW, DEW, DWL and DWC and their slopes are reported in Supplementary Fig. 35.

Phylogeny and population structure were assessed using a representative set of LD-based pruned SNPs[132] (r^2 0.50). To avoid overestimation of number of clusters due to the background LD and deviation from Hardy-Weinberg equilibrium we assessed different complimentary methods of analysis as detailed in section *1.3.4.* The analyses resulted in highly concordant picture and gave complementary information (*PCA vs*. *NJ* analysis *vs*. clustering). Overall, the four-main tetraploid germplasm groups (WEW, DEW, DWL, DWC) appeared largely differentiated, suggesting strong

demographic and founder effects and little evidences for polyphyletic origin. This observation was evident from both the *NJ* phylogenetic tree analysis as well as from the quantitative non-hierarchical clustering analysis methods.

PCA effectively illustrated the overall extent of genetic diversity for each taxon as compared to the others. DEW germplasm showed the broadest genetic diversity space, followed by DWL. DWC showed a relatively limited genetic diversity and a close relationship to specific DWL populations. *PCA* also highlighted the isolation by distance and evolution/adaptation to specific environmental conditions of Ethiopian populations compared to the main populations. Similar observation of demographic isolation could be made for the *T. turgidum* ssp. *carthlicum* and *T. turgidum* ssp. *turanicum* groups.

The four-independent quantitative non-hierarchical clustering methods converged to a highly similar and concordant overall population structure representation (Supplementary Data Set 3). The analyses classified the accessions into population clusters consistent with the origins and the expected genetic relationships. Pearson's *r* correlation coefficients and RMSE values calculated for all pair comparison among the Q matrices are reported in Supplementary Data Set 3, with the *r* coefficients ranging from 0.59 ($DAPC_{Kmeans}$ *vs.* $DAPC_{Ward}$ and $DAPC_{Ward}$ *vs. sNMF*) to 0.83 (*sNMF vs. ADMIXTURE*). RMSE ranged between 0.018 ($DAPC_{Kmeans}$ *vs.* $DAPC_{Ward}$) and 0.011 (*SNMF vs. ADMIXTURE*). *DAPC* classified the accessions into clusters (both *K*-means and Ward methods) by allowing much less quantitative admixture and cross-relationships than *sNMF* and *ADMIXTURE*. Therefore, *sNMF* and *ADMIXTURE* gave more cross comparable and relatively more informative data than *DAPC*. Additionally, the two methods explored in *DAPC* showed a relatively low correlation value as well as a low correlation to *sNMF* and *ADMIXTURE*. Based on all these considerations the results from *ADMIXTURE* were preferred for the detailed investigation at single taxon level.

Since the tetraploid wheat germplasm is highly structured, a well-defined population structure cannot be captured by a single *K* value. Rather, both WEW and DEW appeared to be highly structured based on layers of main populations (lower *K*) and additional well-defined populations (higher *K*), mostly related to the geographical origin of the accessions and, for DEW and DWL, to the human-driven dispersal process. In the DWL germplasm, admixture among main populations due to the cross Mediterranean exchange appeared to be an important component of the diversity. The detailed population structure obtained with the four methods is reported as Supplementary Data Set 3.

Though the *ADMIXTURE* and *fineSTRUCTURE* analysis assume a Hardy-Weinberg equilibrium, we found the results obtained from these analyses highly valuable for the tetraploid wheat germplasm structure description. *ADMIXTURE* was able to capture efficiently most of the

geographical-based structure while accurately estimating admixture events. The software detected the strongest population structure differentiation with *K* from 2 to 10, while, meaningful population differentiations in agreement with taxonomy, geographical origins and/or passport-pedigree information were generated at highest *K* (up to 20). Overall, at *K* equal to 20, as much as 1,053 (56.58%) accessions had membership *Q* value >0.7, or 1,440 (77.38%) accessions showed *Q* >0.5. *FineSTRUCTURE* further divided the whole dataset into 110 subgroups. The result was highly comparable to *ADMIXTURE* while dividing the germplasm at a finer level, thus a joint representation of population structure was reported by plotting the *ADMIXTURE* results for *K* from 2 to 20 (Supplementary Fig. 24 and Supplementary Data Set 2). At *K* = 2 WEW, DEW, *T. turgidum* ssp. *carthlicum* and the Ethiopian DWL were grouped together in one main group, while the second group includes *T. turgidum* ssp. *durum* landraces, cultivated accessions and other closely-related durum taxa. The Ethiopian DWL were then clearly discriminated from all the other accessions early at *K* = 3 and were further subdivided into two populations at *K* = 17 (pops. *Q9* and *Q10*). DEW separated from WEW at *K* = 5. Two main Western and Eastern DEW populations were divided at *K* = 6, while a further separation of Ethiopian and Indian DEW was evident at *K* = 7. *K* = 12-20 consistently structured the DEW germplasm into five populations. These DEW populations were genetically well characterized as highlighted by the high frequencies of non-admixed, highly-structured accessions (in DEW the frequency of accessions with *Q* membership value >0.7 was = 0.80). Based on *K* = 4-13, five DEW populations were clearly distinguished into: *i*) Western population from Fertile Crescent to Europe (*Q3, I*); *ii*) Western population mainly from Fertile Crescent with connection to European accessions (*Q8, II*); *iii*) Western populations from Turkey to Balkans and Russia (*Q6*); *iv*) Eastern population including Iran, Transcaucasia, Russia and Asia (*Q4*); *v*) Eastern population including Ethiopia and India (*Q5*). These results are remarkably similar to those obtained from previous diversity studies with cytogenetical structural variants[136]. The *T. turgidum* ssp. *carthlicum* population (*Q7*) was clearly distinguished from DEW at *K* = 12.

Most of DWL from various origins along with other *Triticum* taxa related to durum wheat were mainly delineated from the modern elite durum germplasm cultivated worldwide at *K* = 4. DWL were further subdivided at higher *K* values into subdivisions related to Western-Eastern spread and diffusion routes associated to the human migration and trade, confirming and detailing the earlier observations carried out in emmer and hexaploid wheat[107,136]. The division between Western (Mediterranean) and Eastern (Asian continental) populations was evident at *K* = 6-8. One of the two *T. turgidum* ssp. *turanicum* populations which was clearly distinct from the all the other DWL populations could be defined at *K* = 10 while other *T. turgidum* ssp. *turanicum* accessions classified together with true DWL. One population of landraces mainly from Fertile Crescent (Southern Levant)

was consistently defined at K = 13-20. Western Mediterranean DWL were defined at K = 14 and corresponded to three populations mainly originated from: *i)* Greece to Balkans (pop. *Q11*); *ii)* Fertile Crescent including Southern Levant, Cyprus to North Africa and Iberian peninsula (pop. *Q12*); *iii)* North Africa (Egypt to Morocco) to Iberian peninsula with relationships to a group of *T. turgidum* ssp. *turanicum* accessions, as already stated (pop. *Q13*). The Western Continental landraces included two populations from Russia (*Q16*) and from Turkey, Transcaucasia, Russia and Asia (*Q17*). The genetic and geographical subdivision of durum wheat landraces closely resembled the one previously observed for DEW, though very little residual of genetic cross-talk events between the two germplasms are evident from the *ADMIXTURE* analysis.

DWC mostly clustered into three to five main populations corresponding to three main germplasm pools bred worldwide. The first one includes cultivars and breeding lines bred directly at CIMMYT (The International Maize and Wheat Improvement Center) or, most recently, in Mediterranean breeding programs heavily relying on the innovative semi-dwarf and photoperiod insensitive CIMMYT germplasm (pop. *Q18*). The second one includes the North American germplasm (Canada and northern USA) and subsequently the germplasm bred in France and in Austria (*Q19*); this germplasm included mostly photoperiod sensitive cultivars. The third pool comprises germplasm locally bred in Mediterranean countries such as Italy and at the ICARDA (The International Center for Agriculture Research in the Dry Areas), mostly originating from crosses between the native North African and Syrian landraces and modern semi-dwarf varieties (*Q20*).

Overall, durum wheat accessions (both landraces and modern cultivars) showed a much higher level of admixture as compared to DEW and WEW. This was expected based on the breeding history of durum wheat[137-141] and from previous molecular analysis results from the durum wheat Mediterranean landrace pool[145,142].

Detailed *ADMIXTURE* and *fineSTRUCTURE* results for each of the four taxon groups are reported in Supplementary Data Set 2, together with the *Xusw59* score diagnostic for the *HMA-3B1a/b* allele for each of the 1,856 accessions.

fineSTRUCTURE and *ADMIXTURE* data highlighted 99 accessions that were either taxonomically misclassified or characterized by an excessive admixture or by mismatch between passport information and area of origin as determined by population structure. Therefore, out of 1,856 accessions, only 1,755 were assigned to populations based on the highest Q membership score (Supplementary Data Set 2) and were further considered for detailed population structure analysis and inspection of *TdHMA3-B1a/b* allelic distribution.

A more detailed view of population structure and genetic relationships among populations was obtained by running *ADMIXTURE* and *NJ* analysis within each germplasm group (Supplementary

Data Set 3, Supplementary Figs. 4, 5). The two methods gave very similar results, however the results from the *ADMIXTURE* carried out at increasing number of *K* populations gave a more detailed indication on the most probable historical relationships among taxa and extant germplasm populations. Both WEW and DEW, as compared to durum wheat, showed a highly structured genetic diversity, with a high rate of population assignment at the highest *K* value (*K* = 12 for WEW and *K* = 20 for DEW). The WEW germplasm showed a main division corresponding to the two populations from North-Eastern Fertile Crescent and Southern Levant (WEW-NE and WEW-SL, respectively). WEW-NE was further divided into several populations from Turkey, Iran and Iraq, while the WEW-SL population included distinct populations for Israel (three), Jordan, Syria, Lebanon (Supplementary Fig. 4). In this respect, *ADMIXTURE* and *NJ* results were very consistent and in agreement with previous phylogenetic analysis conducted in WEW with molecular markers at a much lower density[68,143].

Both DEW and DWL followed a very similar Northern-to-Southern Fertile Crescent (FC) and Eastern-to-Western radial dispersal patterns. In DEW germplasm we found six main populations (two from the Northern FC, Turkey-to-Transcaucasia/Iran (DEW-T-TRC-IRN), Turkey-to-Balkans (DEW-T-BLK); three from Southern FC: Southern Europe (DEW-SthEU), Southern Levant-to-Europe1 (DEW-SL-EU1), Southern Levant-to-Europe2 (DEW-SL-EU2), one grouping the Indian, Omani and Ethiopian DEW (DEW-ETH). Our results are in agreement with recent studies on population structure in worldwide emmer[144], and further provide a detailed insight into the reported evidence that emmer germplasm is delineated into four subgroups (Europeans, Balkans, Asians and Ethiopians) based on geographical factors.

In DWL, six main populations were also identified (two from the Northern FC: Turkey-to-Fertile Crescent (DWL-T-FC), Turkey-to-Transcaucasia (DWL-T-TRC), two from the Southern Levant FC: Southern-Levant-to-North Africa (DWL-SL-NA), Greece-to-Balkans (DWL-GRC-BLK), and two highly distinct populations including the Ethiopian landraces (DWL-ETH) and the *T. turanicum* (DWL-TRN), respectively. All the durum-cultivated accessions cluster to a further distinct germplasm that represents a wide branch of the durum North African landrace pool.

The genetic relationships among and within the main tetraploid taxa and populations were further investigated by removing the accessions showing high level of admixture (across taxa and across populations within taxa). Taxa and population differentiation was further tested using hierarchical ANOVA and by computing pairwise F_{st} and *Nei's genetic distances* (*GD*) among all populations (Fig. 4). This provided a confirmation for the northern-to-southern Fertile Crescent and eastern-to-western radial dispersal patterns and phylogeny.

The WEW-NE from Turkey, Iran and Iraq appears as the most probable ancestor of all DEW populations and durum wheat germplasm, as compared to WEW-SL populations (*Fst* and *GD* values consistently lower for all WEW-DEW and WEW-DW pairs). Furthermore, in the second main transition (DEW-DWL), the two *T. turgidum* ssp. *dicoccum* populations from Southern Levant FC that showed prime relationships with European accessions (DEW-SL-EU1 and DEW-SL-EU2) showed the lowest differentiation and genetic distance to all DWL populations (excluding *T. turgidum* ssp. *turanicum* population).

The modern durum wheat germplasm was mostly related to the two DWL populations from North-Africa (DWL-SL-NA) and Turkey-to-Transcaucasia (DWL-T-TRC). Further, DWL-GRC-BLK and to DWL-T-FC were specifically related to the modern durum varieties bred for the dryland areas at ICARDA and to the Italian germplasm adapted to the Mediterranean environments. The Ethiopian and *T. turanicum* durum were the most differentiated among the DWL populations, and their contribution to the modern durum germplasm was minimal. The successful, high-yielding potential modern CIMMYT germplasm released in the '80 (Altar84) appeared to be the most differentiated from all the DWL and DEW pools.

2.1.12. Genome-wide scan for genetic diversity and LD decay rate

An initial survey of genetic diversity based on *D* statistics averaged over 10Mb non-overlapping windows was calculated for each of the four main germplasm pools and averaged on a 10 Mb step (D_{10Mb}). Based on the *D* statistics we assessed the overall depletion in diversity due to domestication and selection. The four germplasm groups showed a reduction trend of whole-genome average genetic diversity: $D_{WEW} = 0.285$, $D_{DEW} = 0.254$, $D_{DWL} = 0.201$, $D_{DWC} = 0.192$. The average rate of diversity reduction from WEW to DWC is of 32.6% (statistics refer to single nucleotide polymorphism loci only).

The overall depletion in average genetic diversity was more evident in the pericentromeric recombination-depleted regions (average $D_{WEW} = 0.269$, $D_{DEW} = 0.220$, $D_{DWL} = 0.161$, $D_{DWC} = 0.151$) then in the distal highly-recombinogenic ones (average $D_{WEW} = 0.287$, $D_{DEW} = 0.295$, $D_{DWL} = 0.268$, $D_{DWC} = 0.250$). Detailed statistics on a chromosome-basis are reported in Supplementary Table 23.

The centromeric regions of WEW showed a relatively high level of diversity, except in chromosome 4A[145]. On the contrary, DEW, DWL and DWC experienced extensive demography/selection effects. Genetic diversity of modern DWC was strongly depleted in the centromeric regions of chromosome 1A, 2B, 4A, 5B, 6A (Supplementary Table 23). Strong

depletions already occurred in DEW for chromosome 4A and 6A and in DWL for chromosome 1A, 2A, 4A, 5B, 6A. In total, 74 regions of diversity depletion were detected (Supplementary Table 20): 19 for the transition from WEW to DEW, mostly peri-centromeric and including a strong depletion in diversity associated with the *BRT-B3* locus; 44 depletions for the DEW-DWL transition; and 11 depletions for the DWL-DWC transition. In six cases, the divergence in genetic diversity was in favor of an increase in genetic diversity for the most recent DWC.

A strong depletion in diversity in the peri-centromeric region of chromosome 1A (230 Mb and 846 HC-genes) specifically marked the DEW-DWL transition. This diversity depletion was already observed with genetic mapping data (as to marker density and presence of ample and recursive gaps in map continuity in the 1A pericentromeric region)[43,146,147]. Nevertheless, the availability of the Svevo genome assembly and the high-density SNP technology made it possible to clearly describe this and other strong bottleneck points.

The LD decay rate over physical distances was investigated both at the whole-genome level by estimating average LD decay rate over physical distances and at local chromosome region level using the focal marker method. LD decay rate was estimated separately for WEW, DEW, DWL and DWC. Based on a whole-genome average estimate of LD decay pattern, the speed of LD decay varied among the considered germplasm groups, as expected, and was most rapid for WEW, followed by DEW/DWL (comparable decay rate) and, lastly, DWC (Supplementary Fig. 36). Considering an LD threshold value of $r^2 = 0.2$, which corresponds to the background baseline LD in DWC, the LD reached this background level at 195 kb in WEW, at 1.4 Mb in DEW, at 1.6 Mb in DWL and at 4.5 Mb in DWC. The local speed of LD decay (focal marker method) was the lowest in the proximal regions of chromosomes (Supplementary Fig. 37). This is in line with strongly reduced recombination frequencies in the genetic centromeres of *Triticeae* chromosomes. This pattern was observed in all chromosomes except for chromosome 4B, most likely due to a paucity of markers at the peri-centromeric region of this chromosome. We note that, for all chromosomes, proximal regions harbored fewer SNPs per Mb, resulting in longer physical windows across which LD decay was estimated. The relatively fast LD decay rate observed in the WEW is similar to what has been observed for other *Triticeae* wild relatives[148,149] and makes this germplasm suitable for high resolution GWAS.

2.1.13. Demography and selection signals in the Global Tetraploid wheat Collection

Various metrics for the identification of divergence/selection signals including indexes less sensitive to ascertainment bias, such as haplotype-based methods[133] were used to reduce the impact of ascertainment bias, population structure and demographic factors on results. Furthermore, the

highly divergent Ethiopian DEW and DWL germplasm was excluded from the analyses. The SNP-based gene diversity index (*D*) averaged over 10Mb non-overlapping windows (Supplementary Table 20) was initially considered to estimate the extent of diversity loss along each transition. Then, the analysis was extended to a survey of five different divergence/selection signal metrics using a constant 25-SNP sliding window average, as represented in Fig. 5 and detailed in Supplementary Data Set 4. We applied the four main approaches currently used to detect selection signatures using five different metrics: *i)* the diversity reduction, assessed using the diversity reduction index (*DRI* = $D_{ancestral}/D_{derived}$), *ii)* the divergence, using both a single site index (F_{st}) and a haplotype-based frequency differentiation index, *hapFLK*, also 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*).

Selection metrics indicated several chromosome regions putatively under differentiation/selection. Results were reported mainly as 25 SNP- or 1Mb- averaged sliding windows to account for single site erraticism and to reveal the presence of strong signals extended to wide regions. Data are summarized in Fig. 5 and detailed in Supplementary Data Set 4.

Frequently, two or more metrics showed outlier signals in overlapping regions and were therefore considered as a single *selection region*. All selection regions identified either by a single selection metrics (singletons) or by multiple metrics were thereafter referred as *unique selection clusters*.

In total, Supplementary Data Set 4 summarizes 104 pericentromeric and 350 distal putative selection signal clusters, for a total of 454 *unique clusters*. On average, the peri-centromeric clusters showed a size of 107.7 Mb (95% size distribution: 2.7 to 369.1 Mb), while the distal clusters showed 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 to DWL-DWC (from 10.2 to 15.3 Mb for distal regions). A Venn diagram in Supplementary Data Set 4 summarizes the co-occurrence of metrics into *selection clusters*.

DRI metric signals. Among the four germplasm groups, WEW showed the highest average gene diversity and an evenly distributed diversity patterns across the whole genome, except for the pericentromeric regions of chromosome 2A and 4A which showed locally reduced diversity[146]. The WEW diversity pattern thus provided a valuable reference for cross-comparisons with DEW, DWL and DWC. Each of the subsequently domesticated/improved germplasm groups showed strong depletions in diversity widespread across the genome that independently occurred and progressively consolidated throughout the crop improvement process. With a few notable exceptions, we observed that once a strong depletion of diversity occurred (either during the WEW-DEW or DEW-DWL transition) the diversity did not recover in the subsequent derived germplasm (Figs. 5, 6). At the end

of the domestication, evolution and breeding process, the genome of the elite DWC progressively accumulated many regions showing near-fixation of diversity (Fig. 5). Notable exceptions were observed for chromosome 2A and 3A in the pericentromeric region where the DWC showed an increase in diversity as compared to DWL and DEW.

The fixed 25 SNP-averaged *DRI* metric confirmed the high rate of domestication-related diversity depletions in the pericentromeric regions compared to the distal regions, as already shown with the 10 Mb-survey (Supplementary Table 20). Chromosome regions showing adjacent non-interrupted SNP with *DRI* value >2 (equivalent to a diversity reduction of 50% or more), were 65, 111 and 75 for the WEW-to-DEW, DEW-to-DWL and DWL-to-DWC transitions, respectively (Supplementary Data Set 4) and accounted for 1,999.2, 2,138.5 and 1,086.6 Mb, respectively. As a result, on average the modern durum germplasm cumulated ca. 5 Gb of sequence with less than halved diversity compared to the ancestral WEW.

The projection onto the Svevo genome of 41 loci known to be under selection during emmer domestication, durum wheat evolution and breeding (Supplementary Table 12) gave a potential explanation for several clusters. Most of the strongest, pericentromeric diversity depletions (*DRI* >4) occurred already in the first WEW-to-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 coincident with the location of *VRN-A1*[150]. Furthermore, one of the two brittle rachis loci marking the early domestication process (harboring *BRT3-B1* at 96.2Mb[1]) showed a localized sharp reduction in diversity highlighted by F_{st} and *XP-CLR* metrics (Supplementary Data Set 4), but not detected with the 25-SNP based *DRI* window. The same region, then, underwent to a subsequent more extreme diversity reduction in the DEW-to-DWL transition (DRI_{max} = 3.4 in a region of 79.1-125.8 Mb).

The transition from domesticated emmer to durum was marked by two major depletions (DRI_{max} >4) in pericentromeric region of chromosomes 1A (one single region of 185 Mb) and 2B (two regions of 12.5 and 34.9Mb). Numerous other depletions were also observed in the non-pericentromeric regions, 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* was mapped between 31.9 and 36.1 Mb in Svevo genome[151,152] (Supplementary Data Set 4) and we observed in chromosome 2BS two severe diversity depletions (DRI_{max} = 4.3) in the regions 25.1-26.4 Mb and 33.3-49.1 Mb. The *Tg-2A* homoeolog, genetically mapped between position 21.2 and 31.7 Mb in Svevo based on the same mapping populations, was found associated to threshing-related traits in the 27.9-32.0 Mb region. The locus *Glu-1*, coding for glutenin subunits and located at 500.8Mb on chromosome 1A, is reported to be nearly fixed in modern germplasm for null allele *Glu-A1c*[153], was

associated to a local strong DRI_{max} signal = 3.2 in a 8Mb window. The domestication-related *Q-5A* locus (positioned at 608.8Mb), could not be directly related to any strong diversity selection signal, except for a local peak of diversity in DWC. While, the *Q-5B* harboring region[154] (positioned at 650.1 Mb) showed *D*, *DRI* and *XP-EHH* signals (Supplementary Data Set 4). Further extreme reduction in diversity associated to the DWL-to-DWC transition were observed in chromosomes 2B, 5B, 6A and 7B. The latter overlaps with several disease resistance (*Lr14a*) and grain yellow pigment content loci, including *Psy-B1*.

Divergence/haplotype based metric signals. Regions with allele frequency differentiation were searched using the popular F_{st} index 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. The pericentromeric regions showed extensive signals of divergence/selection that were simultaneously pointed out by two or more overlapping metrics, particularly in WEW-to-DEW and DEW-to-DWL transitions. This observation highlights that most of the loss-of-diversity and divergence signatures occurred during domestication.

Prioritization of selective signatures can be achieved by selecting the top ranking 1% distribution (Supplementary Data Set 4) and by searching for co-occurred signal clusters. Out of 454 selection signals identified by at least one metrics, 96 were identified by *DRI,* 184 by F_{st}, 167 by *XP-EHH* and 153 by *XP-CLR* (Venn diagram in Supplementary Data Set 4). Notwithstanding, 68 *DRI* signals co-occurred with at least another metric (71% of all *DRI* signals), particularly with F_{st}, followed by *XP-CLR* and *XP-EHH*. These signal clusters, combining both diversity reduction and divergence effects, can be considered as some of the most interesting putative selection clusters for prioritization.

We also provide a comparative alignment between selection signals and wheat genes and QTLs relevant for domestication/improvement. Among a set of 41 previously cloned loci, that are most probably target of selection, many loci co-located with regions marked by strong selection metrics (Supplementary Data Set 4). *TaGW2-A1*[155] on chromosome 6A (235.3 Mb) was associated to a strong signal detected by all metrics in the WEW-to-DEW transition, also associated with a steep decline of diversity in DEW. Similarly, *TaGW2-B1*[155,156] on 2B (300.8Mb) was mapped to a region with top *Fst* and *hapFLK* (WEW-to-DEW) and *XP-EHH* (DEW-to-DWL) signals. Additionally, *TaSus2-A1*, *TaSdr-A1*, and *TaCWI-A1* on chromosome 2A and their homoeologs on 2B were associated to multiple extended signals in WEW-to-DEW and in DEW-to-DWL transitions, while the durum germplasm showed extended regions of low diversity.

Among the loci mapped to non-pericentromeric regions:

- On chromosome group 3, *BRT-A1*[1] was associated to *XP-CLR* and *hapFLK* signals in WEW-to-DEW transition, while *BRT-B1* was associated to F_{st} and *XP-CLR* (WEW-to-DEW).

- On chromosome 5B, *Q-5B* was associated to a *XP-CLR* signal in WEW-to-DEW transition, nevertheless no evident selection signal could be found for *Q-5A*, possibly because of its interactions with other regulatory elements, such as the *miR172*, that could have diluted the signals.
- On chromosome 1A, *Glu-A1* was associated to *XP-CLR* signal in DWL-to-DWC transition.
- On chromosome 5A, *VRN-A1*[150] was associated to an F_{st} peak.
- On chromosome 2A, *Ppd-A1*[157] was associated to *XP-EHH* signal in DWL-to-DWC transition.
- On chromosome group 7, *TaTGW-7A*[158] mapped central to a *XP-EHH* signal in WEW-to-DEW transition, and to a close region with multiple *DRI, F_{st}, XP-EHH, XP-CLR* signals in DEW-to-WEW. The surrounding region was highly depleted in diversity in durum. Similarly, *TaTGW-7B* mapped to *XP-EHH* signal in both DEW-to-DWL and DWL-to-DWC transitions and in a *DRI* region in DWL-to-DWC.
- On chromosome 7B, the *Psy-B1* region[159] was coincident with two signals: *XP-CLR* in DEW-to-DWL and F_{st} in DWL-to-DWC transitions.
- On chromosome 4B, *Rht-B1* was not associated to any signals, probably because in the elite germplasm of durum wheat *Rht-B1* has not yet reached fixation (the North American germplasm is mostly composed of cultivars of conventional height). However, *Rht-B1* was mapped closely (<2 Mb) to an extended region with strong increase in diversity in DWC as compared to DWL.

Co-location between domestication-related QTLs and selection signals. QTLs from mapping populations obtained by crossing WEW or DEW with a modern durum (most frequently Langdon or Svevo) were potentially informative for the domestication-related QTLs and therefore were considered for co-location with the signals detected in the WEW-to-DEW and DEW-to-DWL transitions. The most prominent examples were the 46 QTLs for shattering (brittle rachis phenotype), threshability, threshing time, threshing efficiency and tenacious glumes phenotypes (see Supplementary Data Set 1 for QTL references). Twenty-one out of 46 were detected within selection cluster intervals, most of them located in the distal, recombining regions.

One brittle rachis QTL, not coincident with *BRT-3B*, mapped in Langdon x G18-16 on chromosome 1B at 556 Mb was close to a restricted region of 10 Mb showing multiple selection signals for both WEW-to-DEW and DEW-to-DWL transitions. A similar observation can be made for a second brittle rachis QTL mapped on chromosome 1B at 651 Mb in Svevo x Zavitan and for a threshability QTL mapped in Langdon x G18-16 chromosome 2A at 81 Mb.

Three QTLs mapped in the tough glume *Tg-2A* locus, marking the DEW-to-DWL transition and not yet cloned, were projected in the interval 31-36 Mb on chromosome 2B. This region showed a marked multiple *DRI*, F_{st}, *XP-EHH* and *XP-CLR* cluster specific of the DEW-to-DWL transition.

These data could contribute to a further characterization of this main domestication locus, mainly responsible for the selection of naked, free threshing tetraploid forms from ancestral DEW.

A threshing efficiency QTL mapped on chromosome 4A at 61 Mb is also coincident with a selection cluster specific for the DEW-to-DWL transition, observed between 67-99 Mb interval (*DRI* and F_{st} signals).

The grain protein content (GPC) is thought to have undergone a progressive dilution during domestication and subsequent evolution under domestication, due to the continuous unconscious selection for yield capacity mainly driven by starch accumulation. Four GPC QTLs were associated to strong and narrow selection signals in WEW-to-DEW and DEW-to-DWL transitions. For instance, we observed a QTL located on chromosome 1B at 10 Mb, centered in a narrow 5 Mb-wide region characterized by extensive WEW-to-DEW and DEW-to-DWL selection signals. A second GPC QTL on chromosome 2A mapped at 140 Mb in coincidence with a strong selection cluster. Finally, a third GPC locus mapped on chromosome 6A at 37 Mb was coincident with strong F_{st} and *hapFLK* signals in DEW-to-DWL transition.

A total of 175 QTLs related to grain yield components and mapped in WEW or DEW × DWC populations are listed in Supplementary Data Set 1. Forty-one of them were co-located or in close proximity of selection signal clusters detected mainly with *XP-CLR* and *XP-EHH*, and much less by *DRI* or F_{st} which were more relevant for domestication QTLs.

Finally, it is noteworthy to observe that the genome-projected position of the QTL peaks corresponding to *Ppd-A1*, *Ppd-B1* and *Sr13* from three previous QTL mapping studies based on the 90K SNP array, were highly predictive of the physical positions of the corresponding genes in Svevo, with a precision of 1 to 2 Mb maximum (Supplementary Data Set 4). This is encouraging for future activities of gene and QTL dissection and cloning and highlights the importance of relying on a high-quality reference genome.

Co-location of breeding-relevant QTLs and selection signals. QTLs for yield-related traits and disease response detected in populations obtained by crossing two durum wheat landraces or modern cultivars, were compared for co-location with the signals detected in the DWL-to-DWC transition.

A total of 192 yield-related QTL were projected on the selection signal map, 48 of them were coincident with selection signals in DWL-to-DWC transition (11 QTL were associated to *hapFLK*, 7 to *XP-EHH*, 6 to *XP-CLR*, 7 to multiple signals in clusters, and 7 to F_{st} and 10 to *DI/DRI* signals).

For the disease-response QTLs, a total of 179 QTL were considered and 59 of them were coincident with selection signals in DWL-to-DWC transition (6 QTL were associated to *hapFLK*, 15 to *XP-EHH*, 4 to *XP-CLR*, 10 to multiple signals in clusters, and 9 to F_{st} and 15 to *DI/DRI* signals).

2.2 Identification of a locus controlling cadmium accumulation in durum wheat grain

2.2.1 Cloning and functional analysis of Cdu-B1

The 90K SNP array results for the Svevo × Zavitan and Kofa × W9262-260D3 (G9586) populations were in strong agreement and indicated that there were significant markers for Cd accumulation in the grain located on chromosome 5B near the *Cdu-B1* QTL interval described by[78] (Supplementary Fig. 6a). Additional markers for *Cdu-B1* from the Kofa × W9262-260D3 population were used to screen 5,081 F_2 plants derived from 8982-TL-L × 8982-TL-H, and the locus was resolved to an interval flanked proximally by *ScOpc20* and distally by *Xusw14* and *Xusw53.* The low-Cd accumulators had the same molecular variants within the flanking markers and contained nearly one third the amount of Cd in grain when compared to the high-Cd accumulators, which contained alternate molecular variants across the interval (Supplementary Table 21). The flanking molecular markers *Xusw53* and *ScOpc20/Xusw49* were inconsistent with respect to Cd accumulation in grain, indicating that they are outside of the *Cdu-B1* locus. Thus, these flanking markers for *Cdu-B1* were mapped to the DW genome, which resulted in a physical interval spanning positions 563,586,136 to 567,855,527 bp (Supplementary Table 24, Supplementary Fig. 6b). Exome sequencing of the high- and low-Cd accumulators derived from 8982-TL-L x 8982-TL-H confirmed that between the flanking markers, there were several markers that were in agreement with the parental line that had the same Cd accumulation phenotype (Supplementary Fig. 6b). This region was also coincident with the marker peaks identified by single marker regression in the Svevo × Zavitan and Kofa × W9262-260D3 populations (Supplementary Fig. 6a).

To identify specific genes from the refined interval for *Cdu-B1* that may contribute to the differences in Cd accumulation within grain, we performed a comparative analysis of *Cdu-B1* between the genomes of Svevo and Zavitan. The markers co-localized to a region approximately 4.5 Mbp in length on chromosome 5B spanning positions 572,890,476 to 577,183,949 bp in the WEW genome (Supplementary Table 24). Alignment of *Cdu-B1* between Svevo and Zavitan revealed a region of increased nucleotide variation between the flanking markers *Xusw49* and *Xusw59* and ended near the markers *Xusw50* and *Xusw51* (Supplementary Fig. 6c). The gene *TRITD5Bv1G197370* (WEW ortholog *TRIDC5BG060070*) was located within the region of increased nucleotide variability and is annotated as a P_{1B}-type Heavy-Metal ATPase 3 (HMA3) transporter (*TdHMA3-B1* herein) and is located in the best matching region of *HMA3* from both rice and *Brachypodium*. Furthermore, the marker *Xusw59* spanned a 17 bp duplication within the first exon of the gene in Svevo compared to Zavitan (Supplementary Fig. 7); and when screened against wheat populations this marker perfectly discriminated high- and low-Cd accumulators (Supplementary Fig. 8).

The *Cdu-B1* physical interval includes 48 high-quality gene models not annotated as transposable

elements (Supplementary Table 13). We investigated potential candidate genes, other than *TdHMA3-B1*, in this region that could contribute to Cd sequestration in root tissue, the mechanistic basis for the *Cdu-B1* phenotype[83]. Examination of AHRD (Automatic assignment of Human Readable Descriptions), *Pfam*, *InterPro*, and gene ontology (GO) annotations identified two additional putative genes with annotations relevant to a Cd sequestration phenotype: *TRITD5Bv1G197460* and *TRITD5Bv1G197500* (both annotated as putative ATP synthase subunit alpha; GO:0006811 and descendants, ion transport). However, ATP synthesis coupled proton transport is unlikely to account for Cd sequestration in root tissues. Only *TdHMA3-B1* is functionally consistent with the *Cdu-B1* phenotype.

We cloned the homoeologous pairs of full-length *TdHMA3* cDNAs (*TdHMA3-A1* and *TdHMA3-B1*) from low (8982-TL-L) and high (8982-TL-H) Cd accumulating isogenic DW lines. The *TdHMA3-A1* cDNA of low- and high-Cd isogenic lines (KF683291 and KF683292, respectively) are identical to each other and are identical or nearly identical to the CDS of genes *TRITD5Av1G202240* (Svevo) and *TRIDC5AG056040* (Zavitan, 3 SNPs, 2 missense), respectively. In contrast, the low and high Cd isogenic lines have different alleles for *TdHMA3-B1* (Supplementary Fig. 7a). The low-Cd allele (*TdHMA3-B1a*, KF683294) is nearly identical to Zavitan *TRIDC5BG060070* (4 SNPs, 3 missense), whereas the high-Cd allele (*TdHMA3-B1b*, KF683295), which is identical to Svevo *TRITD5Bv1G197370*, has a 17 bp duplication in exon 1 (Supplementary Fig. 7a). Except for the 17 bp duplication, the *TdHMA3-B1a* and *TdHMA3-B1b* cDNA are otherwise identical. Comparisons using sequence data from Chinese Spring and Sumai 3 suggests that these hexaploid wheats also carry the low grain Cd allele *HMA3-B1a* (Supplementary Fig. 7a). Thus, the source of low Cd is likely to be shared between hexaploid and tetraploid wheat, which conflicts with a recent theory of alternate sources between species[160]. Although allele *TdHMA3-B1b* has an intact exon-intron structure (Supplementary Fig. 7b), the 17 bp duplication causes a frame-shift and premature stop codon that results in a severely truncated predicted protein in the high-Cd genotypes (ORF1, 60 aa). Phylogenetic analysis (Supplementary Fig. 7c; Supplementary Data Set 12) places *TdHMA3-A1* and *TdHMA3-B1a* in the Poaceae *HMA3* clade within the P_{1B-2}-ATPase sub-group[161], which show substrate specificity for Cd, Zn, Co, and Pb[162]. *TdHMA3-A1* and *TdHMA3-B1a* have eight predicted transmembrane helices (Supplementary Fig. 9, TM1-8), a topology typical of P_{1B}-ATPases[163-165]. The *TdHMA3* proteins display features typical of the P_{1B-2}-ATPase sub-group (Supplementary Fig. 9), including the transmembrane metal binding site motif, CPC(x)4SxP, in TM6[166], the N(x)7K(x)10,20DxG(x)7N signature sequence in TM7 and TM8[167], and the N-terminal heavy-metal-associated (HMA) domain (Prosite: PS50846, Lys-28..Val-94) motif, GxCCxxE[166]. Many plant P_{1B-2}-ATPases, such as AtHMA2 and AtHMA4, have long C-terminal regions after the last transmembrane domain[166] that,

due to the presence of multiple cysteine pairs and histidine residues, exhibit metal-binding activity[168,-170]. The C-terminal regions of TdHMA3-A1 (102 aa) and TdHMA3-B1 (114 aa), which each contain two Cys-Cys pairs and three His residues (Supplementary Fig. 9), are unlikely to exhibit metal binding activity. In contrast, the C-terminal regions of bread wheat TaHMA2-A1 (KF933095), TaHMA2-B1 (KF933096), and TaHMA2-D1 (KF933097) are between 281 and 322 aa residues and contain 2–4 Cys-Cys pairs and 60–68 His residues.

The phylogenetic and structural characterization of TdHMA3 proteins suggests they should function as tonoplast-localized Cd and Zn transporters. We tested these predictions using heterologous expression in yeast. The low-Cd allele, *TdHMA3-B1a*, complemented Cd- and Zn-sensitivity of yeast knockout strains *ycf1* and *zrc1cot1*, respectively (Supplementary Fig. 10, Supplementary Fig. 11), and TdHMA3-B1a-GFP was observed to localize to the tonoplast (Supplementary Fig. 12). In addition, *TdHMA3-B1a*-mediated complementation of *ycf1* and *zrc1cot1* was attributable to P-ATPase metal transport activity, rather than to detoxification by metal chelation. Mutation of the conserved phosphorylatable aspartate residue (D411A) in the DKTGT motif (Supplementary Fig. 9), which is necessary for P-ATPase transport activity[94], abolished complementation of *ycf1* and *zrc1cot1* (Supplementary Fig. 11) and reduced cellular accumulation of Cd and Zn (Supplementary Fig. 13). These results suggest that TdHMA3-B1a contributes to vacuolar sequestration of Cd and Zn.

The *TdHMA3-B1b* ORF is highly truncated (183 bp) and only the initial 10 aa residues of the translated product are conserved with TdHMA3-B1a (Supplementary Fig. 9). Unsurprisingly, *TdHMA3-B1b* failed to complement *ycf1* and *zrc1cot1* (Supplementary Fig. 10) and did not localize to the tonoplast (TdHMA3-B1b-GFP was consistent with nuclear localization; Supplementary Fig. 12). Given that the disruptive duplication in *TdHMA3-B1b* occurs near the 5' terminus of exon 1, we tested if an alternative, 5'-truncated ORF of *TdHMA3-B1b* could functionally substitute for *TdHMA3-B1a*. The largest alternative ORF for *TdHMA3-B1b* (ORF2, Supplementary Fig. 7b) is truncated at the 5' end by 449 bp. Translation of ORF2 begins at Met-145 (relative to *TdHMA3-B1a*; Supplementary Fig. 9, and thus retains the six core transmembrane domains (TM3 to TM8) and catalytic sites that constitute the minimal functional transport unit within P-ATPases[170]. Although TdHMA3-B1b-ORF2-GFP localized at the cell periphery and at a perinuclear location (Supplementary Fig. 12), which is consistent with membrane localization of TdHMA3-B1b-ORF2 at the plasma membrane and endoplasmic reticulum respectively, *TdHMA3-B1b*-ORF2 failed to complement *ycf1* and *zrc1cot1* (Supplementary Fig. 10). Metal-induced growth inhibition of *ycf1* and *zrc1cot1* expressing *TdHMA3-B1b*-ORF2 was equivalent to that of these strains expressing empty vector (EV). The *TdHMA3-B1b* allele is unable to transport Cd or Zn and is therefore non-functional.

In contrast to *TdHMA3-B1a*, *TdHMA3-A1* complemented Zn-sensitive *zrc1cot1*, but not Cd-sensitive *ycf1* (Supplementary Fig. 10). Compared to the empty vector control, expression of *TdHMA3-A1* in *ycf1* resulted in a Cd-hypersensitive phenotype (Supplementary Fig. 10, Supplementary Fig. 11). This is different from Cd-hypersensitivity seen for *OsHMA3* expressed in yeast[171,172], which occurs when OsHMA3 fails to localize to the tonoplast. OsHMA3 remains in the endomembranes and transports Cd to the ER lumen, which is highly sensitive to Cd accumulation[173]. Yeast codon-optimized *TdHMA3-A1* and *TdHMA3-B1a* both localize to the tonoplast of *ycf1* (Supplementary Fig. 12), a site where HMA3-mediated Cd transport into the vacuole would be expected to increase Cd tolerance of *ycf1*, as seen for *TdHMA3-B1a*, rather than increase Cd sensitivity, as seen for *TdHMA3-A1* (Supplementary Fig. 11). Although the Cd-hypersensitive phenotype is attributable to P-ATPase metal transport activity, as shown by the elimination of Cd-hypersensitivity (Supplementary Fig. 11) and the reduction in cellular Cd accumulation (Supplementary Fig. 13) in *ycf1* expressing *TdHMA3-A1*-D411A (transport activity knockout), TdHMA3-A1 Cd transport activity is not supportive of TdHMA3-A1-mediated vacuolar sequestration of Cd.

The primary *in planta* effect of the non-functional allele, *TdHMA3-B1b*, is reduced retention of Cd in the roots. Near-isogenic lines (NILs) of DW, low and high for Cd accumulation in grain (homozygous for alleles *TdHMA3-B1a* and *TdHMA3-B1b*, respectively), accumulated similar amounts of Cd (whole-plant) when grown to maturity in hydroponic culture, but the high Cd NIL accumulated between 2 to 5-fold more Cd in grain and shoots at all stages during grain filling[83] (Supplementary Fig. 14). In contrast, the low Cd NIL, with the functional *TdHMA3-B1a* allele, consistently retained more Cd in the roots (Supplementary Fig. 14). The differences between NILs in grain and shoot Cd accumulation during grain filling were also observed under field conditions (Supplementary Fig. 15). In both hydroponic and field experiments, Cd transport from the roots to the shoots continued throughout grain filling, as shown by increasing shoot Cd content (Supplementary Fig. 14, Supplementary Fig. 15), indicating that increased mobilization of root Cd pools in high Cd lines could directly contribute to grain Cd accumulation. Alternatively, Cd accumulated in leaf and stem tissues prior to grain filling may be remobilized to the grain during grain filling[174] and high Cd lines have larger shoot Cd pools at anthesis[83] (Supplementary Fig. 14, Supplementary Fig. 15) that could be remobilized to the grain.

The phenotypic effect of the *Cdu-B1* locus is Cd-specific. Low and high Cd-accumulating DW genotypes accumulate equivalent amounts of essential micronutrients in shoots and grain, while differing by more than two-fold in Cd accumulation in these tissues[83,101] (Supplementary Fig. 14, Supplementary Fig. 15, Supplementary Table 14). However, TdHMA3-B1a is a tonoplast-localized

Cd and Zn transporter (Supplementary Fig. 10, Supplementary Fig. 11, Supplementary Fig. 12). The dual Cd/Zn transport activity of TdHMA3-B1a is consistent with other HMA3 P_{1B-2}-ATPases. The Arabidopsis homolog of *TdHMA3-B1*, AtHMA3, is a tonoplast-localized Cd, Zn, Co, and Pb transporter[175,176], and *Noccaea caerulescens* HMA3 transports Cd and Zn[177]. Knockouts of *OsHMA3* are phenotypically similar to *Cdu-B1*; they show increased root-to-shoot translocation and grain accumulation for Cd, but not Zn[171,172]. However, *OsHMA3* overexpression increases accumulation of Cd and Zn in rice root tissue[178], indicating that OsHMA3 is a Cd and Zn transporter that increases Cd and Zn sequestration in vacuoles of rice roots. The fact that the phenotypic effect of suppression or overexpression of *OsHMA3* only effects Cd accumulation in shoots and grain[171,178] is most likely explained by homeostatic compensation of other Zn transport systems as suggested for *AtHMA3*[179] and *OsHMA3*[178]. Similar to *OsHMA3*, the Cd-specific effect of *Cdu-B1* on Cd accumulation in shoots and grain of DW may be attributable to homeostatic responses that maintain normal Zn accumulation patterns in the absence of a functional *TdHMA3-B1*. An alternative hypothesis is that the HMA3 homoeologs, *TdHMA3-A1* and *TdHMA3-B1*, epistatically produce the Cd-specific *Cdu-B1* phenotype. In low Cd genotypes, TdHMA3-A1 and TdHMA3-B1a act redundantly to transport Zn into root vacuoles, while TdHMA3-B1a alone sequesters Cd in root vacuoles. In high Cd genotypes, vacuolar sequestration of Cd is lost (*TdHMA3-B1b* is non-functional), but vacuolar sequestration of Zn is maintained by TdHMA3-A1. The molecular mechanism that enables TdHMA3-A1 to discriminate between Cd and Zn is the focus of ongoing analyses. Knowledge of the mechanism for Zn-specific vacuolar sequestration in TdHMA3-A1 may be applied to other P_{1B-2}-ATPases such as *OsHMA2*, a plasma membrane transporter of Cd and Zn that contributes to their root-to-shoot translocation[180,181]. Generation of Zn-specific HMA2s will support breeding of Zn-enriched crops that show reduced accumulation of Cd, an important goal for the development of safe, micronutrient biofortified grain crops[182].

2.2.2 Allele differentiation and genetic diversity at the Cdu-B1 locus

To trace the diffusion and origin of *TdHMA3-B1a/b* alleles, the tetraploid diversity panel was completely genotyped for the perfect marker *Xusw59*. The distribution of *TdHMA3-B1a/b* based on taxonomy, population genetic structure and geographic origin of the accessions is summarized in Fig. 8 and Supplementary Fig. 16. Detailed results are reported in Supplementary Data Set 2. The distribution based on geography and *fineSTRUCTURE*/*ADMIXTURE* defined populations are summarized in Supplementary Table 15 and Supplementary Table 16.

The complete set of surveyed WEW, covering all of the main domestication areas described for Turkey (North-Eastern Iran) and Southern Levant (Lebanon, Syria, Jordan, Israel) showed no

presence of *TdHMA3-B1b*, with all of the surveyed germplasm showing fixation for the wild type allele. The non-functional *TdHMA3-B1b* allele most probably originated in the DEW germplasm, where it can be observed in the majority of populations. The *TdHMA3-B1b* frequency in the main populations ranged from 0.061 (*Q3_T. turgidum* ssp. *dicoccum*_West_Fertile_Crescent/Southern_Levant_Europe_I) to 0.378 in the DEW from Fertile Crescent Turkey to West-Balkans and Russia (*Q6_T. turgidum* ssp. *dicoccum*_West_Turkey_West-Balkans_Russia). The latter represent the first example of increased *TdHMA3-B1b* allele frequency in the tetraploid wheat germplasm. Overall, the *TdHMA3-B1b* allele frequency in DEW was 0.140. All the six *T. turgidum* ssp. *durum* landrace' populations showed the presence of *TdHMA3-B1b* allele, with an overall frequency of 0.262, a frequency significantly higher (t test, $P \leq 0.0001$) as compared to DEW. Among the DW and other DW-related tetraploid populations, only one of the two populations of *T. turgidum* ssp. *turanicum*, related to durum wheat, showed an absence of *TdHMA3-B1b* in the sampled accessions (*Q14*). Among the durum landrace populations, Western population *Q13* (Egypt to Morocco_Spain, including *T. turgidum* ssp. *turanicum* accessions) had the highest *TdHMA3-B1b* frequency (0.744), followed by the *Q9/Q10* Ethiopian populations (0.324 and 0.263), the Western population *Q12* (0.259), and the Eastern *Q16* population (0.242).

All the three major DWC populations identified showed a considerable increase in *TdHMA3-B1b* frequency. Both the modern North-American (*Q19*) and Mediterranean (*Q20*) germplasms showed a very high *TdHMA3-B1b* frequency (0.805 and 0.897, respectively), while the CIMMYT-related germplasm (*Q18*) showed a lesser prevalence of *TdHMA3-B1b* (0.550).

Based on the iSelect 90K SNP data from the tetraploid diversity panel, linkage disequilibrium extent and decay relative to *TdHMA3-B1* were assessed in the main DEW, DWL and DWC germplasms. For the DEW germplasm, LD decay relative to *TdHMA3-B1* was sharp with no neighboring SNPs having r^2 to *TdHMA3-B1* that was above the background (Supplementary Fig. 38), while in DWL a strong LD to *TdHMA3-B1* ($r^2 \geq 0.50$) extended from 562,676,669 (*IWB48878*) to 567,855,342 (*IWA3226*). In this region 10 SNPs had an r^2 to *TdHMA3-B1* >0.50, with the greatest being 0.79 for *IWA2255* (565,282,544), 0.70 for *IWB51494*, *IWA2565* and *IWA6024* (567,511,583), and 0.84 for *IWB69410* (567,825,666), with a trend of reduced r^2 going towards *TdHMA3-B1*. In DWC, LD to *TdHMA3-B1* extended further, from *IWB57597* (559,772,742) to *IWB71502* (568,395,895), including 20 SNPs strongly associated to *TdHMA3-B1*. Thus, in both germplasms, LD analysis identified an extended chromosome region of 5.2 Mb in DWL and 8.7 Mb in DWC marked by co-selected markers including *TdHMA3-B1* that has undergone common episodes of allele frequency differentiation/fixation.

F_{st} statistic was then used to explore patterns of genetic differentiation/selection among the four main germplasms considered along chromosome 5B (SNP LD to *TdHMA3-B1*, F_{st} in Supplementary Data Set 5). These statistics gave indications concordant with those from the relative *D* genetic diversity patterns, pointing out the presence of an extended region downwards to *TdHMA3-B1* that experienced mild occurrence of genetic drift/selection. The functional annotation of this region was investigated in detail, including ± 2 Mb (Supplementary Data Set 6). The region included 219 unigene groups (present in both DW and WEW) common to DW and WEW, 73 singleton genes for WEW, and 58 singleton genes for DW. Apart from *TdHMA3-B1,* the region included several cloned and well-characterized wheat genes, such as a serine/threonine protein kinase gene (TRITD5Bv1G194940), three lipoxygenase genes, (TRITD5Bv1G195300, TRITD5Bv1G195310, TRITD5Bv1G195400), an NLR disease resistance protein close to *Yr10*, *Sr35*, *Lr10* (TRITD5Bv1G196020), two cytochrome P450 genes (TRITD5Bv1G196140, TRITD5Bv1G196150), one dehydration-responsive element binding factor protein close to CBF-B11 (TRITD5Bv1G198860), two receptor-like kinase (RLK1, PR1-RK1, XA21-like, TRITD5Bv1G199150 and TRITD5Bv1G199280). At the extreme distal side of the region the functional analysis also showed the presence of *PHYC* and *VRN-B1* genes (TRITD5Bv1G200370, TRITD5Bv1G200510). Overall, the GO Slim summary analysis showed that the regions showed enrichment relative to the whole DW Svevo genome for genes having catalytic, protein binding, nucleotide binding, transferase, DNA binding, kinase, transcription factor, signal transducer, receptor and motor activity (Supplementary Data Set 6).

Beside the hypothesis that the non-functional *TdHMA3-B1b* allele could represent a selective advantage per se, the presence of LD in the region gives also some credit to the hypothesis that the high-Cd allele may have been selected inadvertently via linkage with functional mutations in nearby genes of agronomic importance. The Zavitan-Svevo gene atlas of high impact variants (Supplementary Data Set 9) offers a potential hypothesis. The *Cdu-B1* interval contains a series of genes belonging to BTB/POZ-containing proteins annotated as involved in suppression of axillary branching through an increase in apical dominance, which is a commonly bred trait among many domesticated crop plants[183]. These genes could therefore represent a selection target. The BTB/POZ gene cluster was made of 8 HC genes in Zavitan, while only 7 genes (5 HC: TRITD5Bv1G196660, TRITD5Bv1G196820, TRITD5Bv1G196830, TRITD5Bv1G196850, TRITD5Bv1G196870 and 2 LC: TRITD5Bv1G196880, TRITD5Bv1G196910) were present in the same locus in Svevo. One of the Svevo LC genes was shown to carry a putative loss of function variant (a new stop codon) in position 562,601,512 (less than 1.3 Mbp proximal to *TdHMA3-B1*) in comparison with the corresponding HC gene in Zavitan (TRIDC5BG059880). A recent work shows that the suppression

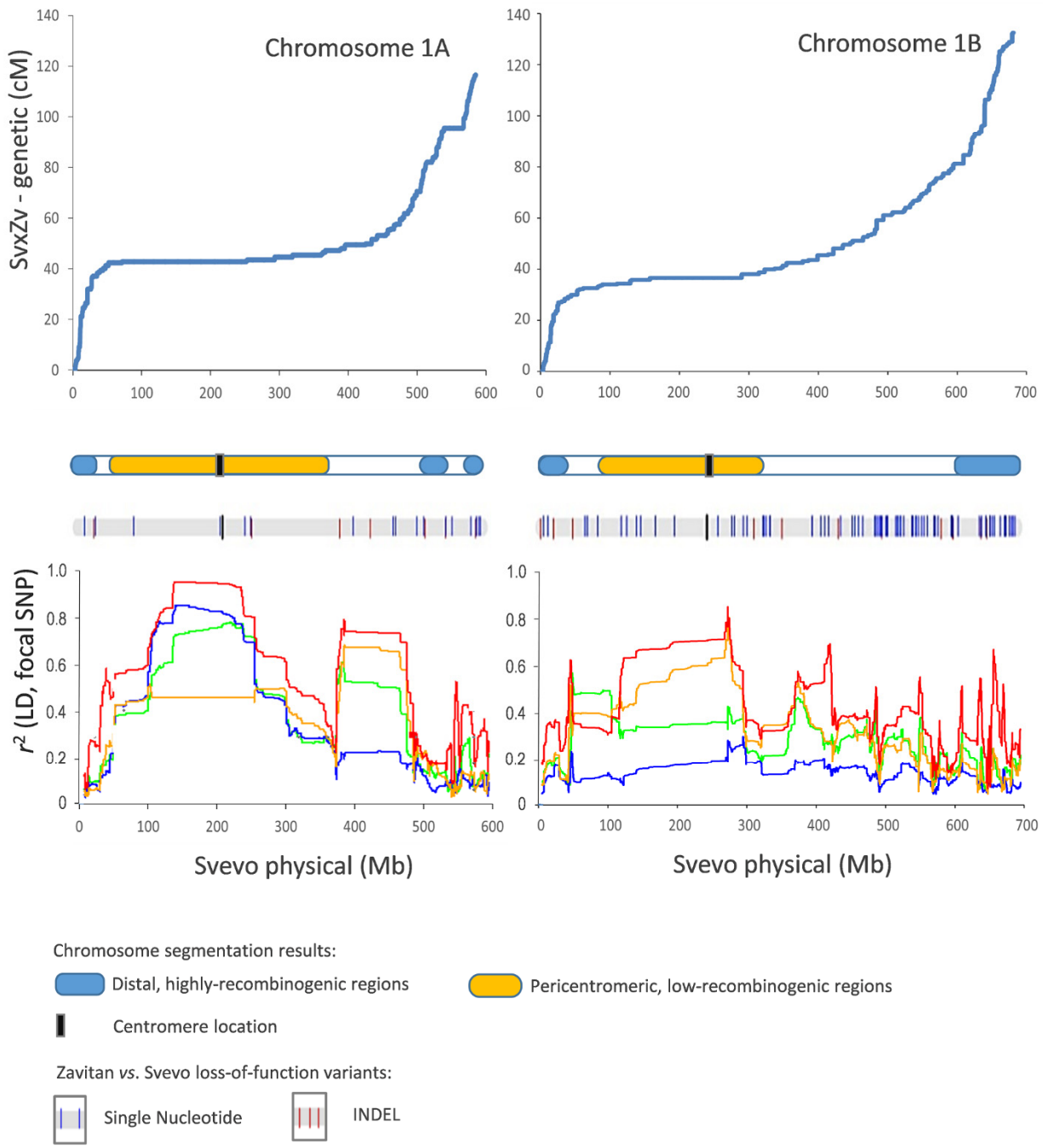
of axillary branching essential to the domestication of maize from teosinte is mediated by the targeting of a BTB/POZ-containing gene in maize (*tru1*) by *teosinte branched1* (*tb1*)[183]. Because of their influence on crop plant architecture, both *tru1* and *tb1* appear to be targets of selection in maize; and the authors suggest the *tru1-tb1* pathway may offer a blueprint for domestication in other grasses.

An inspection of the three main DW germplasm groups (CIMMYT, North American, Mediterranean/ICARDA), suggests that the use of parents with a high frequency of *TdHMA3-B1b* allele in the early breeding phase could have been responsible for the high frequency of *TdHMA3-B1b* allele in modern DW cultivars.

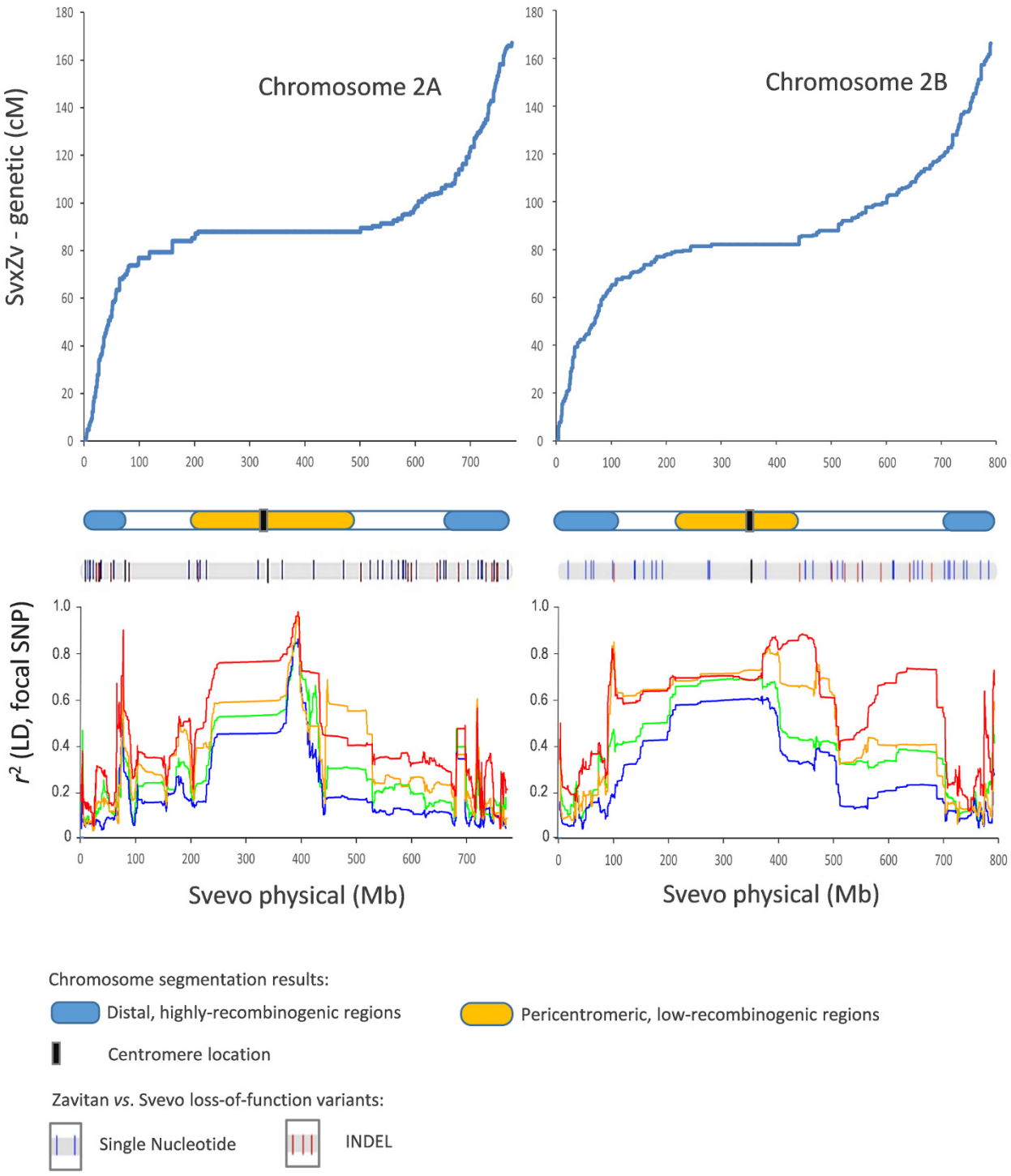
2.2.3. Validation of genome structure within Cdu-B1

Comparative genomics and whole genome chromium sequencing were performed to validate the structure and scaffold placement within *Cdu-B1*, as determined by POPSEQ and Hi-C. *Cdu-B1* in the DW genome includes three scaffolds that were joined by POPSEQ and Hi-C (Scaffold3184-1, Scaffold2417, and Scaffold14160-1) (Supplementary Fig. 39a). These concatenated scaffolds were aligned to assembled bacterial artificial chromosome sequences from the durum cultivar Langdon, and BAC 790-E17 was able to align to both Scaffold2417 and Scaffold14160-1, indicating that these scaffolds are properly placed and oriented in the DW assembly of Svevo. Similarly, separate scaffolds from two hexaploid wheat assemblies, TGAC v1.0 and Triticum 3.1, aligned to Scaffold3184-1 and Scaffold2417, or to Scaffold2417 and Scaffold14160-1, thereby supporting their placement and orientation (Supplementary Fig. 39a). Furthermore, whole genome alignment of WEW and DW showed strong collinearity between assemblies before and after gaps between scaffolds (Supplementary Fig. 39b), which facilitated variant calling between genomes and the detection of the 17bp duplication within *TdHMA3-B1* (Supplementary Fig. 7a). Finally, Chromium sequencing of large DNA molecules from Svevo indicate that single DNA molecules were able to span all scaffolds including breakpoints or gaps between them, providing additional evidence to the high-quality of the Svevo assembly (Supplementary Fig. 39c).

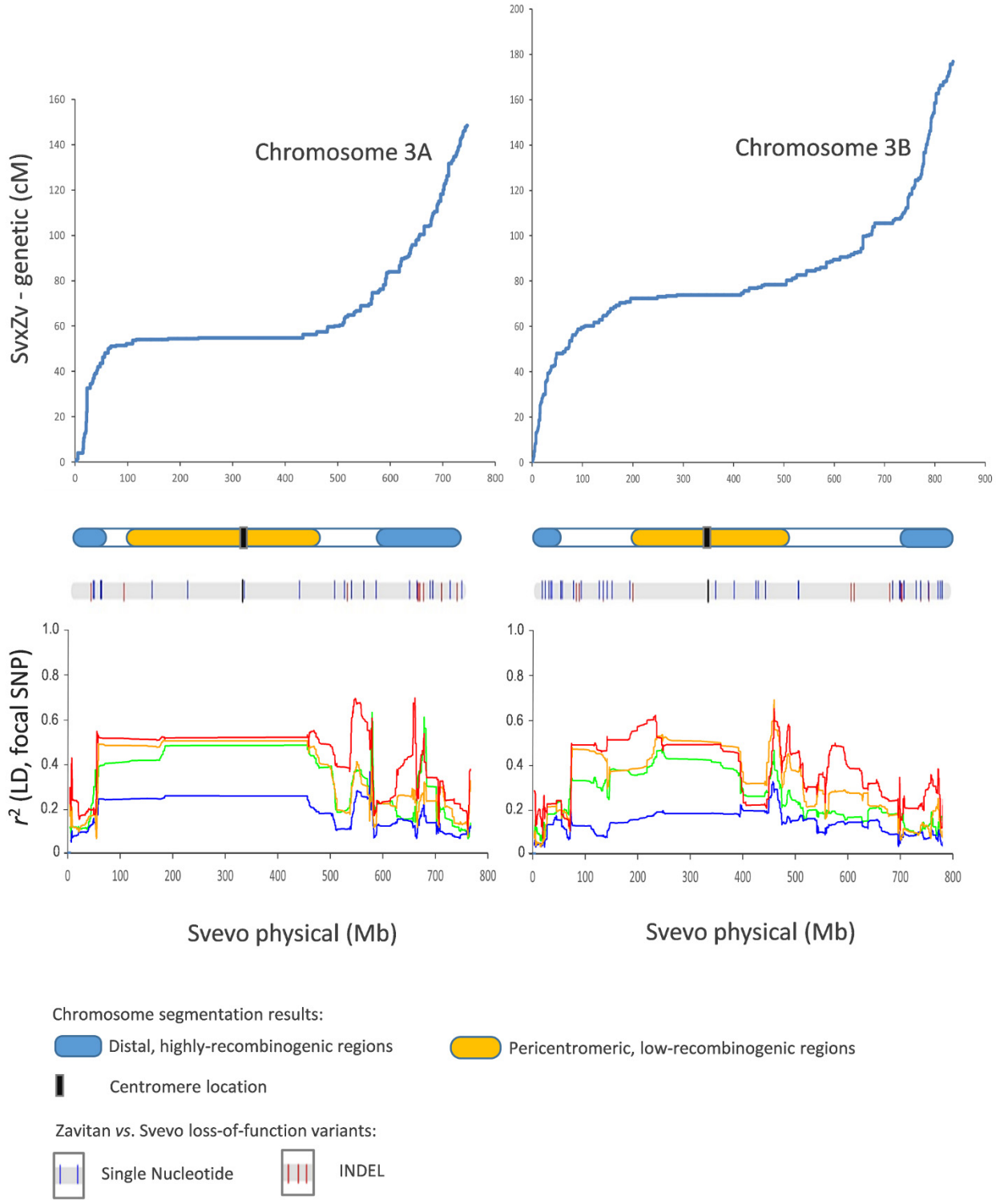
Supplementary Figures: 1 to 39



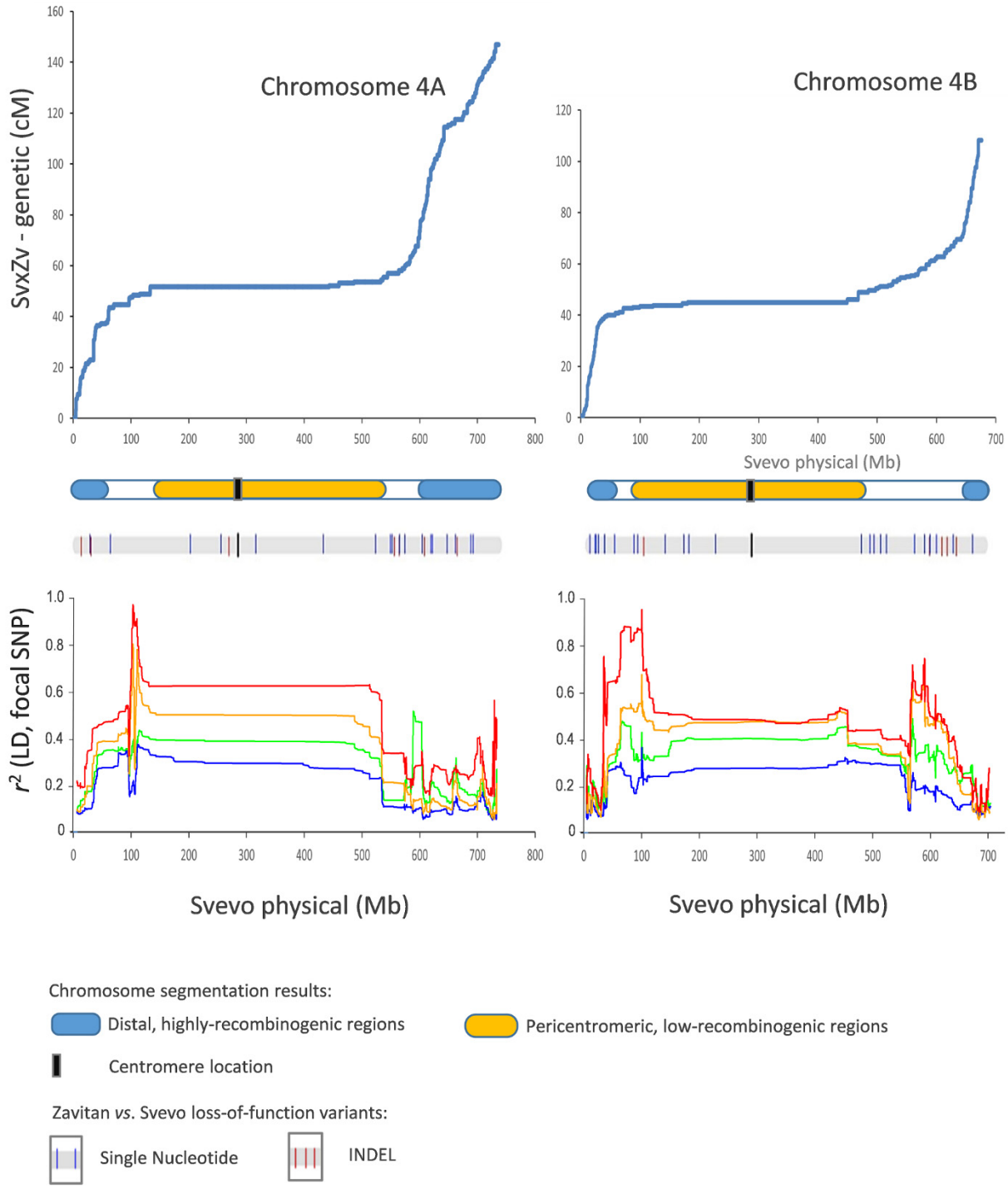
Supplementary Figure 1. Recombination rate, location of high- and low-recombinogenic regions, positions of 597 putative high-impact functional variants (Svevo *vs*. Zavitan - all variants represent putative loss-of-function or severe modification of the functional Zavitan alleles), LD pattern (focal SNP) across the 14 chromosomes of the Svevo genome (blue line: WEW, green line: DEW, yellow line DWL, red line DWC). Bold black ticks = centromeres; raised blue ticks = SNPs; lowered red ticks = indels.



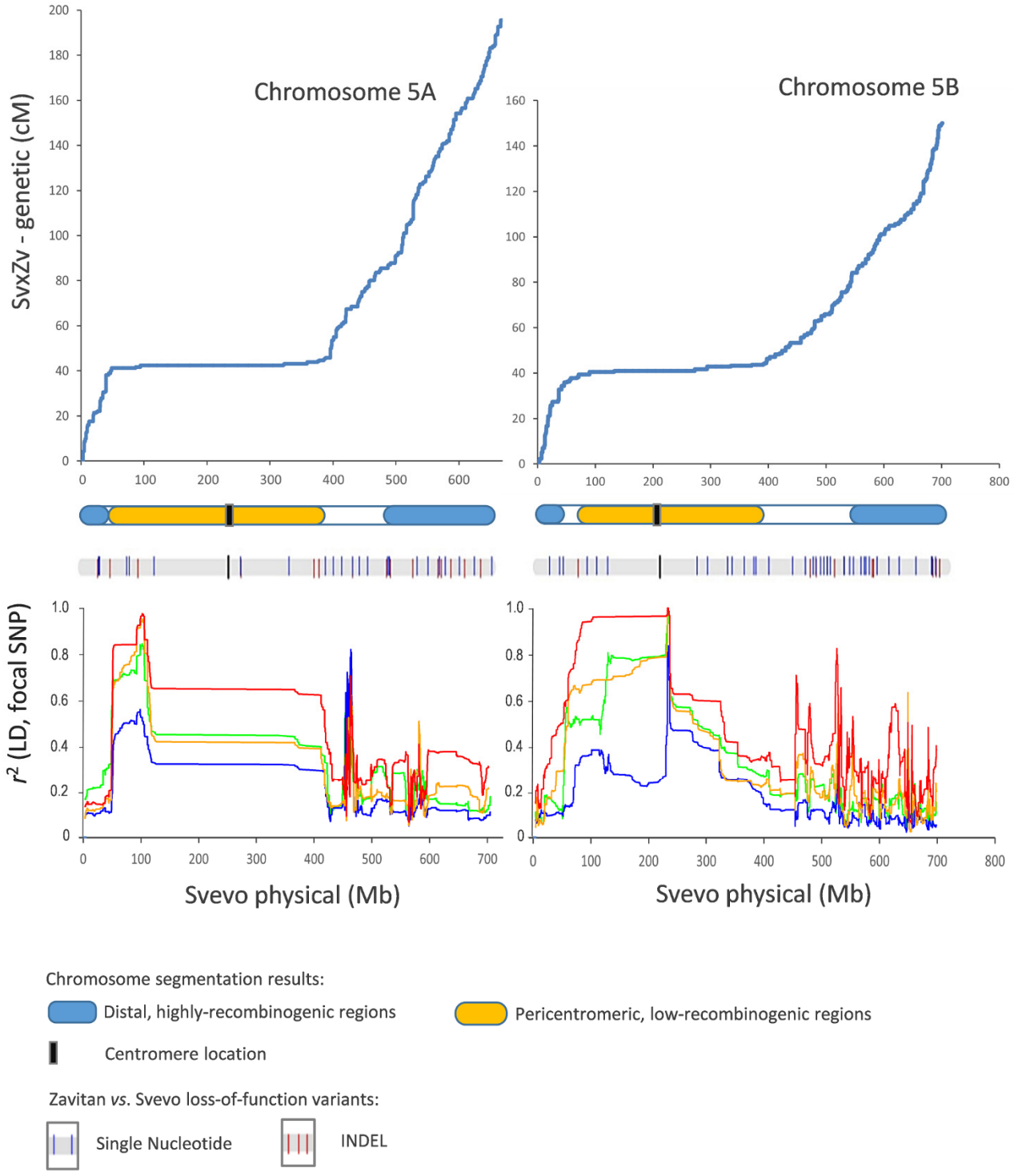
Supplementary Figure 1. Continued.



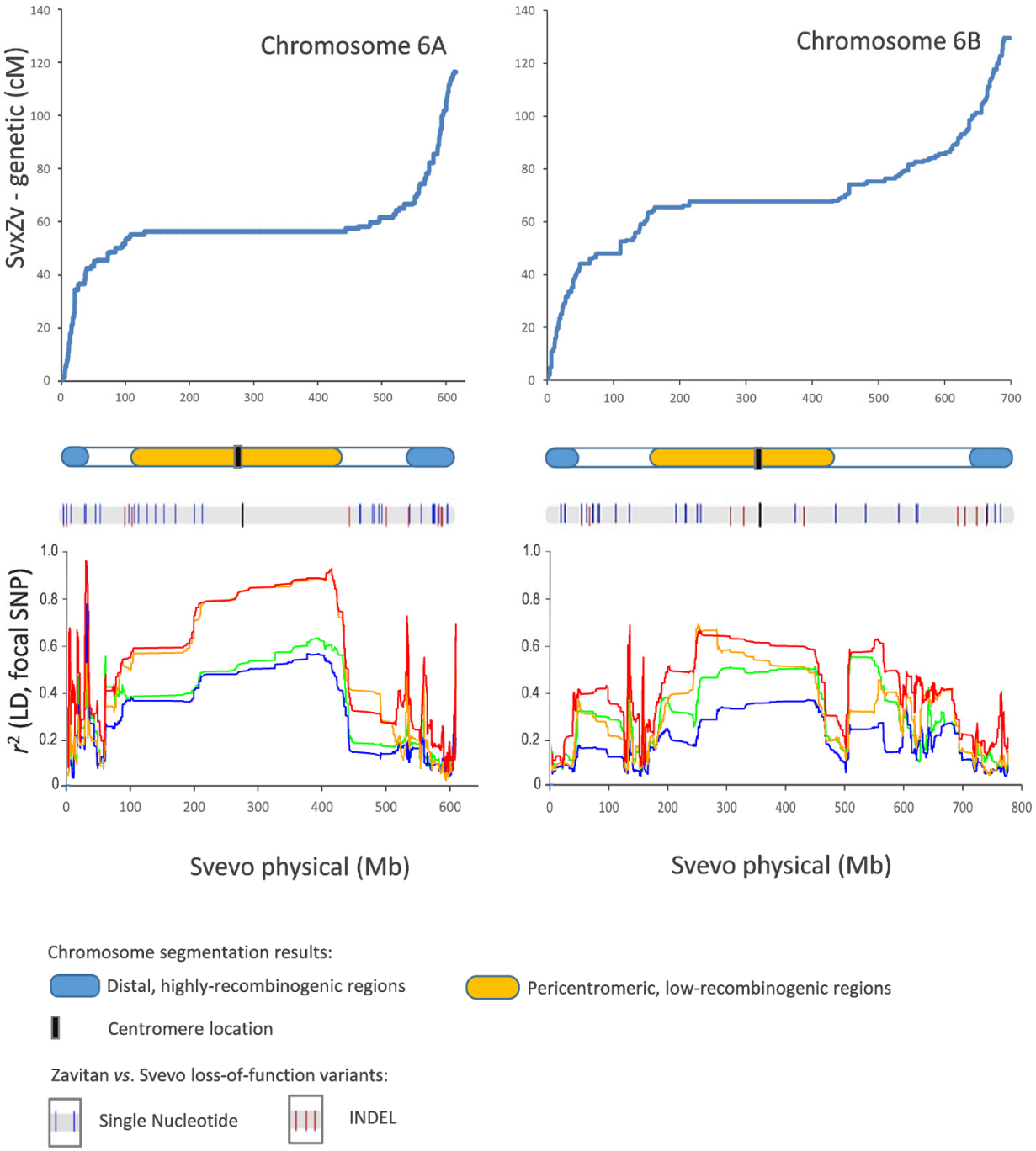
Supplementary Figure 1. Continued.



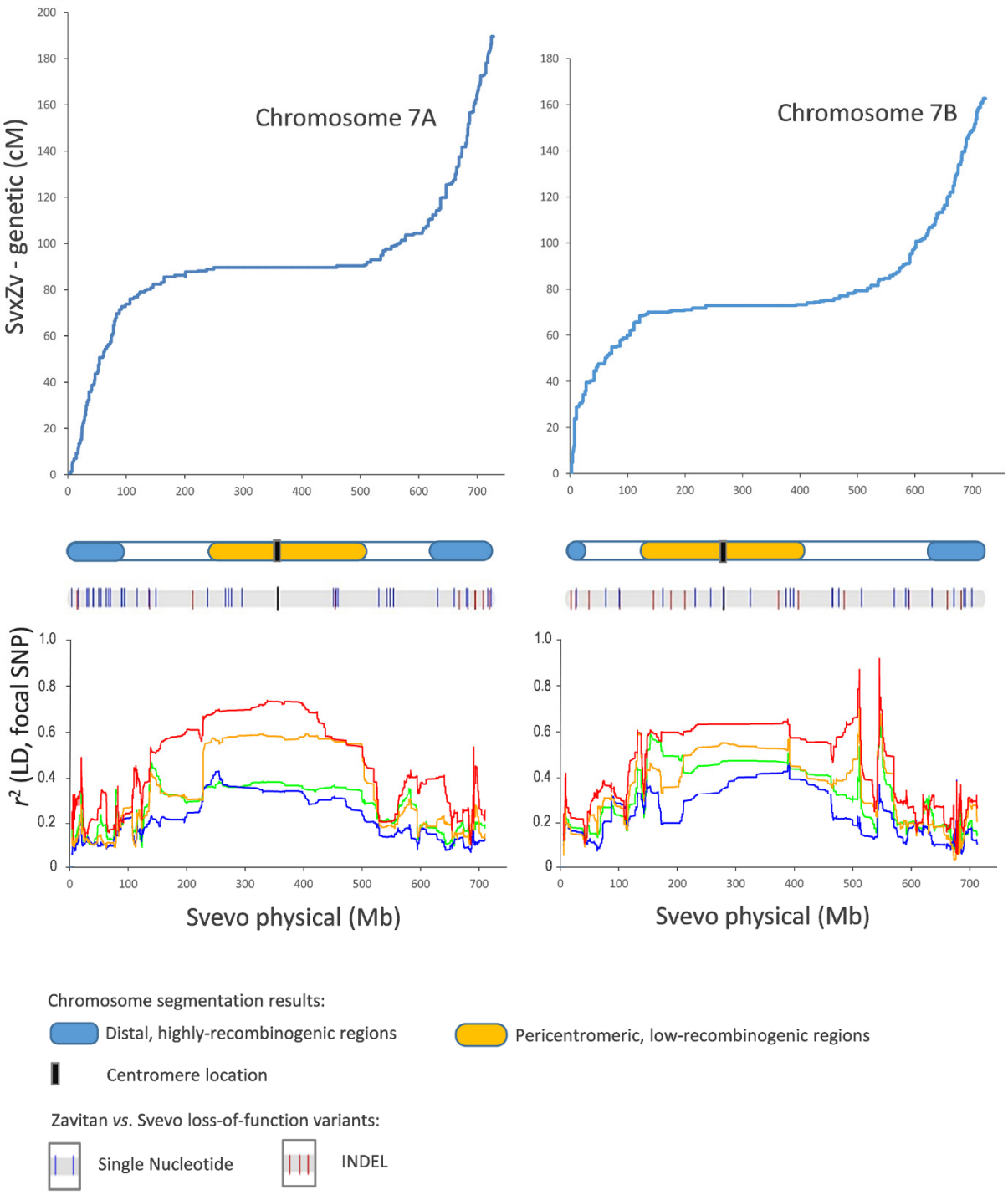
Supplementary Figure 1. Continued.



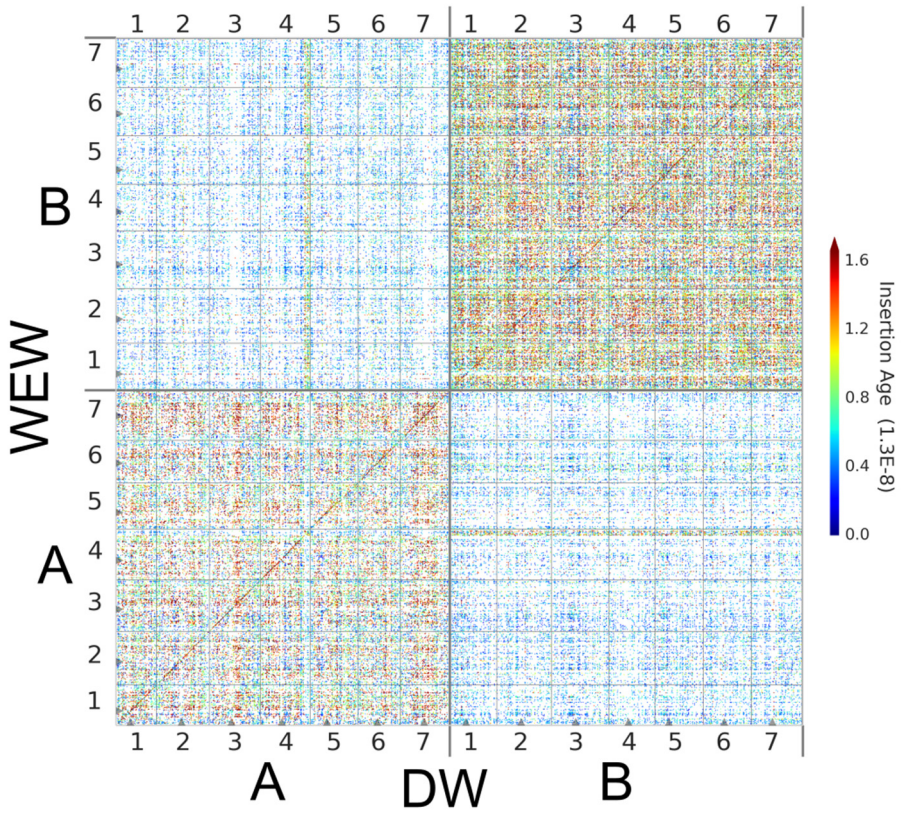
Supplementary Figure 1. Continued.



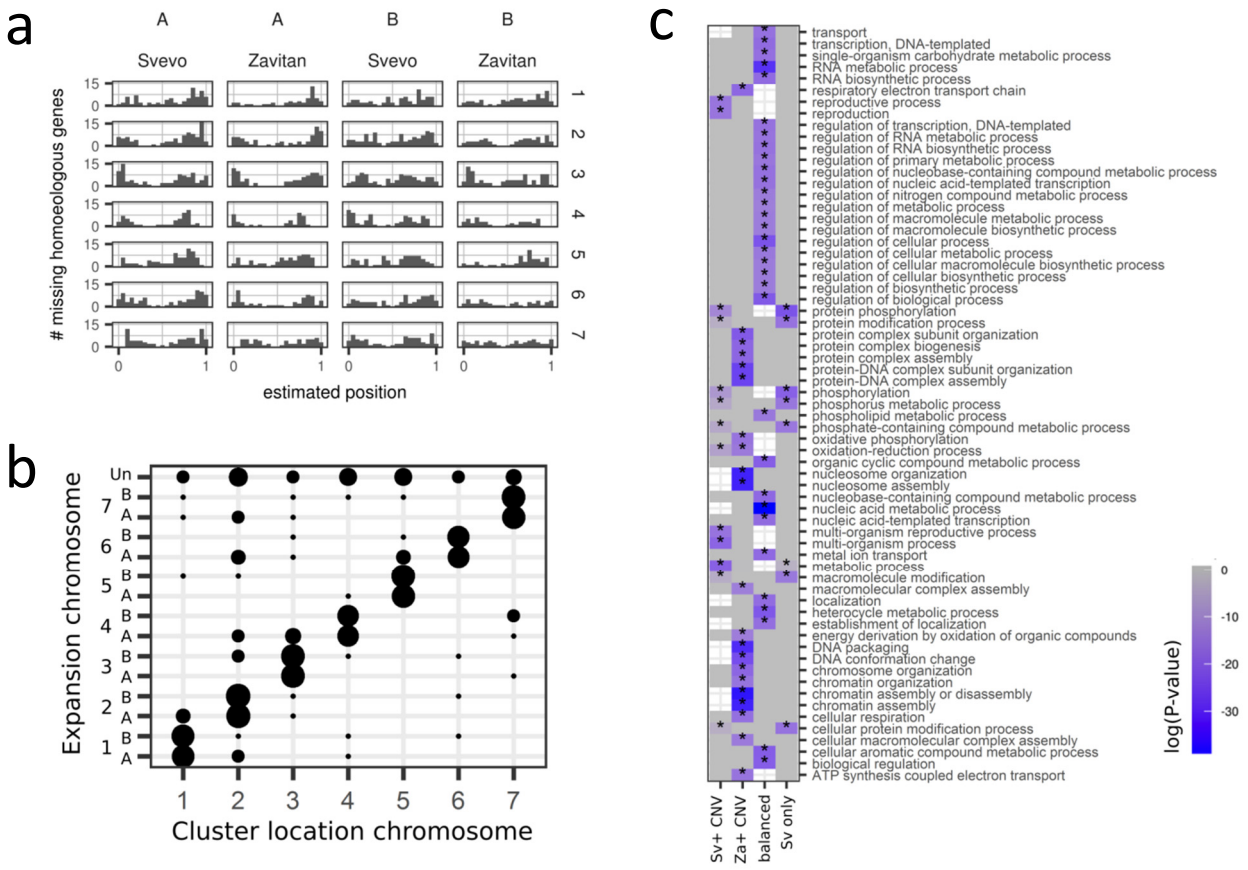
Supplementary Figure 1. Continued.



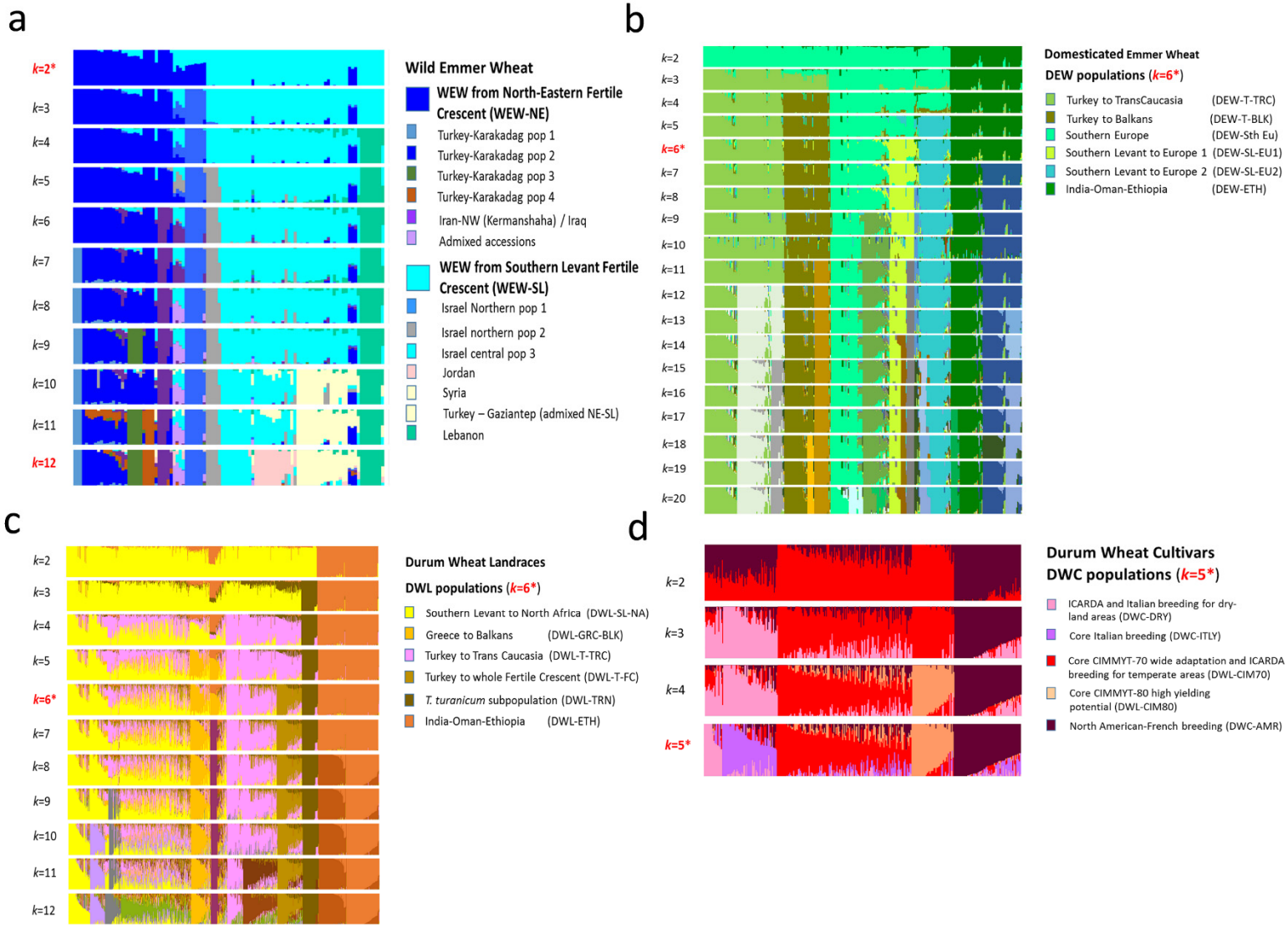
Supplementary Figure 1. Continued.



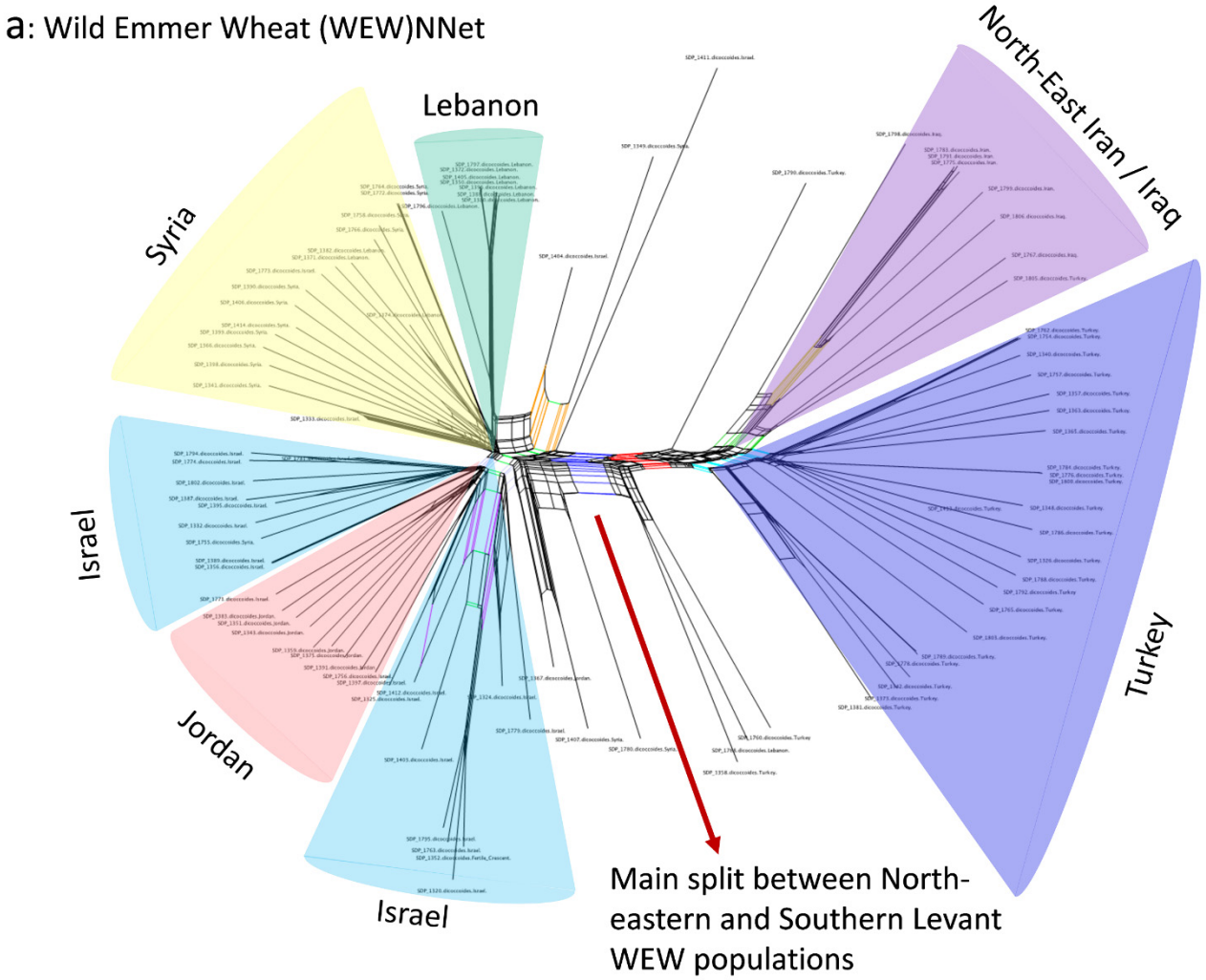
Supplementary Figure 2. Genome wide comparison of DW and WEW full length LTR-retrotransposon (fl-LTR) subfamilies. The dotplot depicts the insertion age (colorscale) and location of fl-LTR subfamilies by pairwise connecting all members of a 90/90 cluster with each other. The clearly visible diagonal represents about 8,000 fl-LTRs with still conserved syntenic positions between DW and WEW. Elements occurring only in the B genome are distinctly younger than those restricted to the A genome. The known translocation from 7B to 4AL stands out by its B genome typic background (older fl-LTRs).



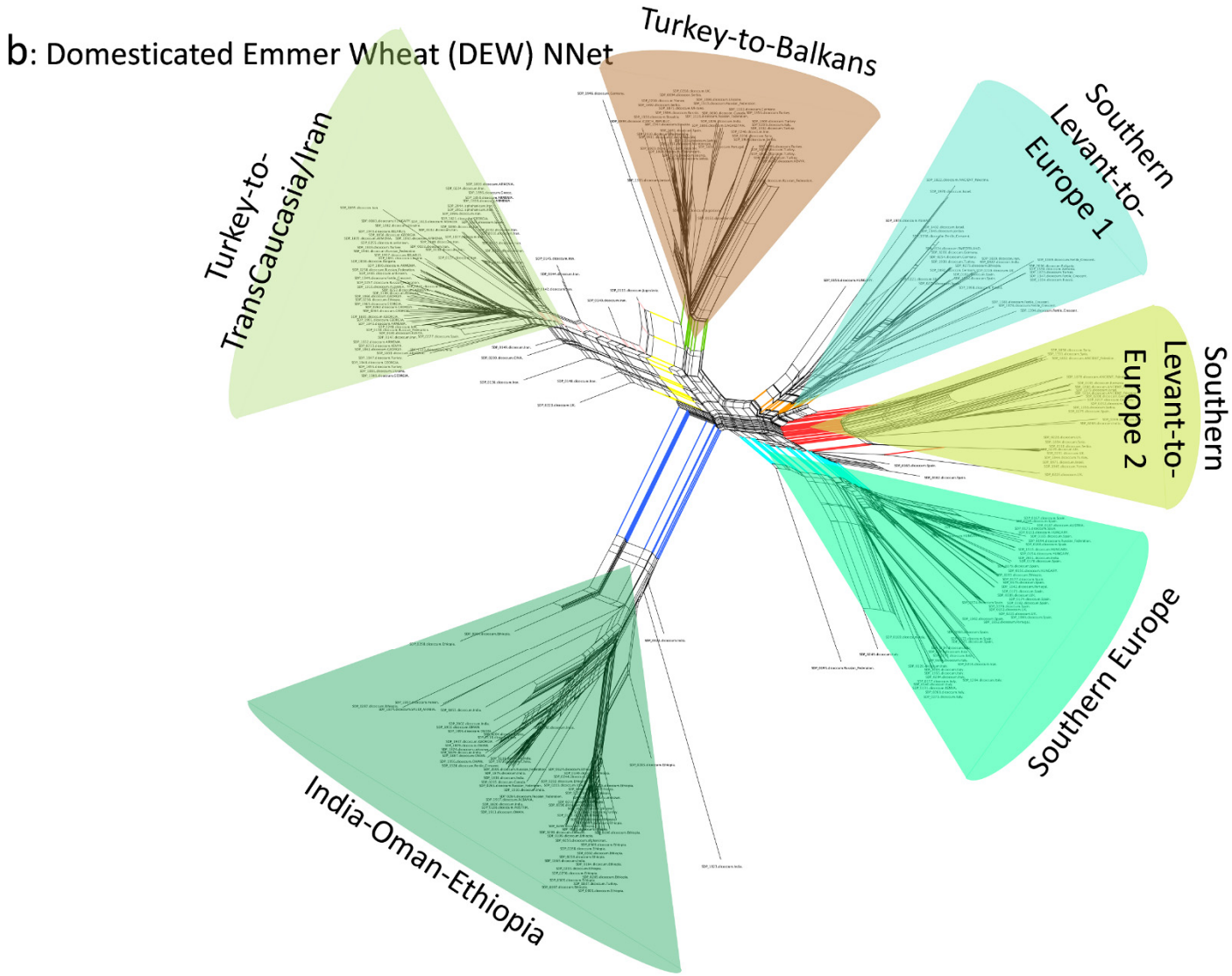
Supplementary Figure 3. Characterization of Svevo and Zavitan HC gene clusters. **a,** Distribution of lost copies along chromosomes. Apparent losses were defined as unigene clusters with 3 members and one missing subgenome member. **b,** Distribution of gained copies from their origin (x axis) to their new location (y-axis). Apparent gains were defined as additions to a four-member cluster with one gene from each subgenome. **c,** Gene ontology (GO) enrichment analyses for Svevo genes belonging to different types of unigene clusters. Sv+: more copies in Svevo, Za+: more copies in Zavitan, balanced: identical copy numbers for Svevo and Zavitan, Sv only: Svevo HC genes without close homeolog in the Zavitan HC gene set. The level of significance (p-value) is depicted by the heatmap color.



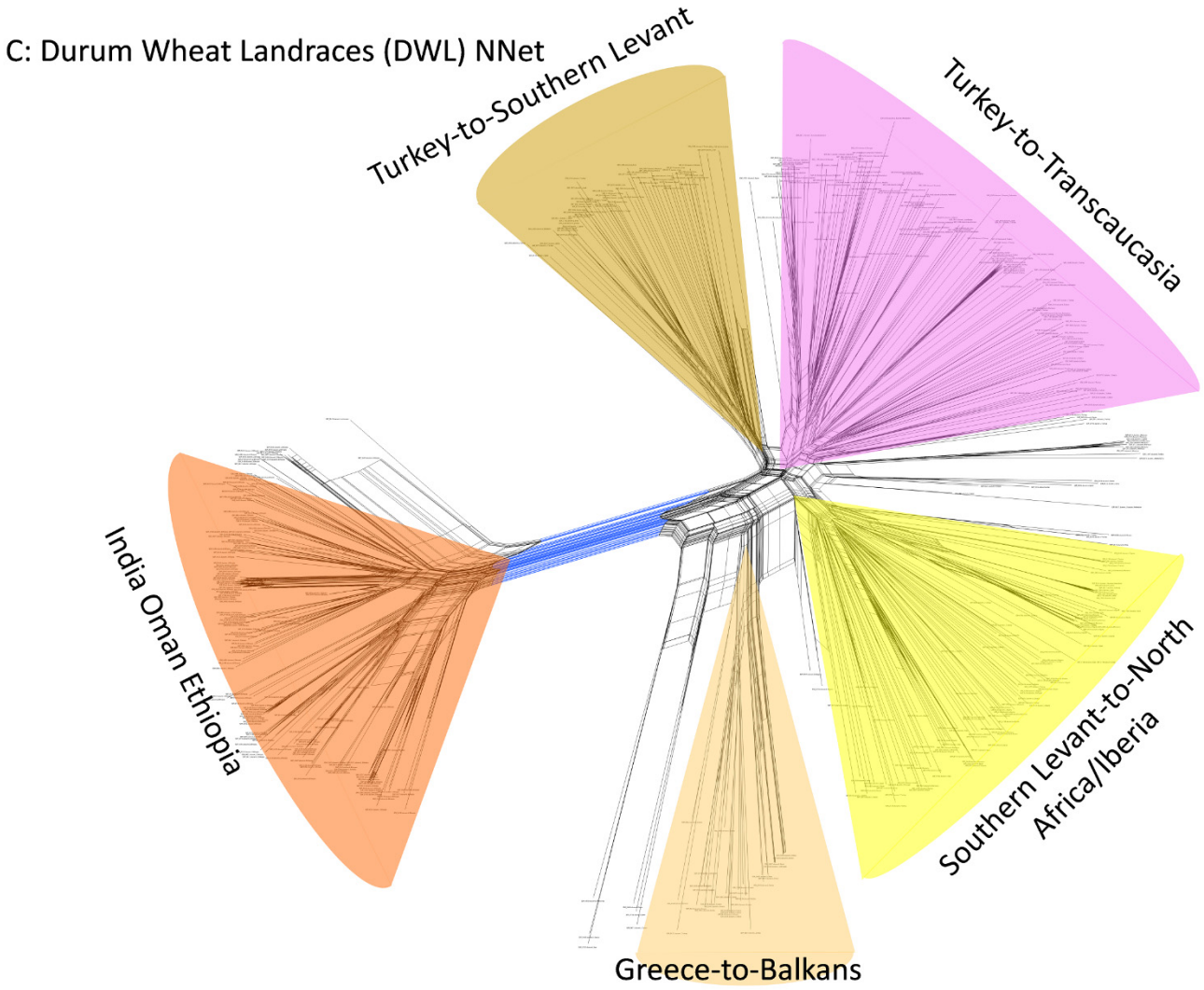
Supplementary Figure 4. *ADMIXTURE* analysis results of wild emmer wheat (WEW) accessions with *K* from 2 to 12 **(a)**, domesticated emmer wheat accessions with *K* from 2 to 20 **(b)**, durum wheat landrace accessions with *K* from 2 to 12 **(c)**, and durum wheat cultivars with *K* from 2 to 5 **(d)**. Results are represented as bar plots of *Q* membership coefficients.



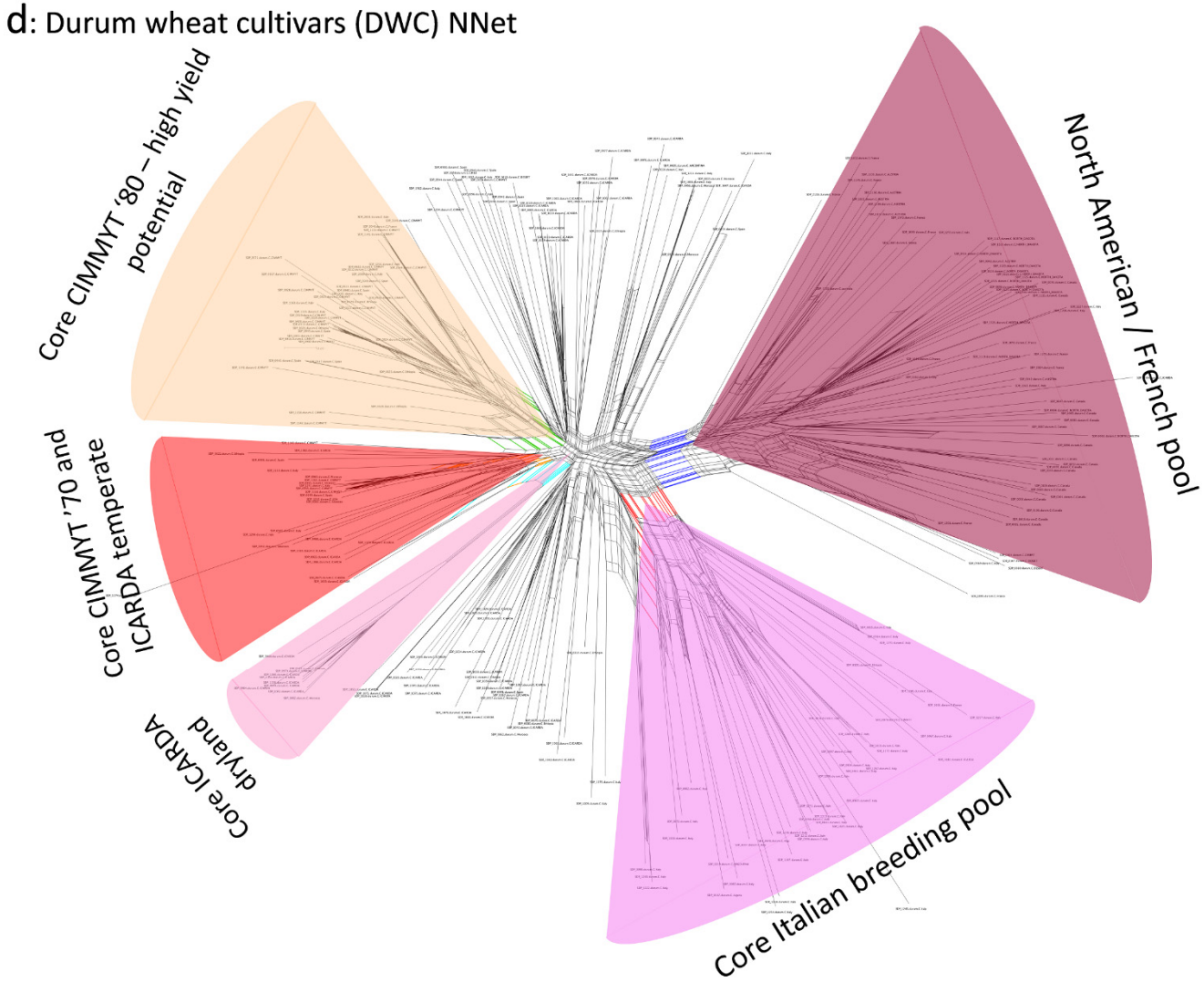
Supplementary Figure 5. NeighborNets of Hamming distances based on 5,775 iSelect 90K wheat SNP polymorphisms for the tetraploid wheat taxa and populations included in this study. **a,** wild ememr wheat, WEW; **b,** domesticated emmer wheat, DEW; **c,** durum wheat landraces, DWL; **d,** durum wheat cultivars, DWC; **e,** combined WEW-DEW; **f,** combined WEW-DEW-DWL. Main populations are indicated in Figure by color-shaded triangles, while major splits are highlighted by colored lines.



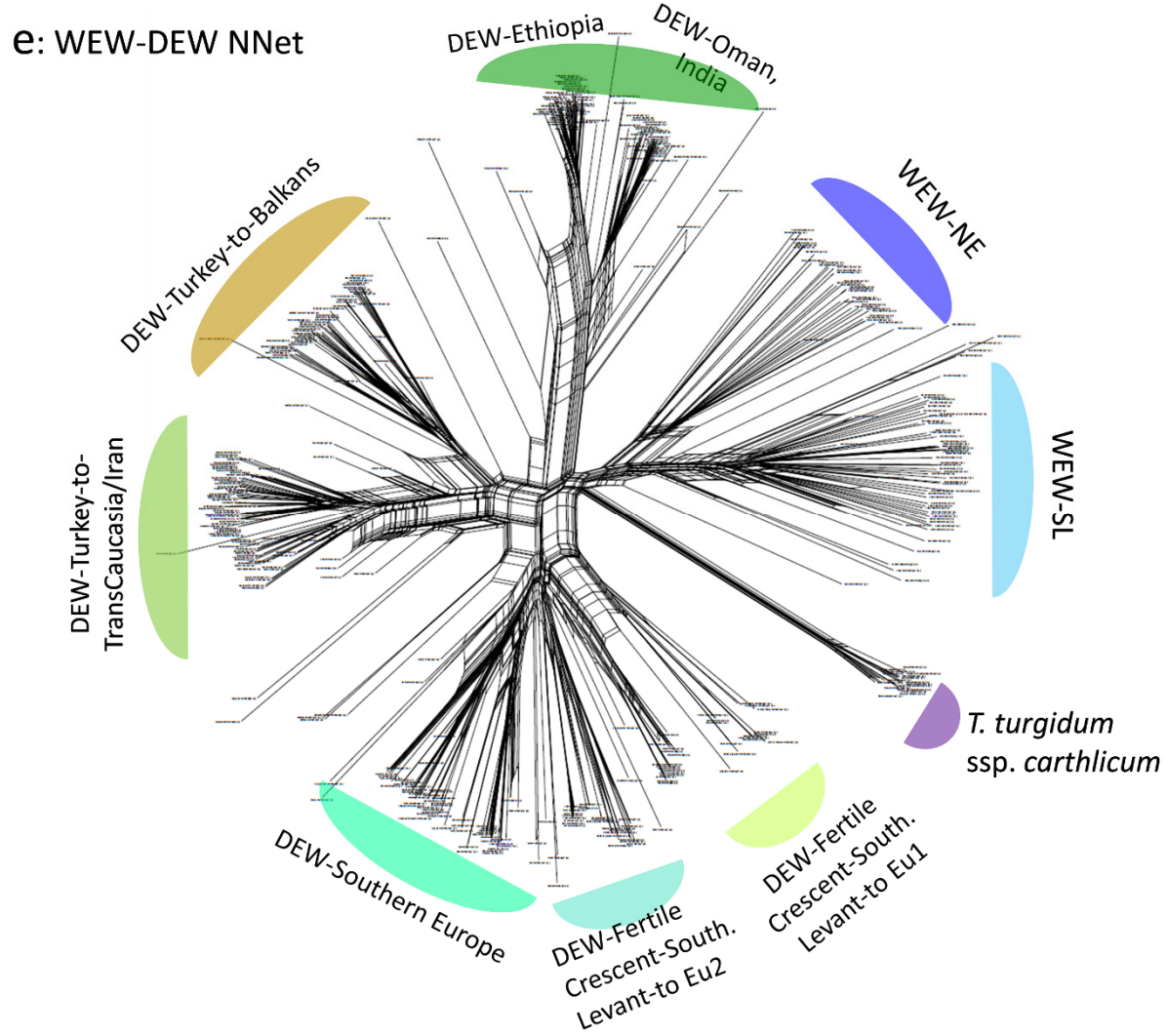
Supplementary Figure 5. Continued.



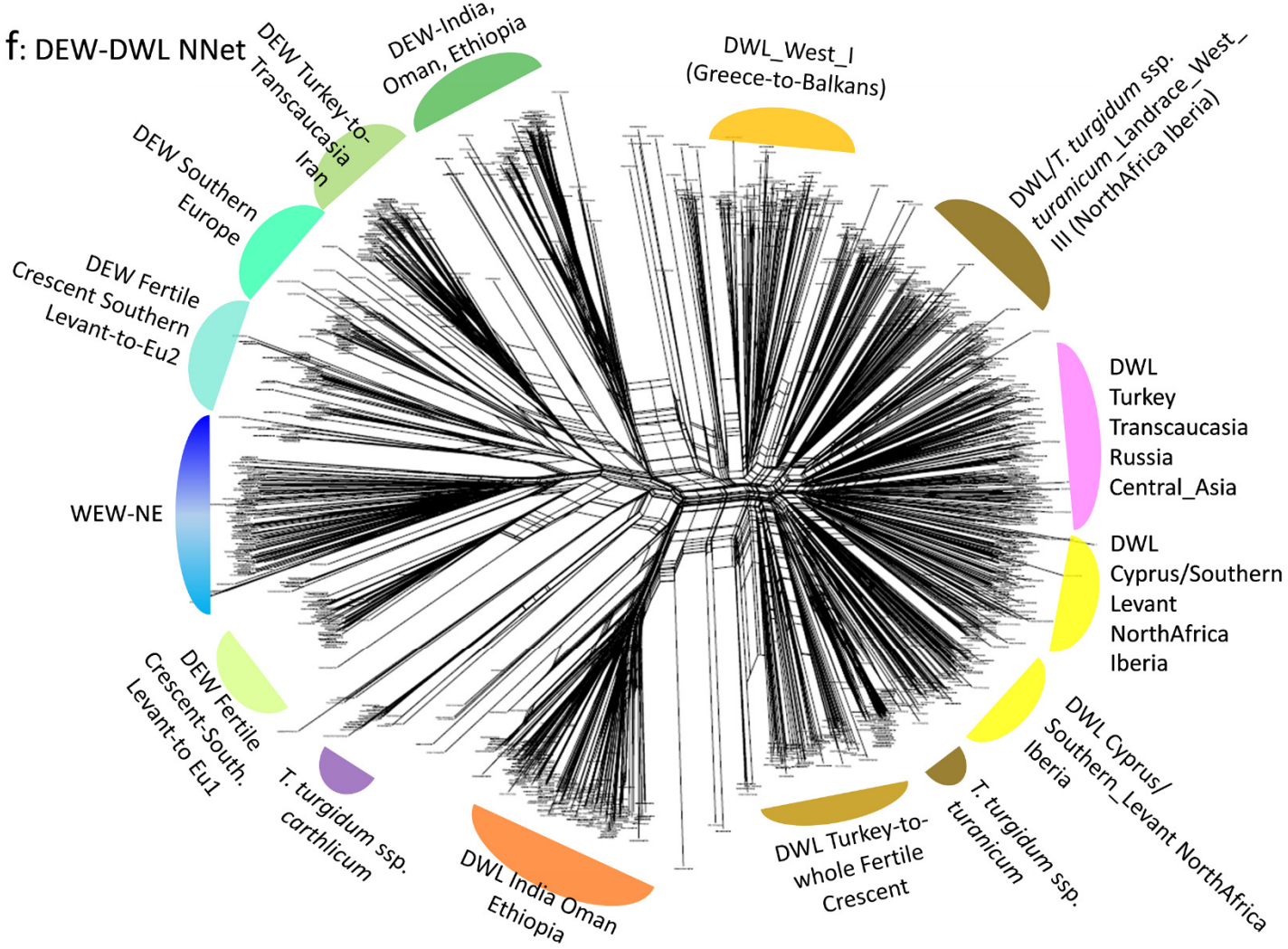
Supplementary Figure 5. Continued.



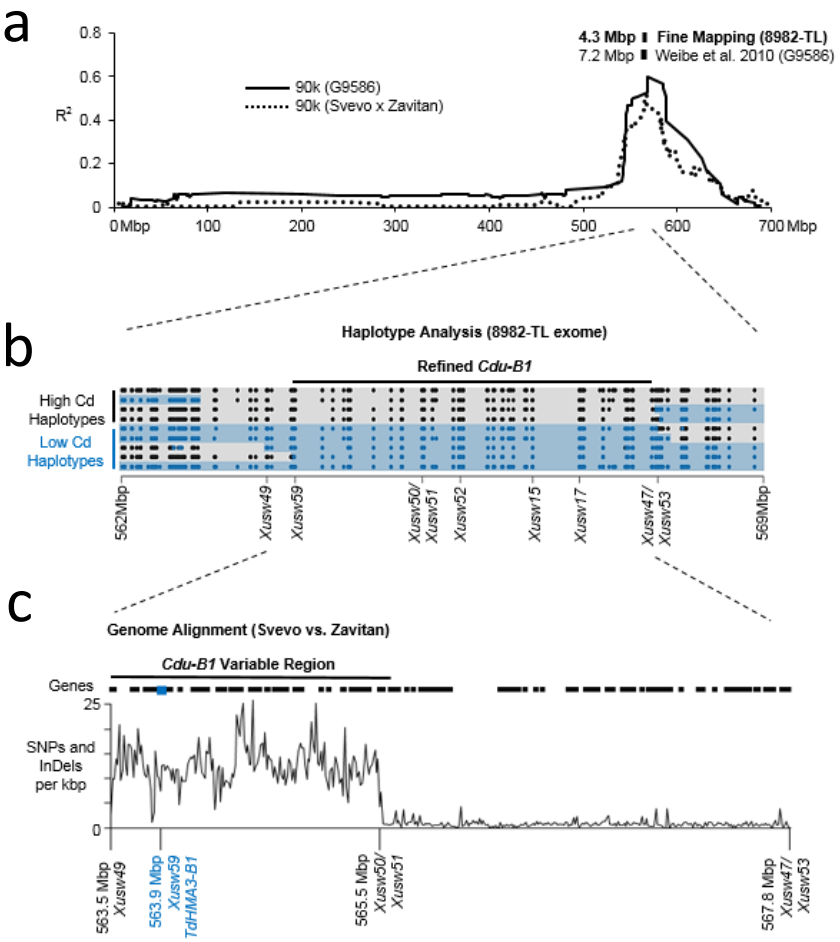
Supplementary Figure 5. Continued.



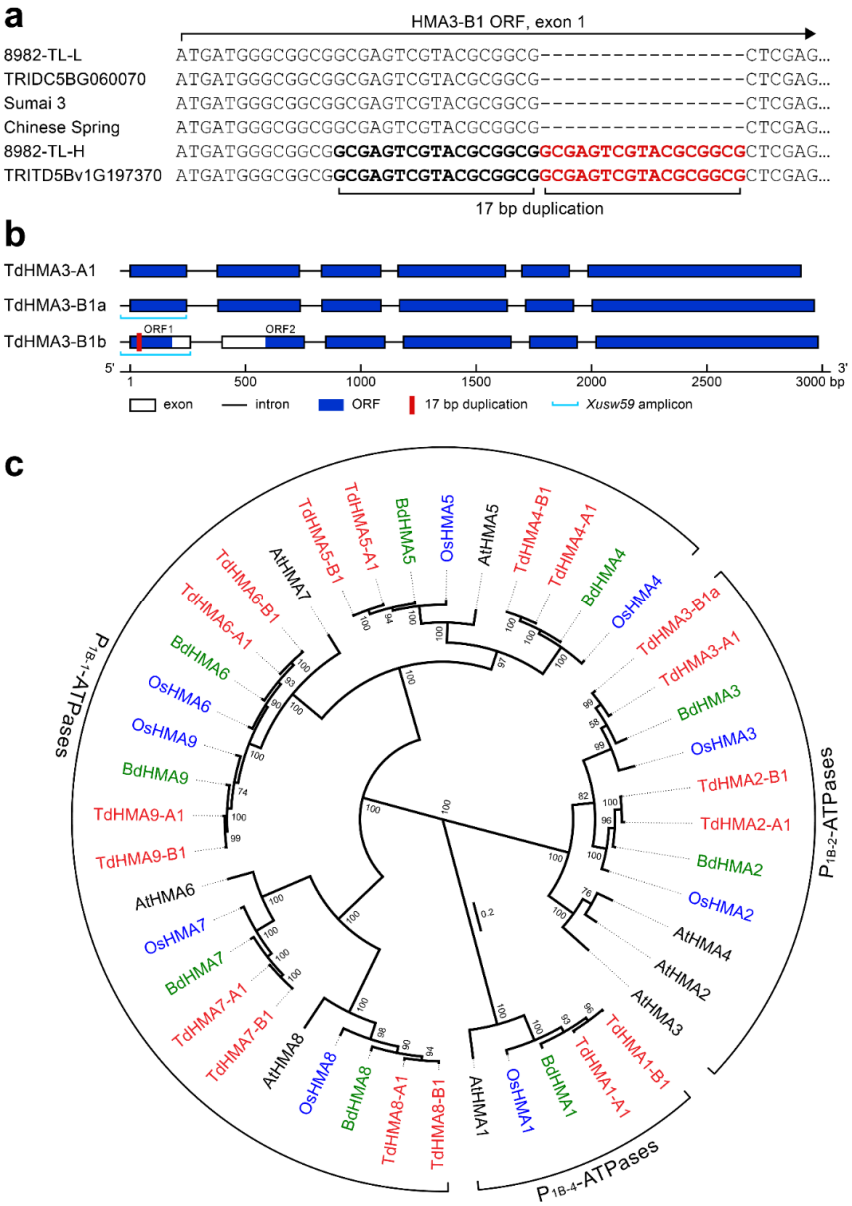
Supplementary Figure 5. Continued.



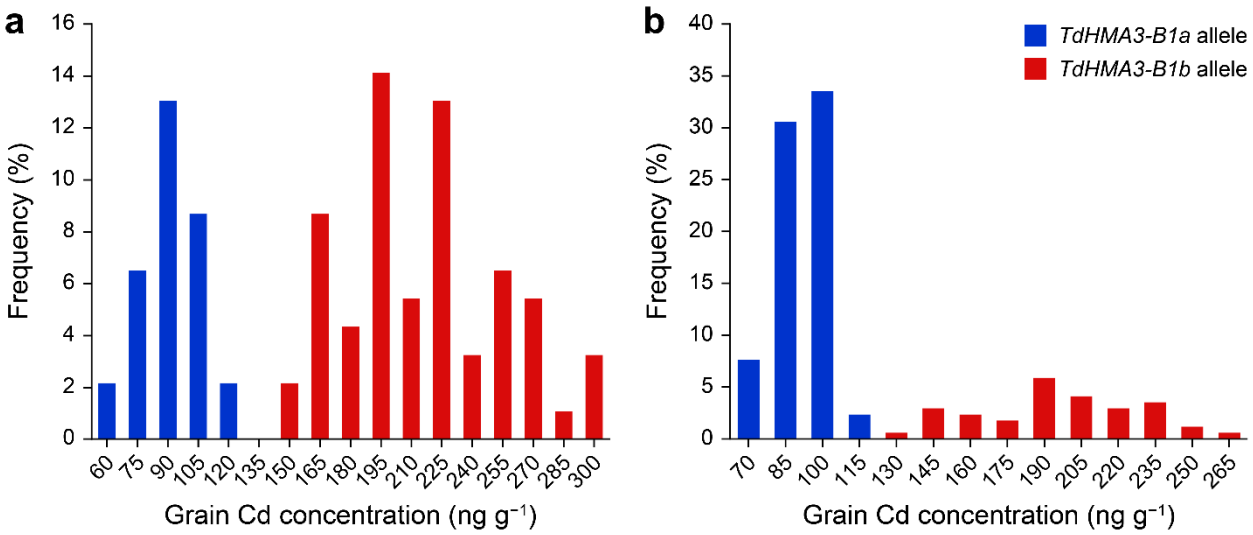
Supplementary Figure 5. Continued.



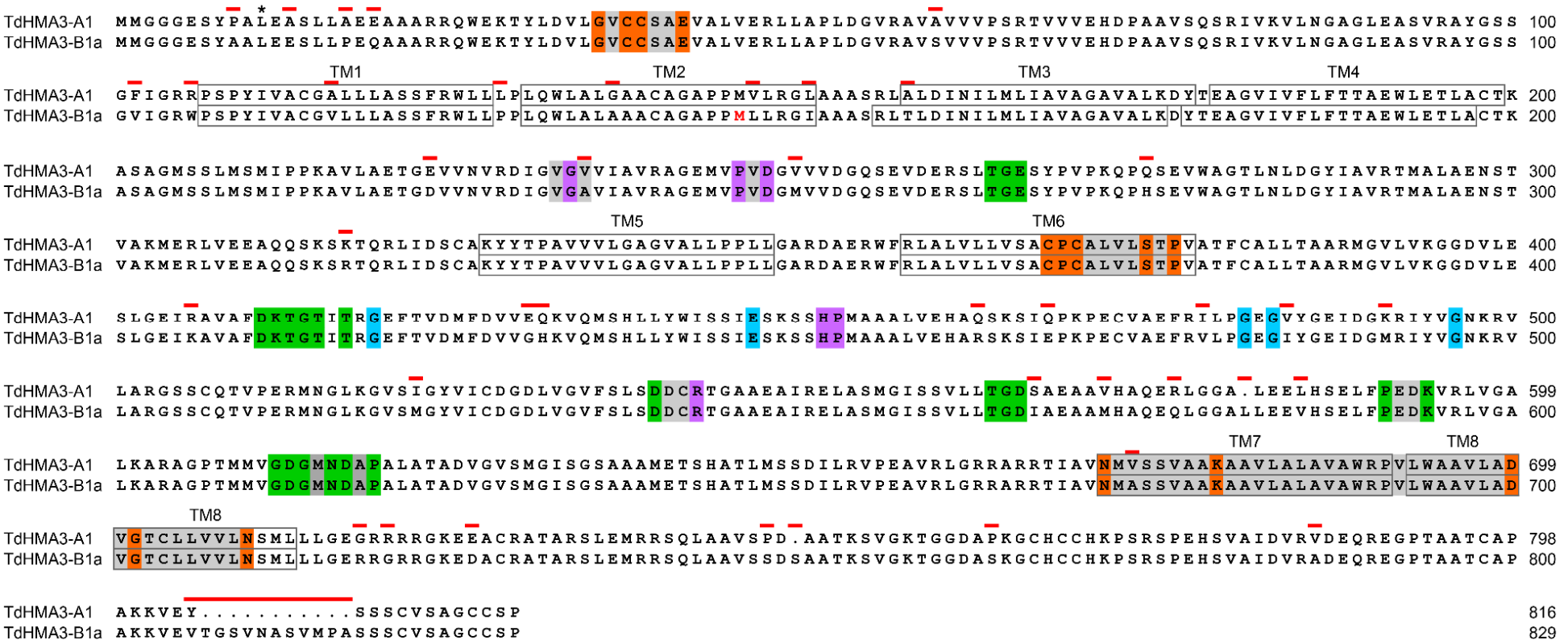
Supplementary Figure 6. Use of the Svevo genome to identify a gene candidate associated with differences in Cd accumulation in the grain of durum wheat. **a,** The *Cdu-B1* locus was mapped in three bi-parental populations (G9586, 8982-TL and Svevo × Zavitan) on chromosome 5B. **b,** High-throughput sequencing of exomes refined the *Cdu-B1* locus by identifying additional polymorphism associated with low (blue) and high (black) Cd accumulation; positions of markers from the original mapping populations are indicated. **c,** Alignment of *Cdu-B1* between Svevo and Zavitan identified a region of increased sequence variation. This region contains a variable gene that encodes a metal transporter, *TdHMA3-B1* (blue).



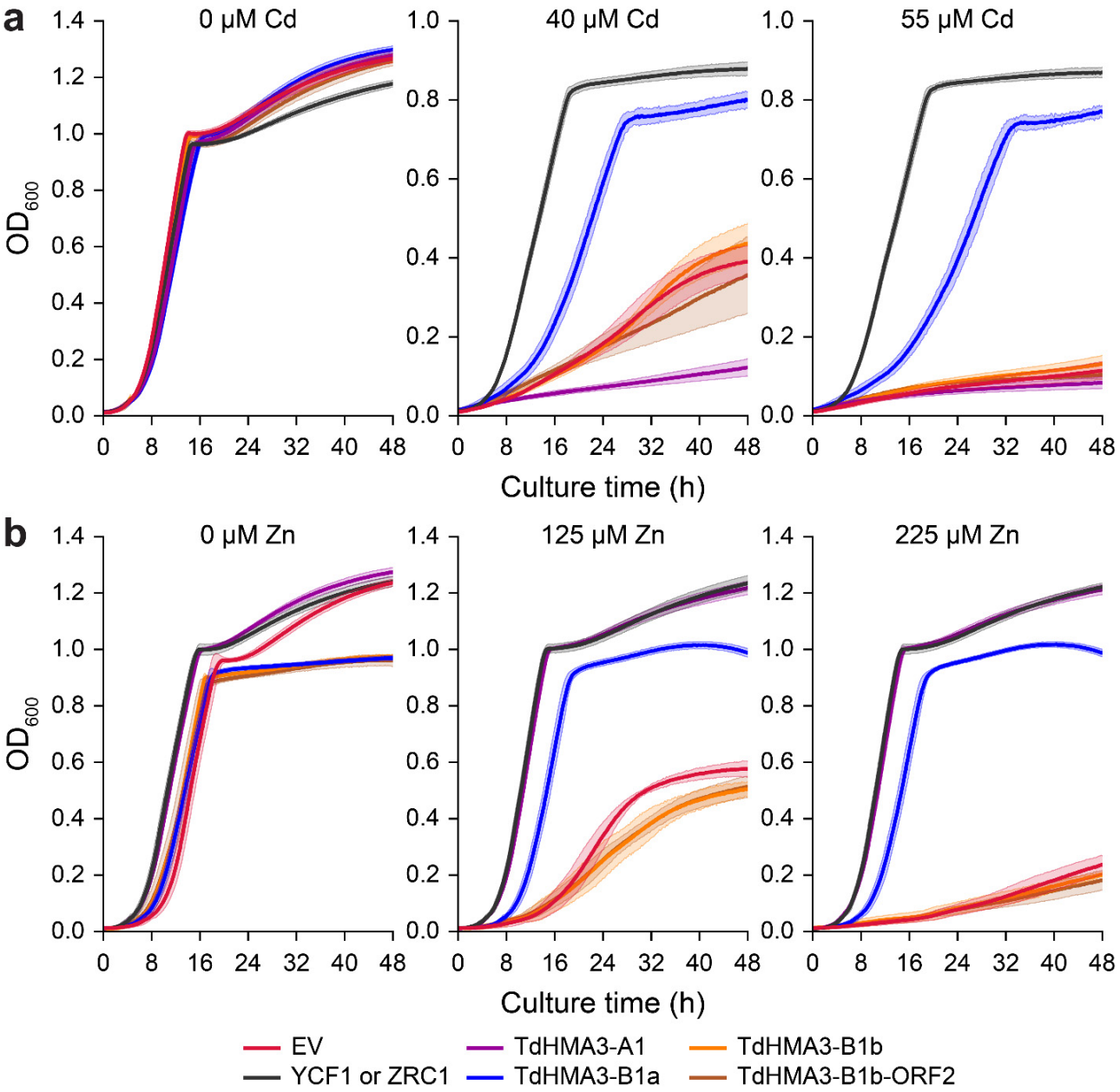
Supplementary Figure 7: Structural and phylogenetic characterization DW *HMA3* P_{1B-2}-ATPases linked to *Cdu-B1*. **a**, Multiple sequence alignment of exon 1 of *TdHMA-B1* from low Cd (8982-TL-L (KF683294), WEW Zavitan TRIDC5BG060070, Sumai 3 (exome sequencing), and Chinese Spring (TGAC v1.0 Scaffold_404346_5BL)) and high Cd (8982-TL-H (KF683295), Svevo TRITD5Bv1G197370) genotypes identifies a 17-bp duplicated insertion (shown in red) in the high Cd genotypes. **b**, Gene structure of *TdHMA3-A1*, *TdHMA3-B1a*, and *TdHMA3-B1b*. The 17 bp duplication in allele *TdHMA3-B1b* results in a frame-shift and premature stop codon in the high Cd genotypes (KF683295:85-267, ORF1, 183 bp). The location of the *Xusw59* amplicon that captures the 17-bp duplication is indicated in light blue. The longest alternative ORF for allele *TdHMA3-B1b* is truncated at the 5' by 449 bp (ORF2: KF683295:534-2591, 2058 bp). **c**, Maximum-likelihood consensus phylogram of P_{1B}-ATPase homologs from *Arabidopsis thaliana* (At, black), *Oryza sativa* (Os, blue), *Brachypodium distachyon* (Bd, green), and *Triticum turgidium* subsp. *durum* (Td, red). P_{1B}-ATPase subgroups[161] are indicated at the periphery. Bootstrap support (%) is indicated at the nodes. The scale shows the estimated branch length measured in the number of substitutions per site. Protein sequences and locus identifiers are provided in Supplementary Data Set 12.



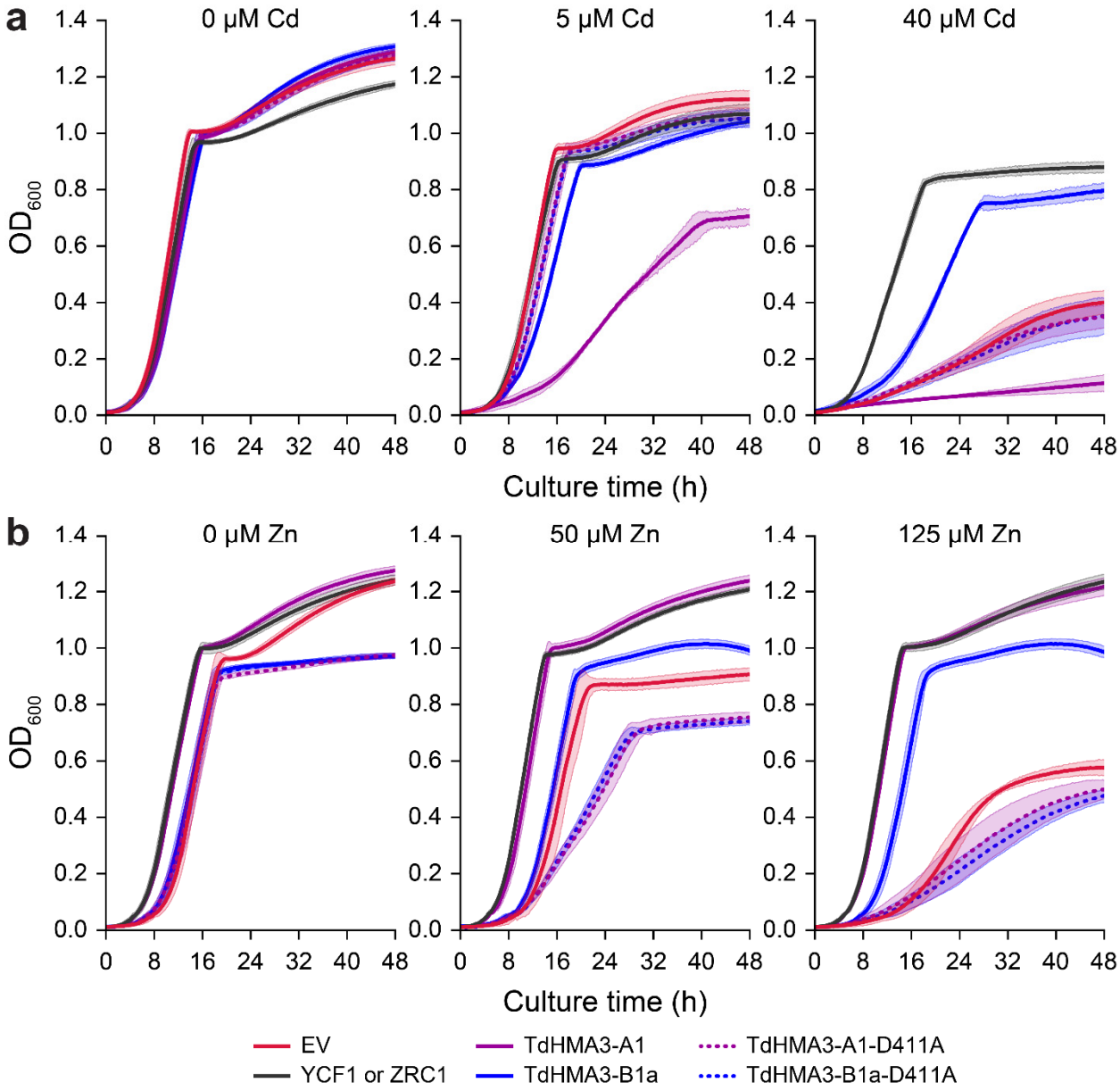
Supplementary Figure 8: *TdHMA3-B1a* and *TdHMA3-B1b* alleles discriminate between diverse low- and high-Cd accumulating genotypes in field trials. a, Frequency distribution of grain Cd concentration in a diversity panel representing 94 cultivars and breeding lines collected from global breeding programs. **b**, Frequency distribution of grain Cd concentration in 174 cultivars and advanced breeding lines developed by Canadian durum wheat breeding programs. Lines carrying the *TdHMA3-B1a* allele are highlighted in blue and those carrying the *TdHMA3-B1b* allele are in red.



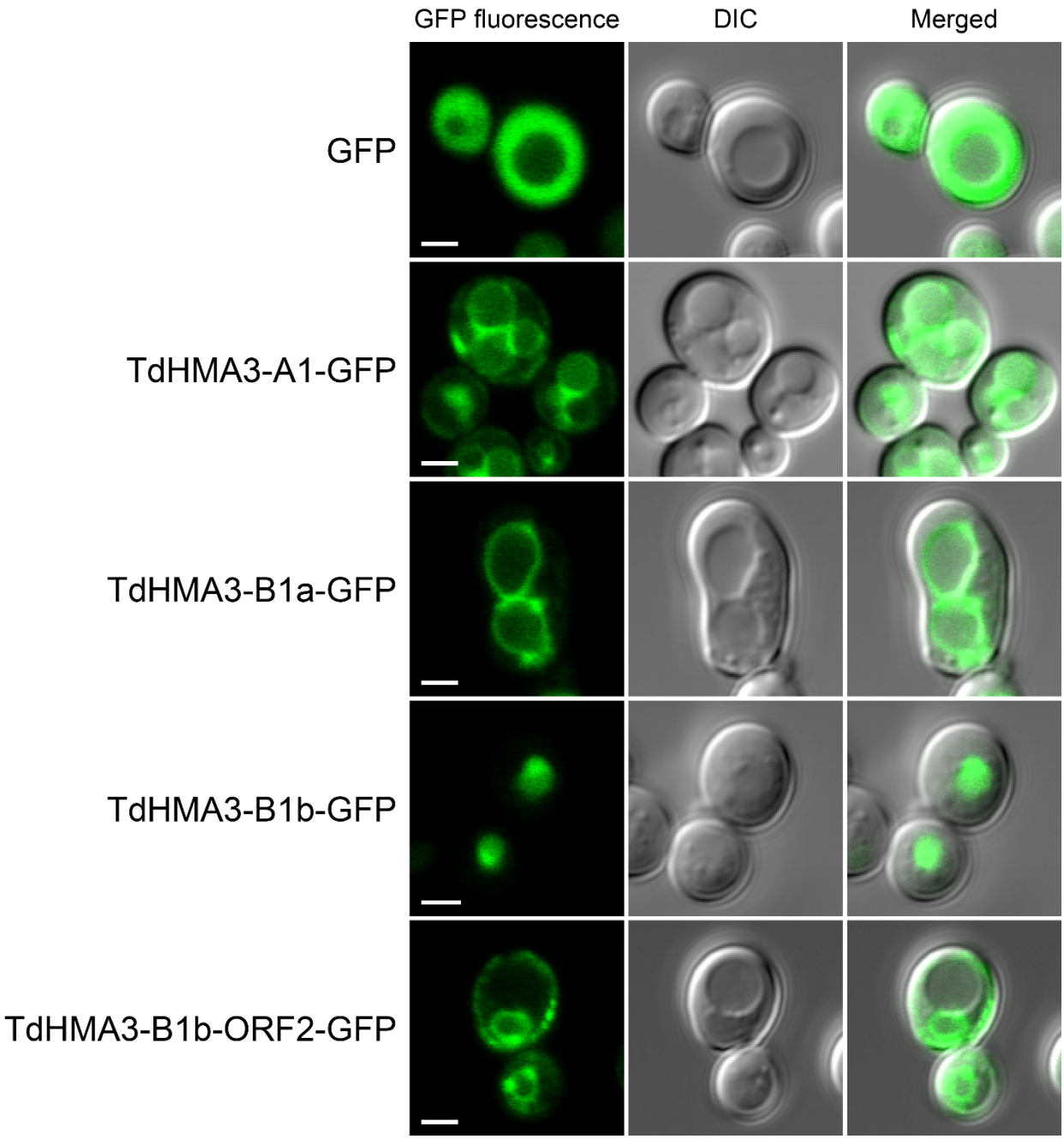
Supplementary Figure 9: Structural characterization of TdHMA3-A1 (AIA57676) and TdHMA3-B1a (AIA57679) P_{1B-2}-ATPases. Transmembrane topology (TM1-8, outlined with boxes) was determined by consensus predictions using TOPCONS (http://topcons.cbr.su.se/). Non-conserved amino acids are indicated above the alignment with a red line. Sequence motifs conserved amongst P-ATPases are highlighted in green, motifs conserved amongst P_{1B}-ATPases are highlighted in blue, motifs conserved amongst P_{1B-2}-ATPases are highlighted in orange, and common sequences not strictly conserved amongst any group are highlighted in purple. Ambiguities within the conserved motifs are highlighted in grey. The asterisk (*) at position 11 indicates the location of the duplication that results in a frameshift in *TdHMA3-B1b*. Met-145 of TdHMA3-B1a in TM2 is highlighted in red to indicate the translation start site of the largest alternative open reading frame after the 17 bp duplication in *TdHMA3-B1b* (ORF2).



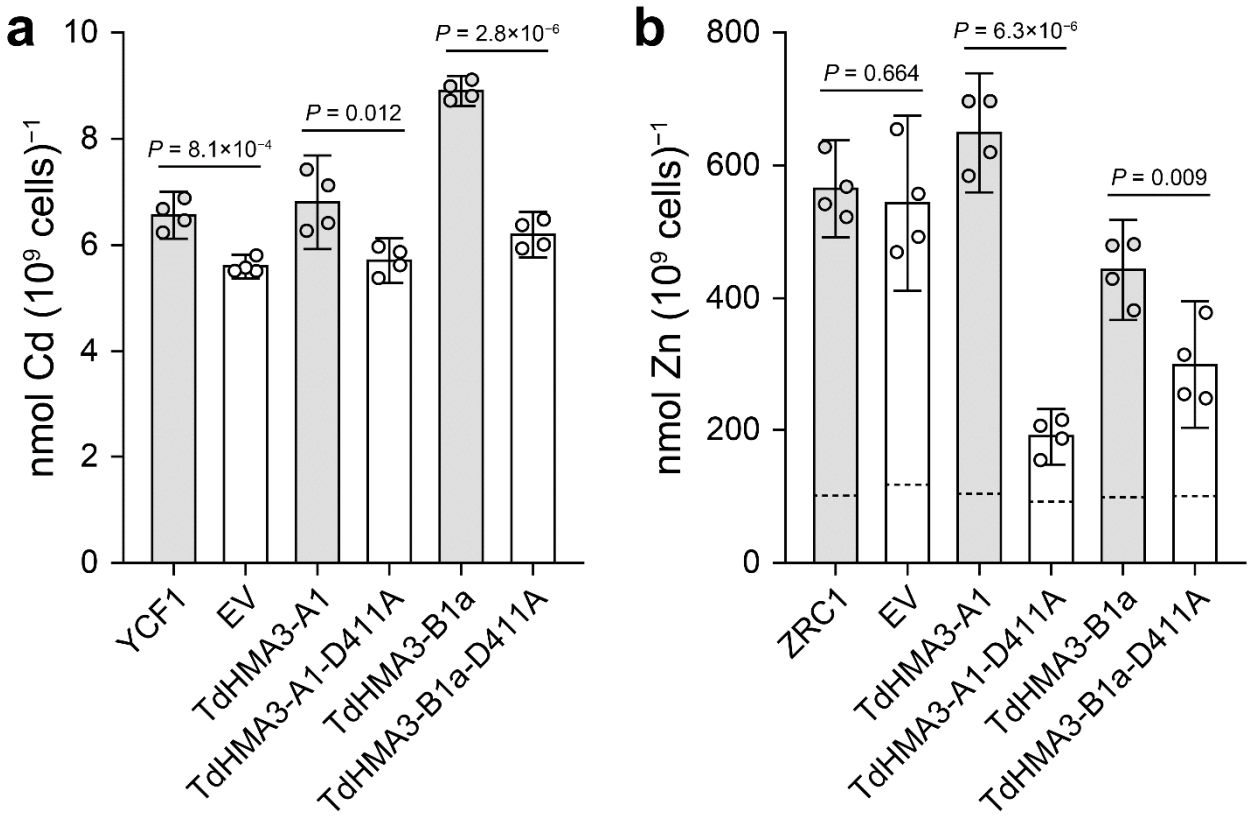
Supplementary Fig. 10: *TdHMA3-B1b* fails to complement Cd- and Zn-sensitive mutant phenotypes of yeast. a, Growth (OD_{600}) of Cd-sensitive *ycf1* yeast expressing empty vector (EV, p413TEF), *YCF1*, *TdHMA3-A1*, *TdHMA3-B1a*, *TdHMA3-B1b*, and *TdHMA3-B1b*-ORF2 in the presence of 0, 40, and 55 µM Cd. Plotted growth curves are means of 4 (ORF2), 5 (*YCF1*, *B1a*, *B1b*), or 6 (EV, *A1*) experiments ± 95% CIs shown as shaded backgrounds. **b**, Growth (OD_{600}) of Zn-sensitive *zrc1cot1* yeast expressing empty vector (EV, p413TEF), *ZRC1*, *TdHMA3-A1*, *TdHMA3-B1a*, *TdHMA3-B1b*, and *TdHMA3-B1b*-ORF2 in the presence of 0, 125, and 225 µM Zn. Plotted growth curves are means of 3 (*B1b*), 4 (ORF2), 5 (EV, *ZRC1*, *B1a*), or 6 (*A1*) experiments ± 95% CIs shown as shaded backgrounds.



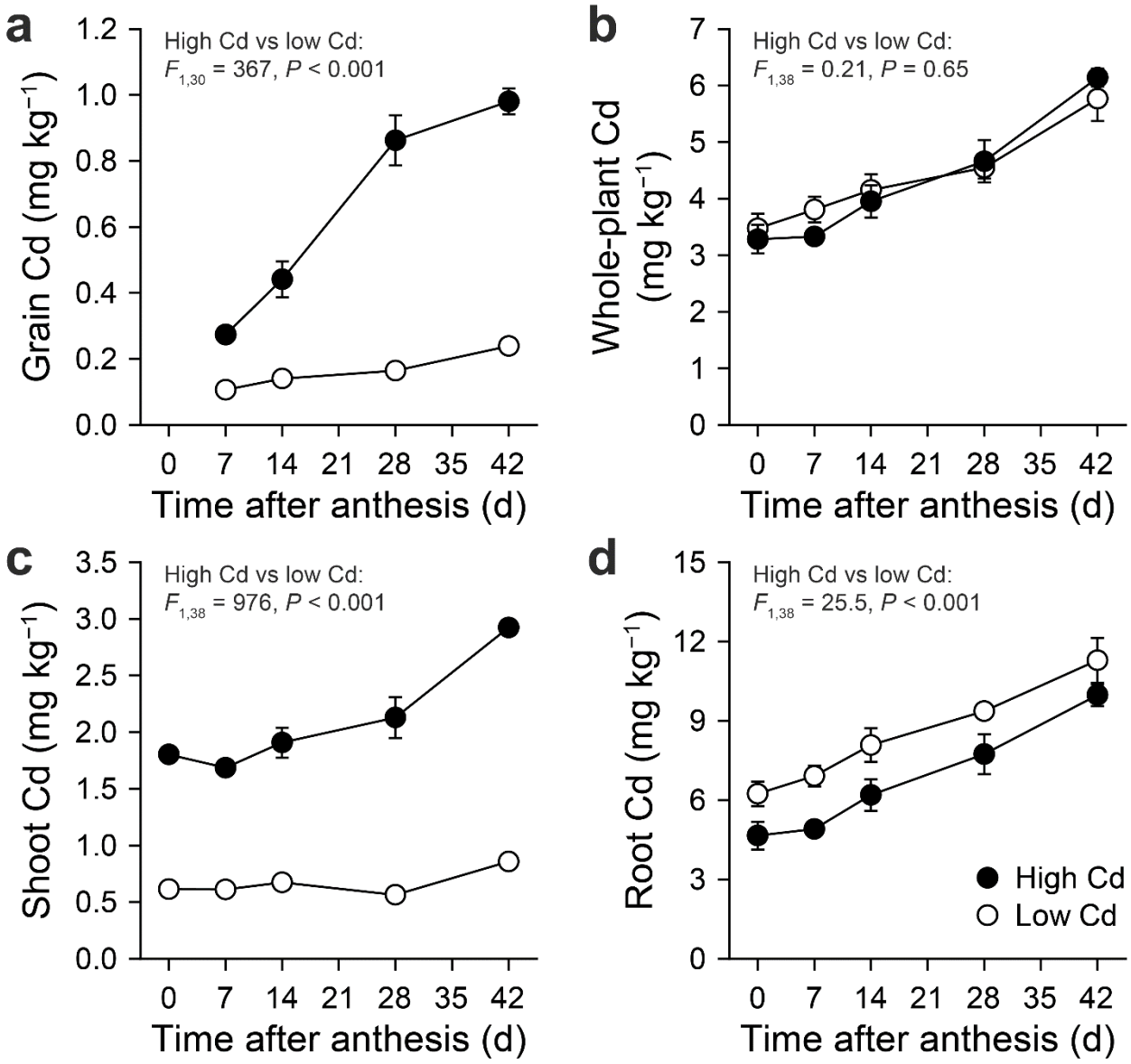
Supplementary Fig. 11: *TdHMA3-B1a*-mediated complementation of Cd- and Zn-sensitive mutant phenotypes of yeast is attributable to P-ATPase ion transport activity. a, Growth (OD_{600}) of Cd-sensitive *ycf1* yeast expressing empty vector (EV, p413TEF), *YCF1*, *TdHMA3-A1*, *TdHMA3-B1a*, *TdHMA3-A1*-D411A, and *TdHMA3-B1a*-D411A in the presence of 0, 5, and 40 µM Cd. Plotted growth curves are means of 3 (*YCF1*), 4 (*A1*, *B1a*, *A1*-D411A, *B1a*-D411A), or 5 (EV) experiments ± 95% CIs shown as shaded backgrounds. **b**, Growth (OD_{600}) of Zn-sensitive *zrc1cot1* yeast expressing empty vector (EV, p413TEF), *ZRC1*, *TdHMA3-A1*, *TdHMA3-B1a*, *TdHMA3-A1*-D411A, and *TdHMA3-B1a*-D411A in the presence of 0, 50, and 125 µM Zn. Plotted growth curves are means of 4 (EV, *A1*, *B1a*) or 5 (*ZRC1*, *A1*-D411A, *B1a*-D411A) experiments ± 95% CIs shown as shaded backgrounds.



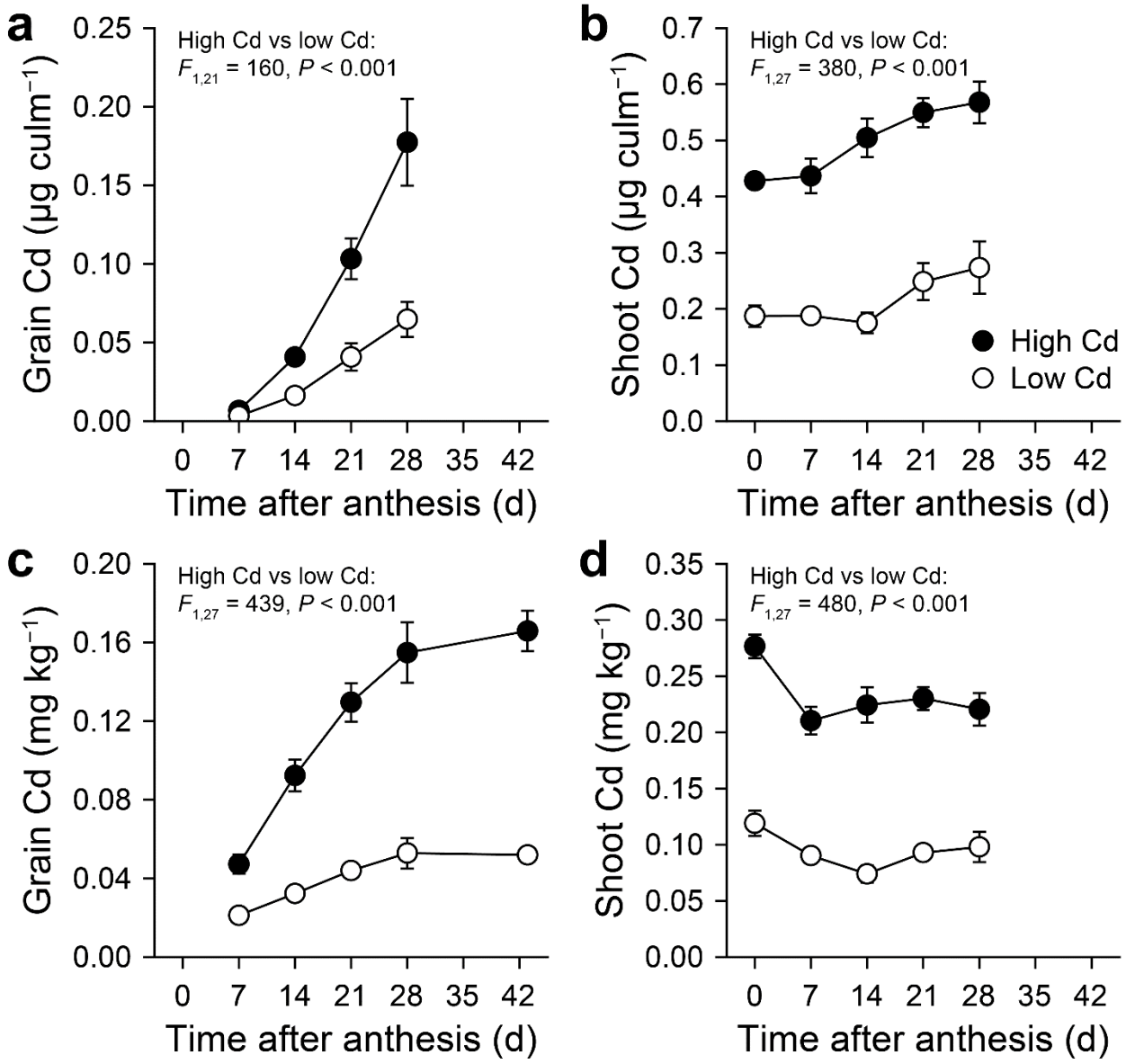
Supplementary Fig. 12: TdHMA3-A1 and TdHMA3-B1a localize to the tonoplast in yeast. Confocal fluorescence of GFP (excitation at 488 nm; detection between 505–530 nm) in *ycf1* cells expressing *GFP*, *TdHMA3-A1-GFP*, *TdHMA3-B1a-GFP*, *TdHMA3-B1b-GFP*, and *TdHMA3-B1b-ORF2-GFP*. Differential interference contrast (DIC) and merged images provide spatial references. Scale bars, 2 μm.



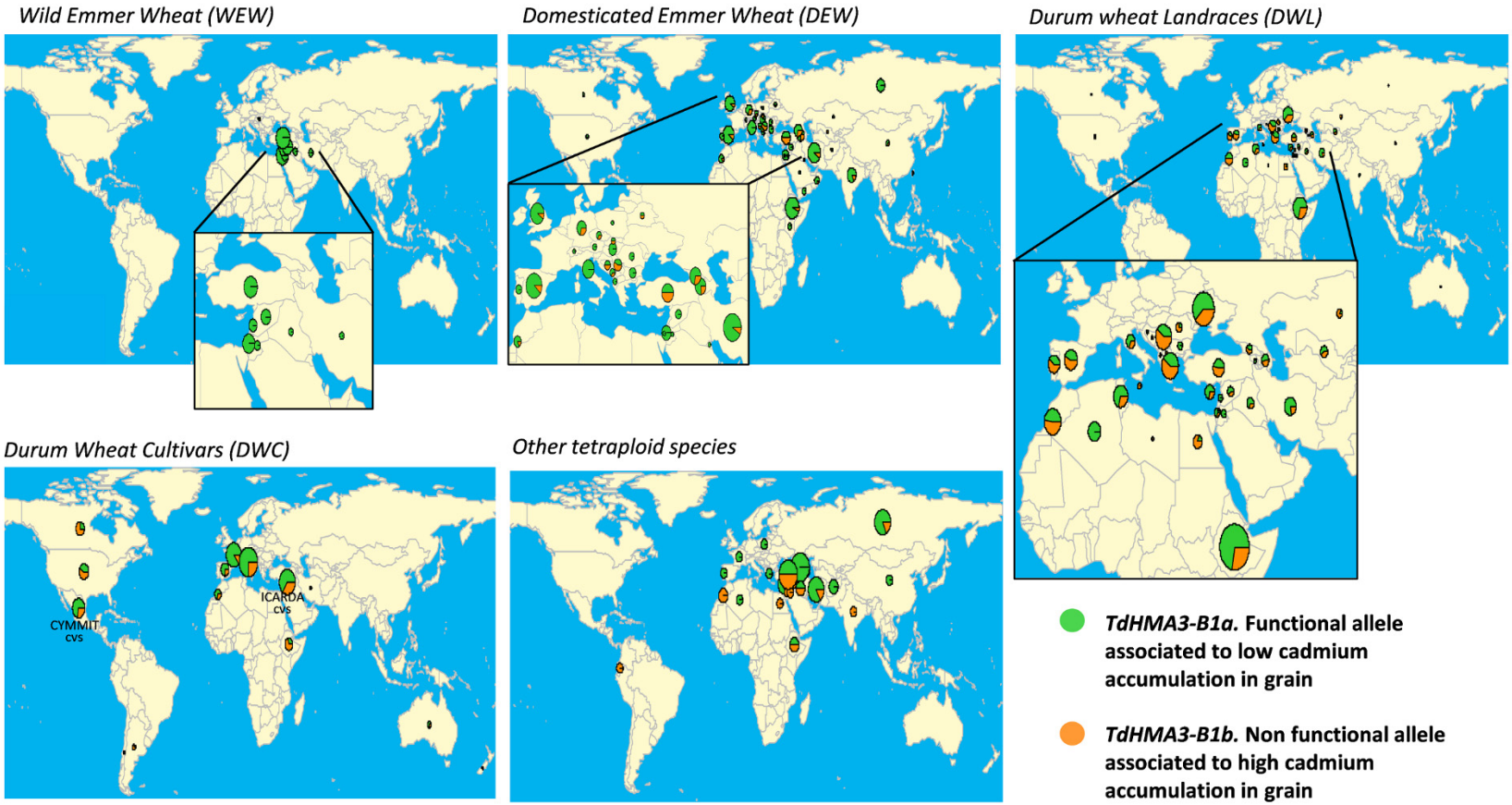
Supplementary Fig. 13: *TdHMA3*-mediated Cd and Zn accumulation in yeast is attributable to P-ATPase ion transport activity. a, Cd accumulation in Cd-sensitive *ycf1* yeast expressing empty vector (EV, p413TEF), *YCF1*, *TdHMA3-A1*, *TdHMA3-B1a*, *TdHMA3-A1*-D411A, and *TdHMA3-B1a*-D411A after exposure to 5 µM Cd for 4 h. **b**, Zn accumulation in Zn-sensitive *zrc1cot1* yeast expressing empty vector (EV, p413TEF), *ZRC1*, *TdHMA3-A1*, *TdHMA3-B1a*, *TdHMA3-A1*-D411A, and *TdHMA3-B1a*-D411A after exposure to 50 µM Zn for 4 h. Data are shown as means ± 95% CIs for $n = 4$ independent cultures (circles). Experiments repeated with similar results. *P* values were calculated by two-tailed, unpaired *t*-tests ($df = 6$). The dashed lines (**b**) indicate the Zn concentrations of the inoculating yeast cultures.



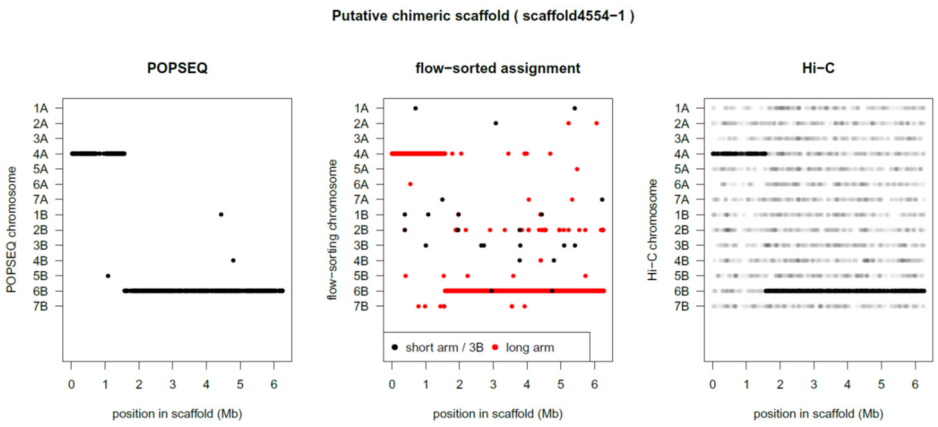
Supplementary Fig. 14: *TdHMA3-B1a* allele reduces Cd transport to shoots and grain during grain filling of DW in hydroponic culture. Developmental changes in Cd concentration of grain (**a**), whole-plants (**b**), shoots (**c**), and roots (**d**) of low Cd (open circles) and high Cd (closed circles) near-isogenic DW lines between anthesis and 42 d post-anthesis. Plants were grown in chelator-buffered hydroponic culture containing 0.5 µM Cd (Harris, N. S. & Taylor, G. J. Cadmium uptake and partitioning in durum wheat during grain filling. *BMC Plant Biol.* **13**, 103, 2013). Plotted values are means ± s.e.m. for n = 5 independent plants (n = 4 for 42 d). Contrasts between low and high Cd lines by two-way ANOVA F test are shown for each variate.



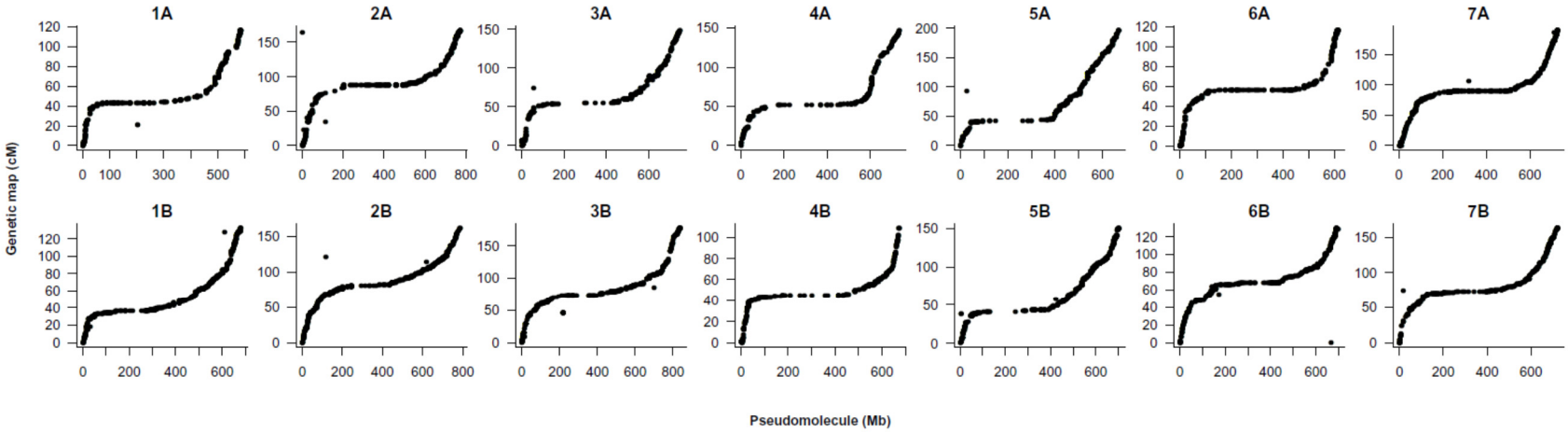
Supplementary Fig. 15: *TdHMA3-B1a* allele reduces Cd accumulation in shoots and grain during grain filling of field-grown DW. Developmental changes in Cd content (**a**, **b**) and Cd concentration (**c**, **d**) of grain (**a**, **c**) and shoots (**b**, **d**) of field-grown low Cd (open circles) and high Cd (closed circles) near-isogenic DW lines between anthesis and 43 d post-anthesis. Only mature grains were collected at 43 d post-anthesis after mechanically harvesting the plots. Plotted values are means ± s.e.m. for $n = 4$ independent plots arranged in a randomized block design. Contrasts between low and high Cd lines by two-way ANOVA F test are shown for each variate.



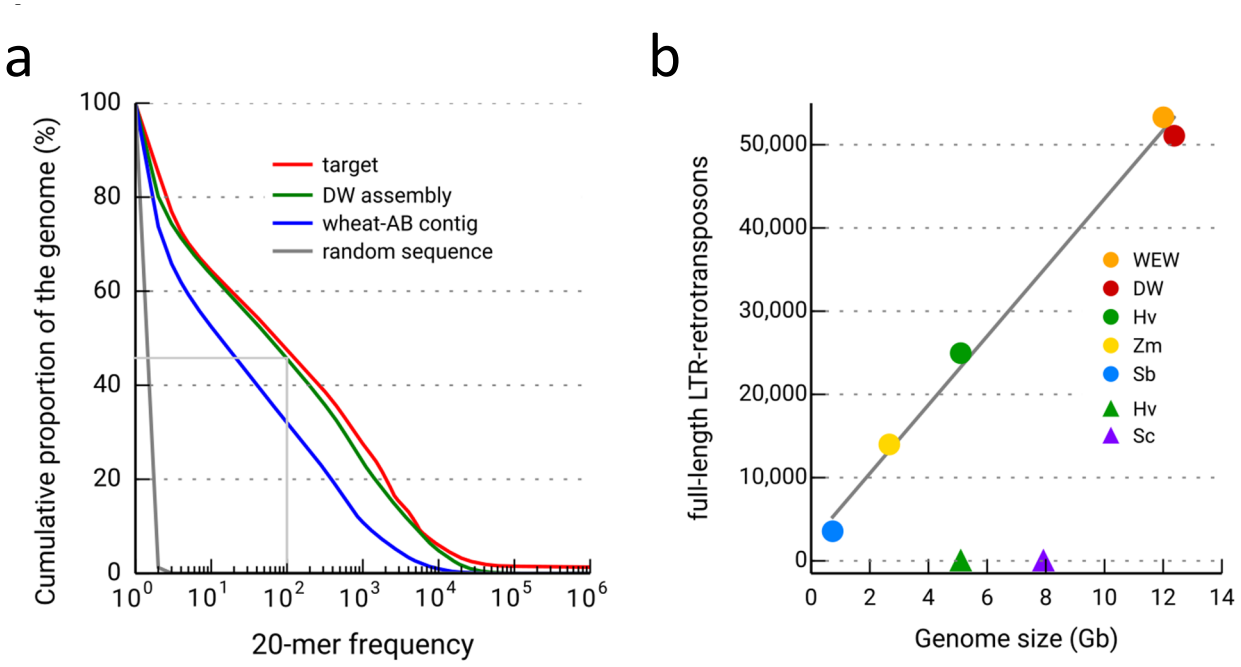
Supplementary Figure 16. *TdHMA3-B1a/b* allelic distribution by tetraploid wheat subspecies and by geographical origin of the 1,854 tetraploid wheat accessions from the world-wide diversity panel. Pie charts indicate the relative number of accessions in specific countries. The size of these pie charts is proportional to the number of accessions per country. *TdHMA3-B1a*, functional allele, equivalent to *Xusw59*, green-filled charts; *TdHMA3-B1b*, non-functional allele, equivalent to *Xusw59*, orange-filled charts. Figure made using the R packages rworldmap (https://github.com/AndySouth/rworldmap) and rworldxtra (https://github.com/AndySouth/rworldxtra).



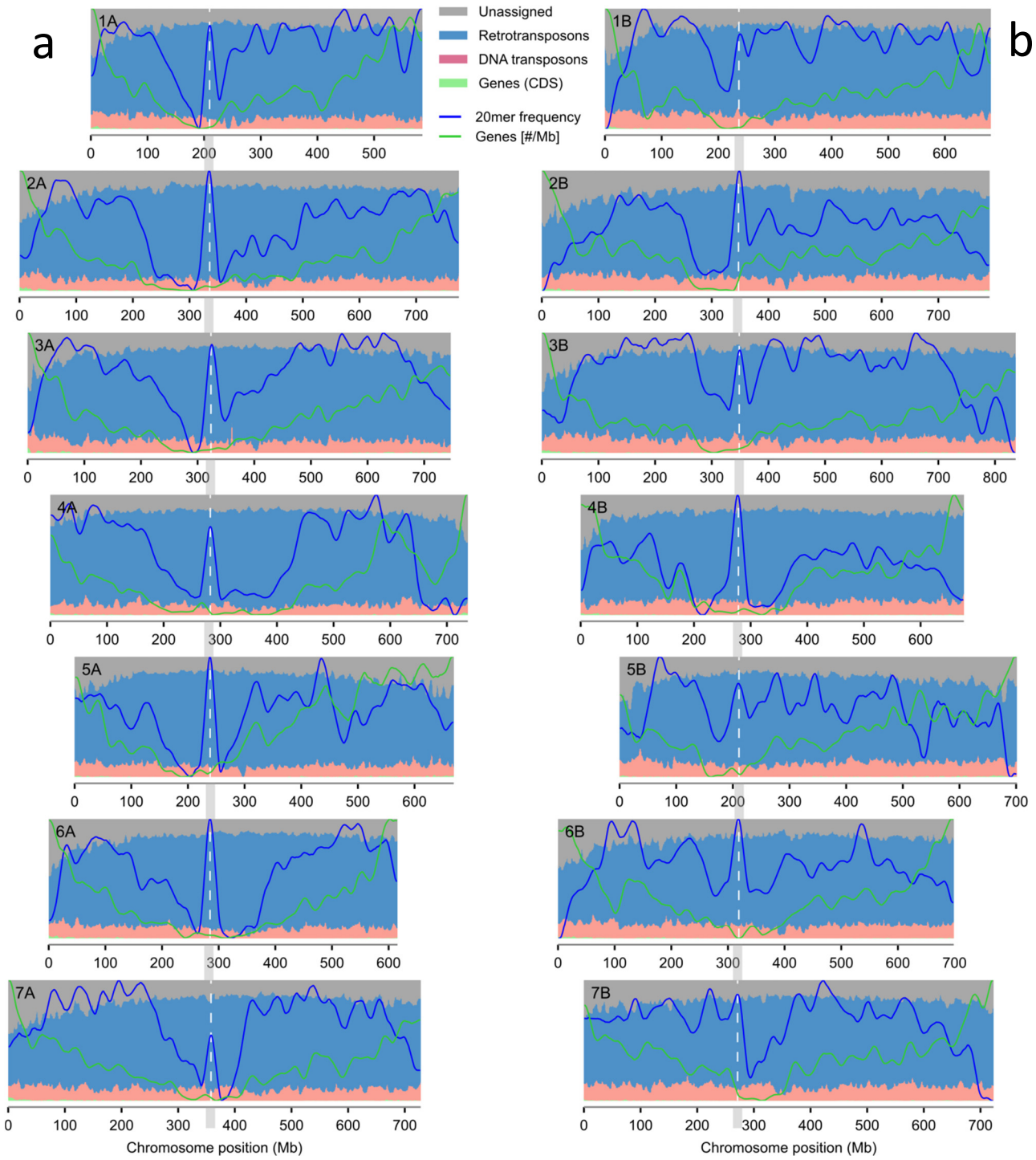
Supplementary Figure 17. Example of a chimeric scaffold. One of the eighteen chimeric scaffolds that were detected in the Svevo assembly v1.0 and subsequently split in the assembly v1.1 is reported. The chimeric nature of this scaffold (6.3 Mb in length) was supported by three lines of evidence: POPSEQ, flow-sorted chromosome assignment and Hi-C link information (panels from left to right, respectively). The first part of the scaffold originates from chromosome 4A, while the second part comes from chromosome 6B.



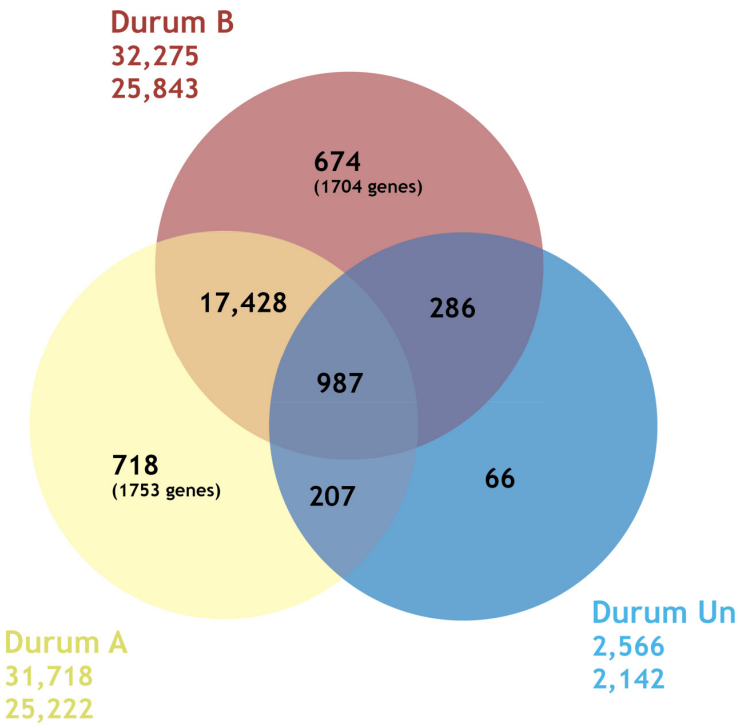
Supplementary Figure 18. Collinearity between the Svevo × Zavitan genetic map and the Hi-C map. The genetic position of the markers in Svevo × Zavitan genetic map (y-axis) are plotted against the coordinates of scaffolds in the final Hi-C-based pseudomolecules (x-axis).



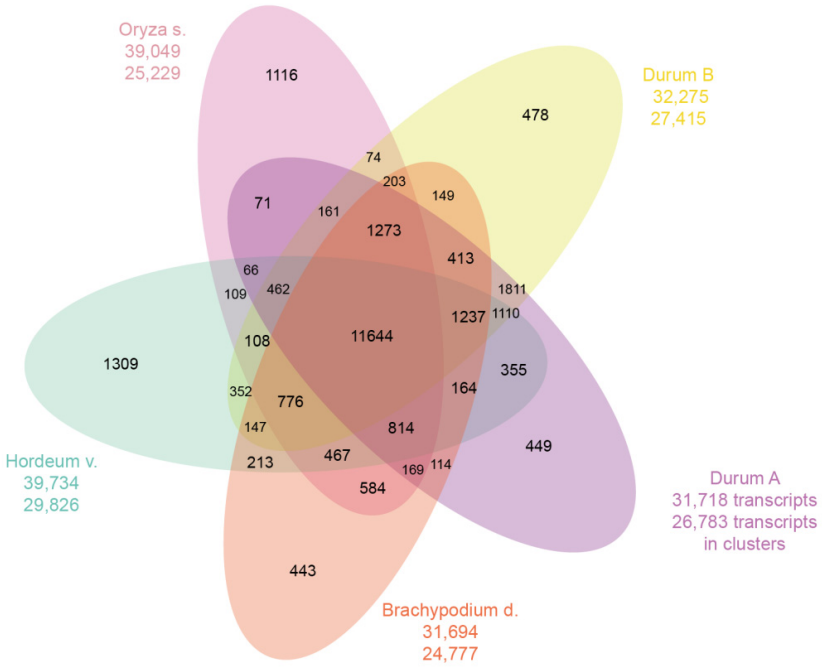
Supplementary Figure 19. Additional metrics supporting the good assembly quality of the highly repetitive Svevo genome**. a,** Mathematically defined repetitivity in form of 20mer frequencies. As an example, 45% of the Svevo asembly consist of 20mers that occur ≥ 100 times. Raw reads of WEW with a 1×-sequence coverage of the estimated genome size (12 Gbp, data.kew.org/cvalues/) served as proxy for total repetitiveness (red target line). This target is almost reached now by the DW assembly in contrast to the AB portion of an older hexaploid wheat assembly[104]. A comparison to random sequence of the same amount (with 4 as highest frequency count) shows that biological sequences are per se highly redundant. **b,** Number of full length LTR-retrotransposons (fl-LTRs) in different genome assemblies. Due to their almost identical 1-2 kb long terminal repeats fl-LTRs were often not well resolved in older contig assemblies (triangles). The amount of retrievable fl-LTRs is directly correlated to genome size and can serve as a complementary metric for the completeness and correct reconstruction of the repetitive space. Circles denote more complete assemblies compared to older contig assemblies (triangles). Sb: *Sorghum bicolor*[6]; Zm: *Zea mays*[184]; Hv: *Hordeum vulgare*[105],[185]; Sc: *Secale cereale*[186]; WEW: wild emmer, accession Zavitan[1], DW: durum wheat, accession Svevo, this paper.



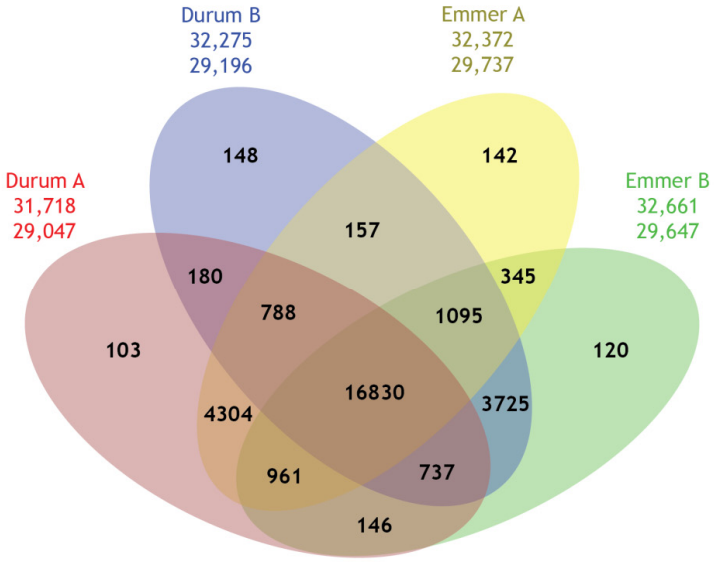
Supplementary Figure 20. Chromosomal Architecture of DW. The seven chromosomes of the A and B subgenome are aligned at their centromers. The background shows the percent composition of the main genomic components: LTR-retrotransposons (blue, 70%, 7.2 Gbp), DNA transposons (red, 11%, 1.2 Gbp) and genes (green, coding sequence without introns 0.8%, 81 Mb). The genes are barely visible at the percent level, their fine scale density variations along the chromosomes are given by the green line. 20-mer frequencies (blue line) are usually low at the gene rich distal ends, increase in the interstitial regions (young TEs), then decrease again (older more deteriorated TEs) in the proximal regions. In most chromosomes the direct centromere regions are highly repetitive due to their tandem array composition.



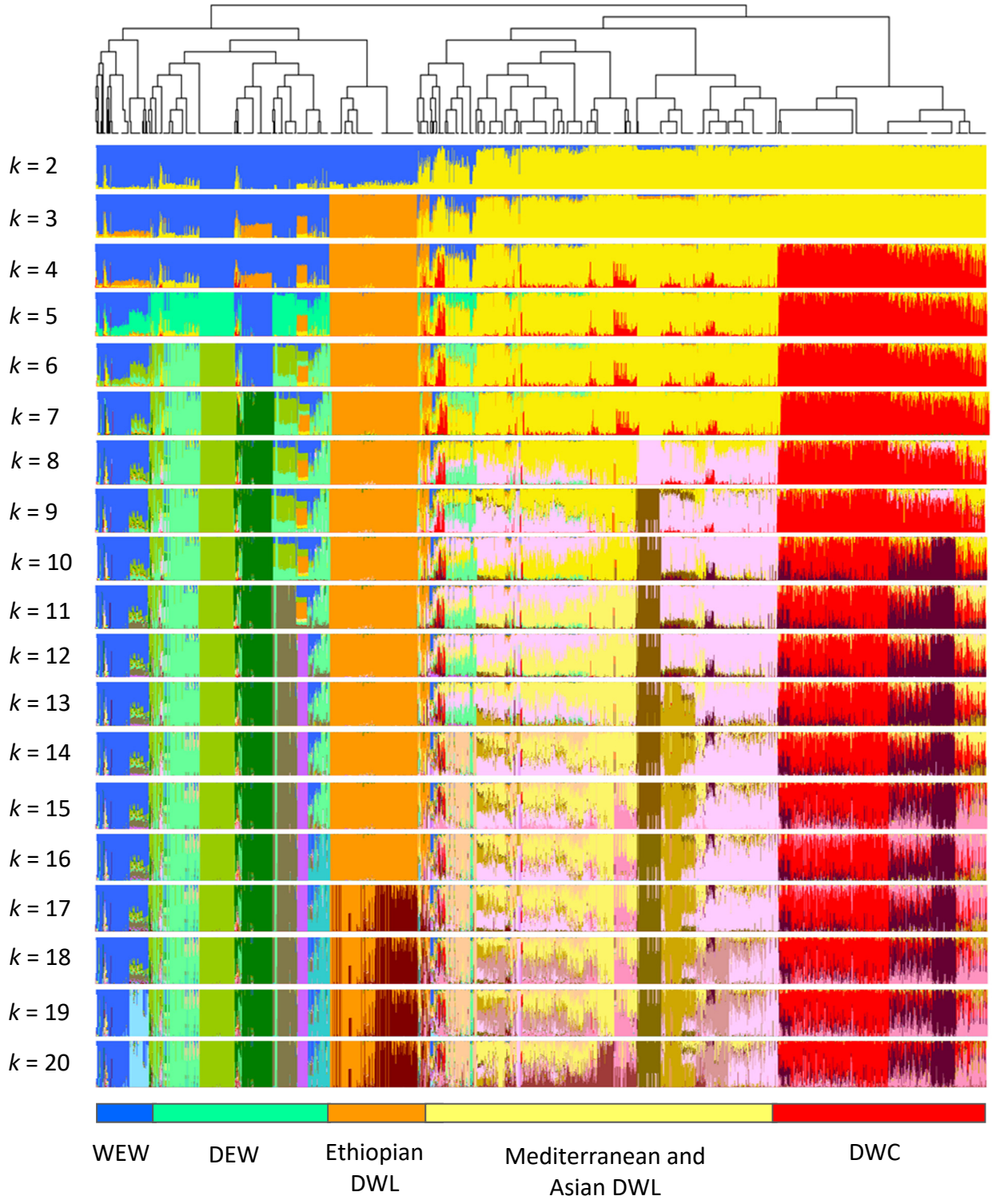
Supplementary Figure 21. OrthoMCL clustering of durum genes from the A and B subgenome and unclassified origin ("Un"). Numbers in the sections of the Venn diagram correspond to numbers of clusters (gene groups). The first number below the subgenome denotes the total number of proteins that were included into the OrthoMCL analysis for each subgenome. The second number indicates the number of genes in clusters for a species. The difference indicates the number of singletons (genes not clustered).



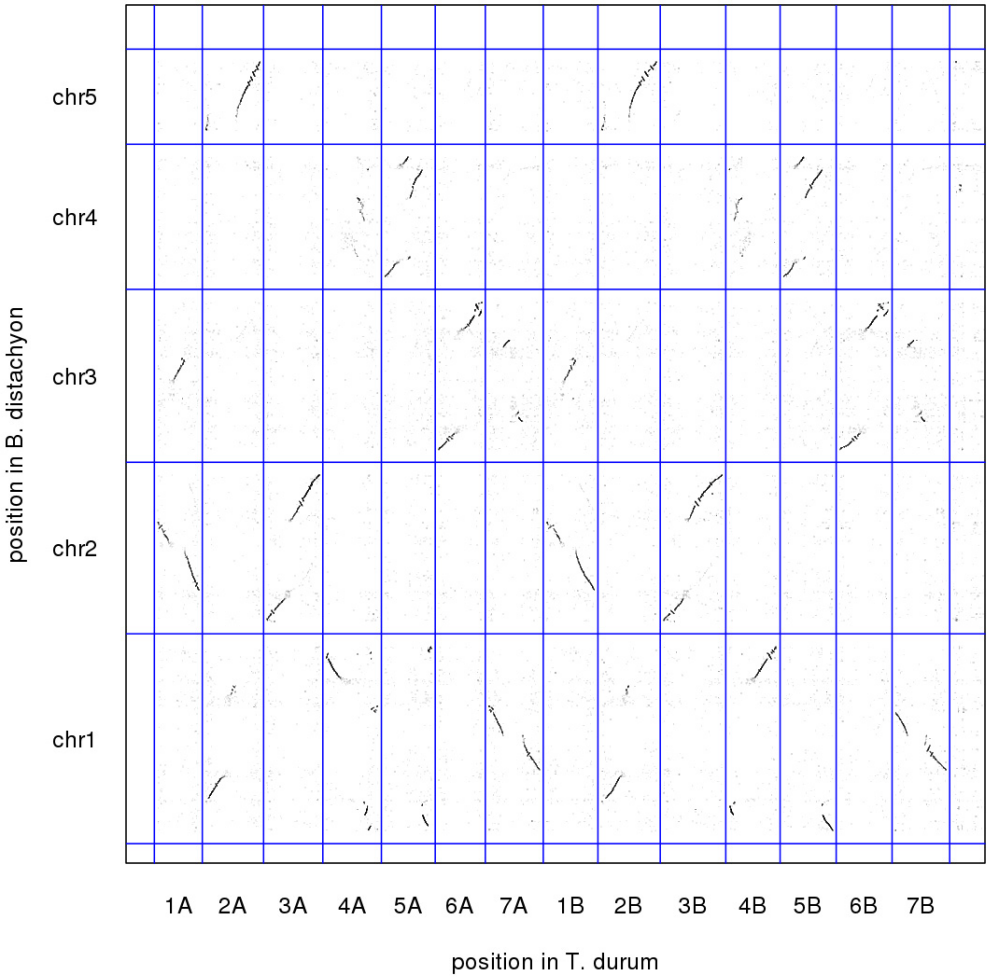
Supplementary Figure 22. OrthoMCL clustering of durum genes from the A and B subgenome and unclassified origin ("Un") with barley, *Brachypodium distachyon* and rice. Numbers in the sections of the Venn diagram correspond to numbers of clusters (gene groups). The first number below the species name denotes the total number of proteins that were included into the OrthoMCL analysis for each species. The second number indicates the number of genes in clusters for a species. The difference indicates the number of singletons (genes not clustered). Please note that genes from the durum unknown subgenome origin are not shown together with the other entities.



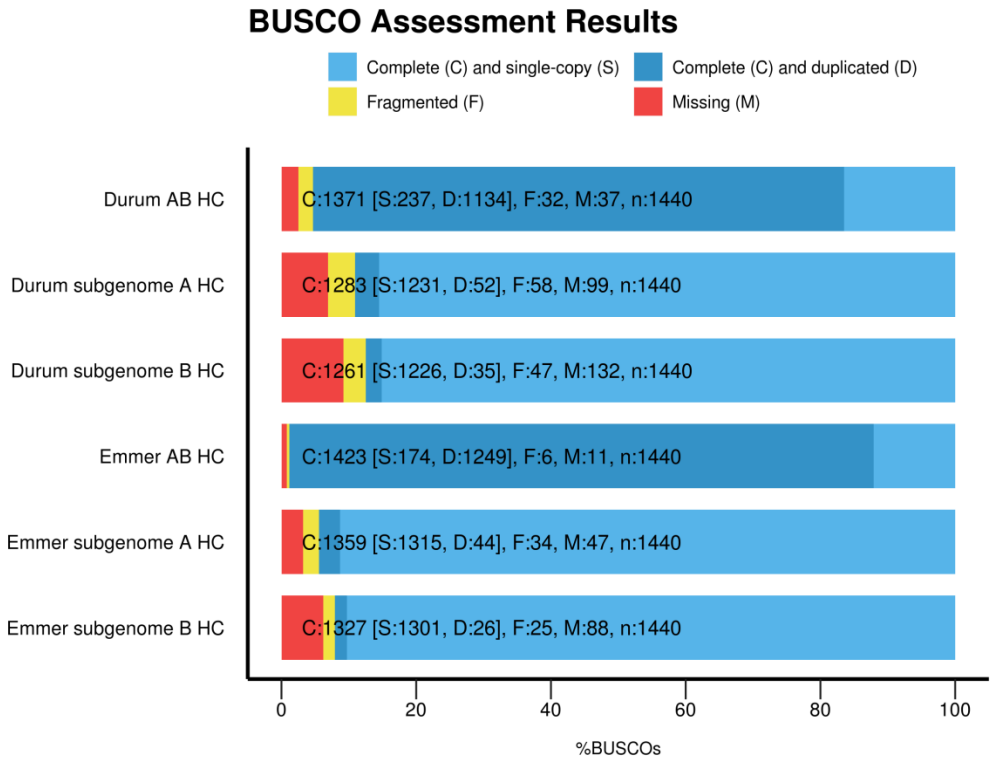
Supplementary Figure 23. OrthoMCL clustering of durum genes from the A and B subgenomes and unclassified origin ("Un") with wild emmer (WEW) genes from the A and B subgenomes and unclassified origin ("Un"). Numbers in the sections of the Venn diagram correspond to numbers of clusters (gene groups). The first number below the species name denotes the total number of proteins that were included into the OrthoMCL analysis for each species. The second number indicates the number of genes in clusters for a species. The difference indicates the number of singletons (genes not clustered). Please note that genes from the durum and emmer unknown subgenome origin are not shown together with the other entities.



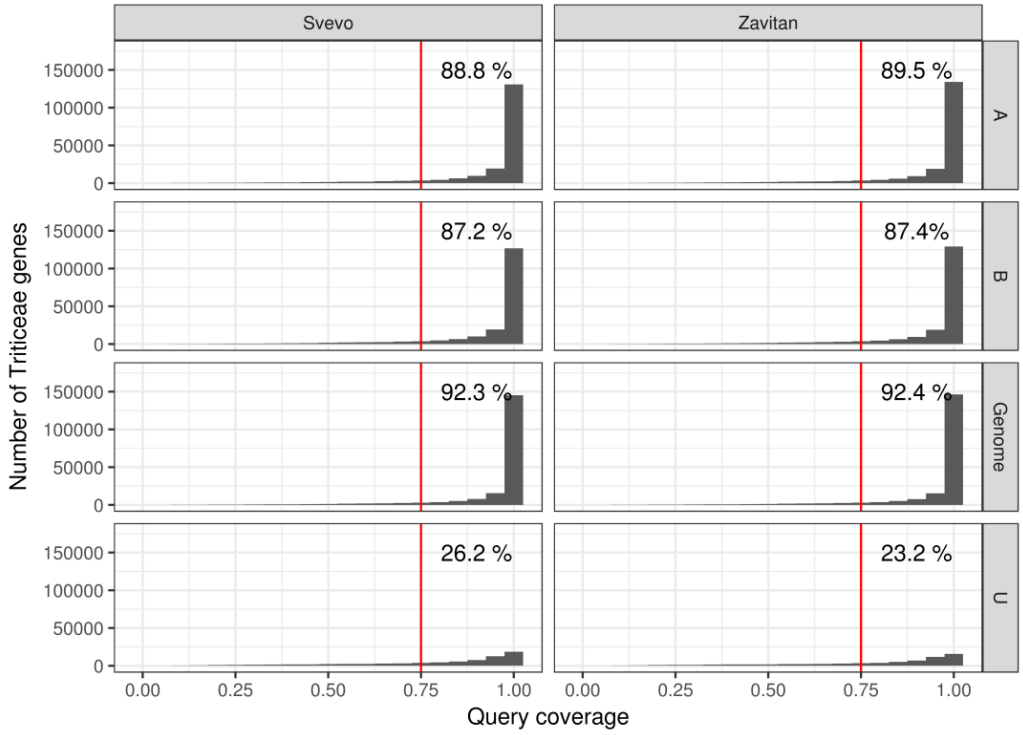
Supplementary Figure 24. Population structure of the tetraploid diversity panel as assessed globally by a single round of *ADMIXTURE* and *FineSTRUCTURE.* Genetic population structure analysis was carried out starting from a dataset of 17,416 iSelect 90K polymorphic SNP filtered for Mendelian segregation, failure rate and presence of singletons. *ADMIXTURE* runs for *k* hypothetical subpopulations from 2 to 20 were executed on a LD-pruned SNP dataset ($r^2 = 0.5$). *FineSTRUCTURE* was used to substructure the accessions at a finer level (110 subgroups). *ADMIXTURE* results were re-plotted after clustering and reordering the accessions to match the order of *FineSTRUCTURE* output. WEW: wild emmer wheat, DEW: domesticated emmer wheat, DWL: durum wheat landraces, DWC: durum wheat cultivars.



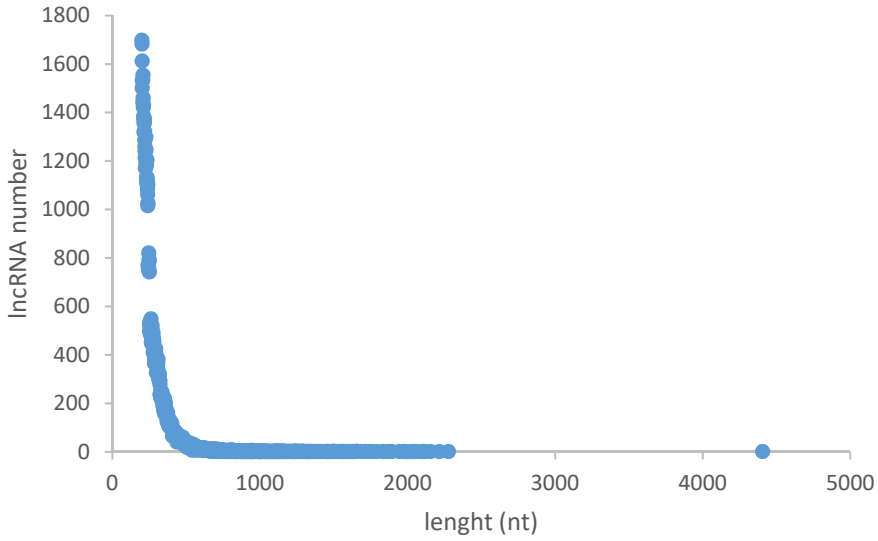
Supplementary Figure 25. Conservation of gene order between *T. durum* cv. Svevo and *Brachypodium distachyon*. Predicted *T. durum* proteins were aligned to *B. distachyon* protein sequences[4] with BLAST and the best hits were selected for visualization.



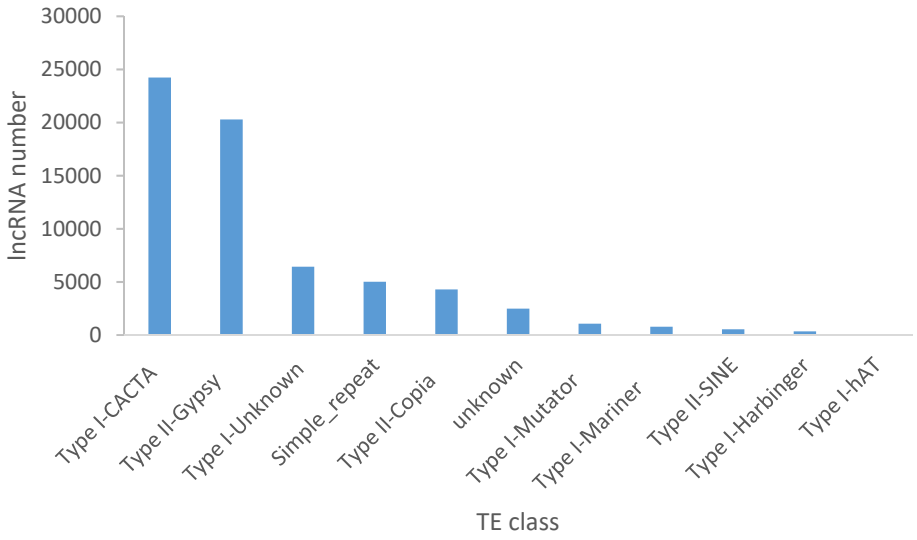
Supplementary Figure 26. Assessment of annotation completeness via BUSCO (Benchmarking Universal Single Copy Orthologs[14]. The plant reference set of BUSCO contains 1,440 genes. 95.2% of them are found in the HC gene sets of DW and respectively 99.8% in WEW. In both tetraploids most of them (79% for DW and 87% for WEW) are still present as duplicated copies, whereas the situation is reversed for the single subgenomes. Here most of the found reference genes are single copy, the single subgenomes only harbor between 2 and 4% duplicates. The BUSCO numbers are consistent with an expected gene loss after polyploidisation and give estimates for the losses to be around 6%: DW-A 6.1%, DW-B 7.6%, WEW-A 4.4%, and WEW-B 6.7%. Compared to the A genomes of both DW and WEW the losses are larger in the B genome.



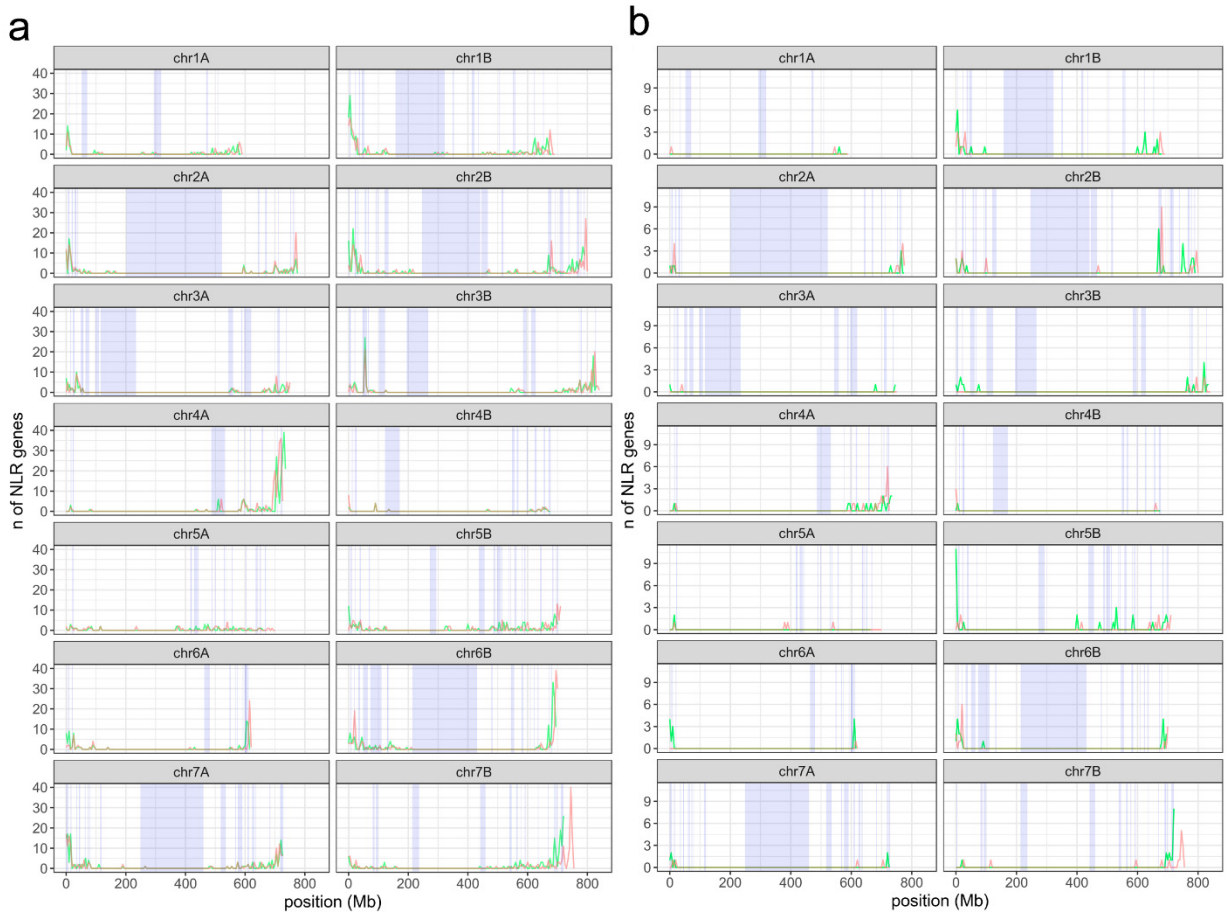
Supplementary Figure 27. Representation of the *Triticeae* reference gene set in the DW and WEW annotations.



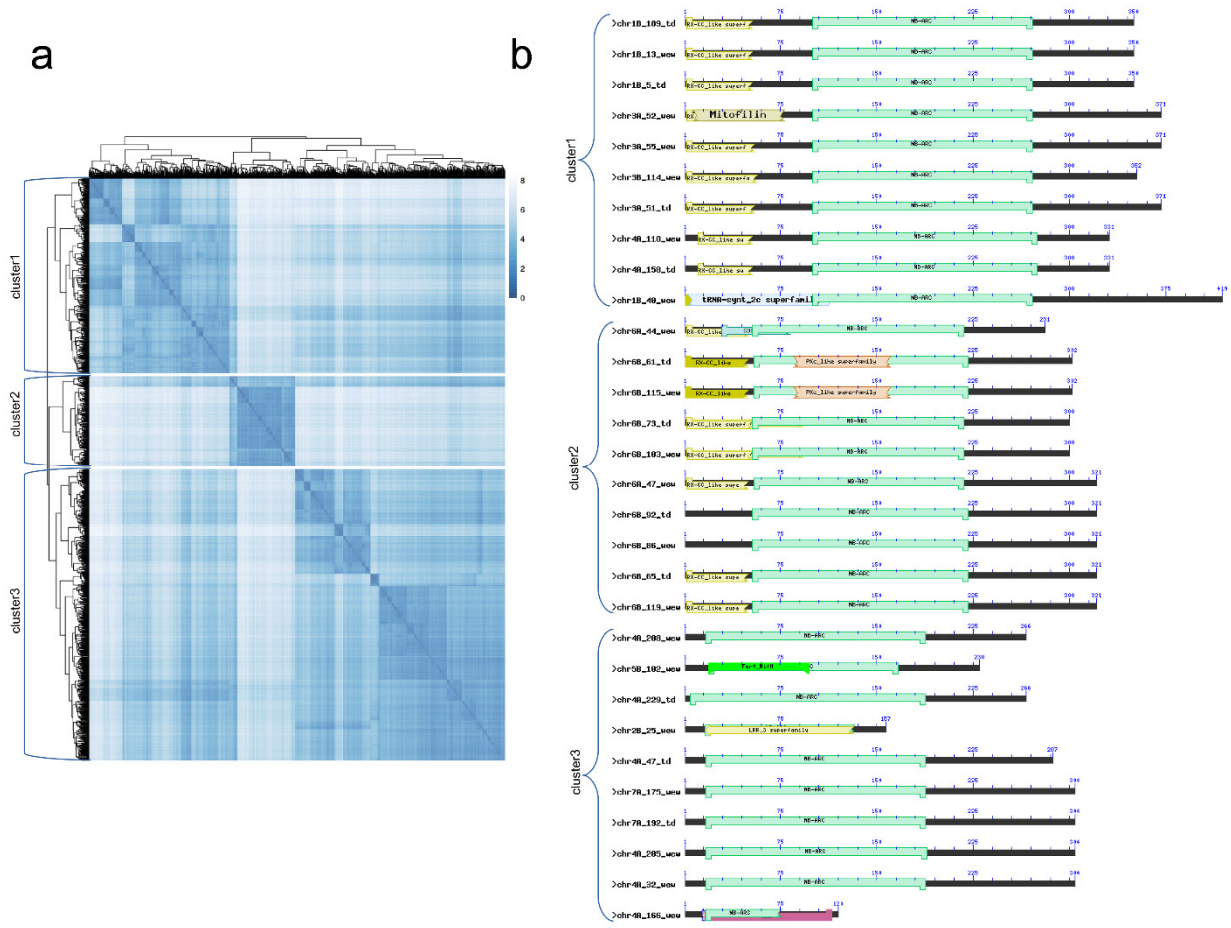
Supplementary Figure 28. Length distribution of lncRNAs.



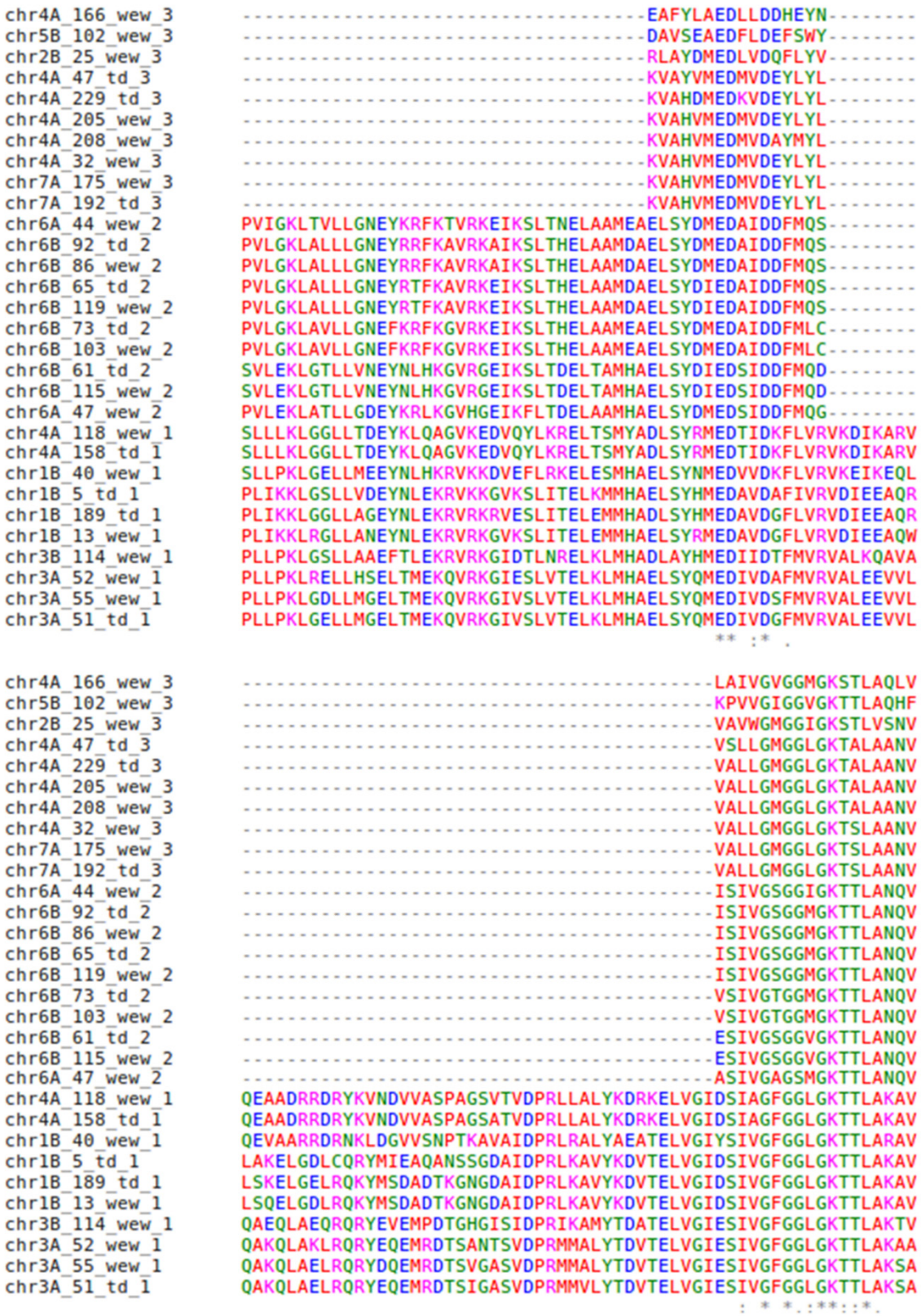
Supplementary Figure 29. TE association of identified lncRNAs. The most abundant transposable elements were represented with the base-pair coverage of lncRNAs.



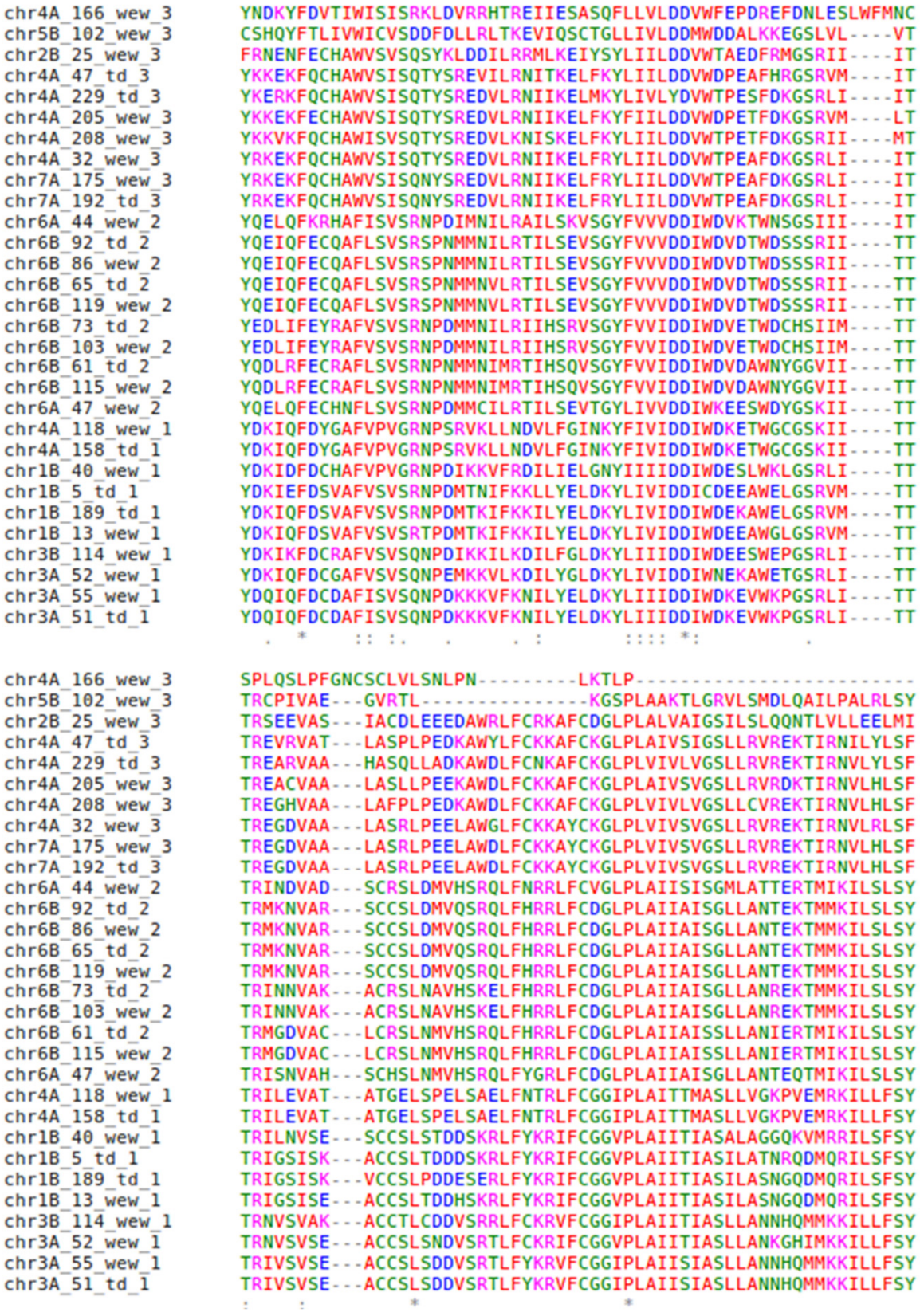
Supplementary Figure 30. Whole genome NLR gene density graph. Blue transparent area represents confidence intervals of published disease resistance QTLs. **a,** Nucleotide-binding leucine-rich repeat (NLR) gene density graph at 5 Mb window. Green colour represents Svevo DW; red, Zavitan WEW. **b,** Gene density of 172 NLR genes specific for DW and 136 for WEW at 5 Mb window (green, DW; red, WEW).



Supplementary Figure 31. NLR gene sub-groups. **a,** Similarity heatmap of complete sets of DW and WEW NLR genes. Orange dashed lines delimit the three NLR clusters. **b**, Domain composition of 10 most representative sequences for 3 clusters.

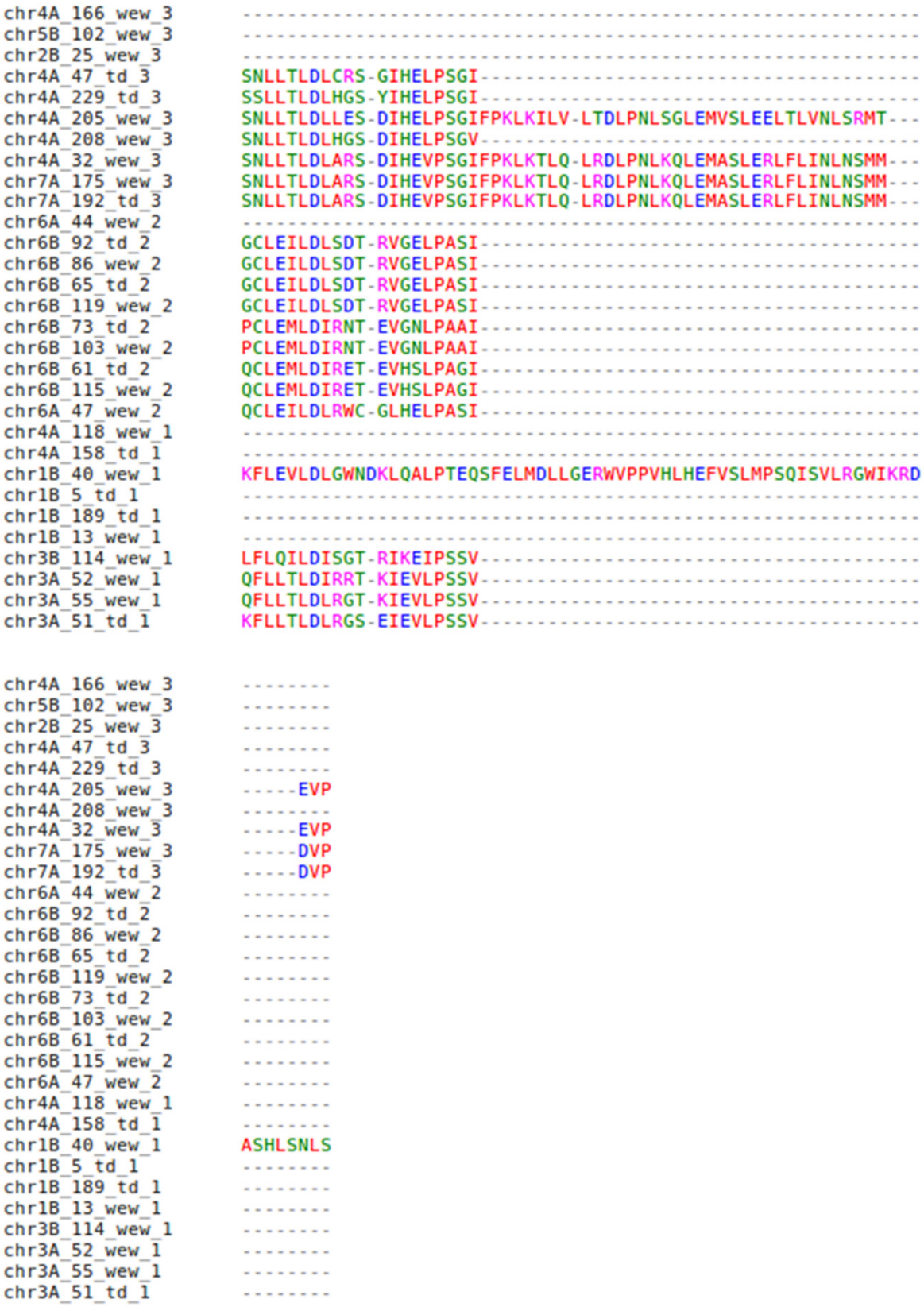


Supplementary Figure 32. Multiple Sequence Alignment of NLR genes. Ten most representative NLR gene sequences for each of the three clusters illustrated in Supplementary Fig. 31a are reported. For each gene, only the region spanning from first to last motif associated with NLR. The left side indicates the chromosome name, the NLR sequence number and the genome where the locus has been found (td: durum wheat; wew: wild emmer genome). 1, 2, 3, represent the cluster number as in Supplementary Fig. 31a.

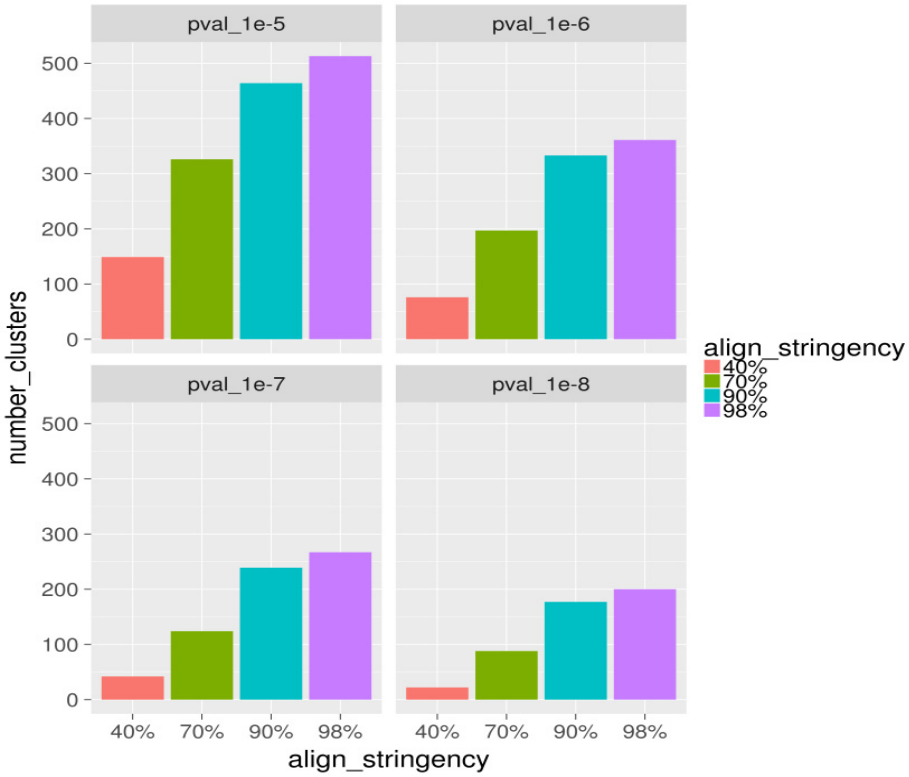


Supplementary Figure 32. Continued.

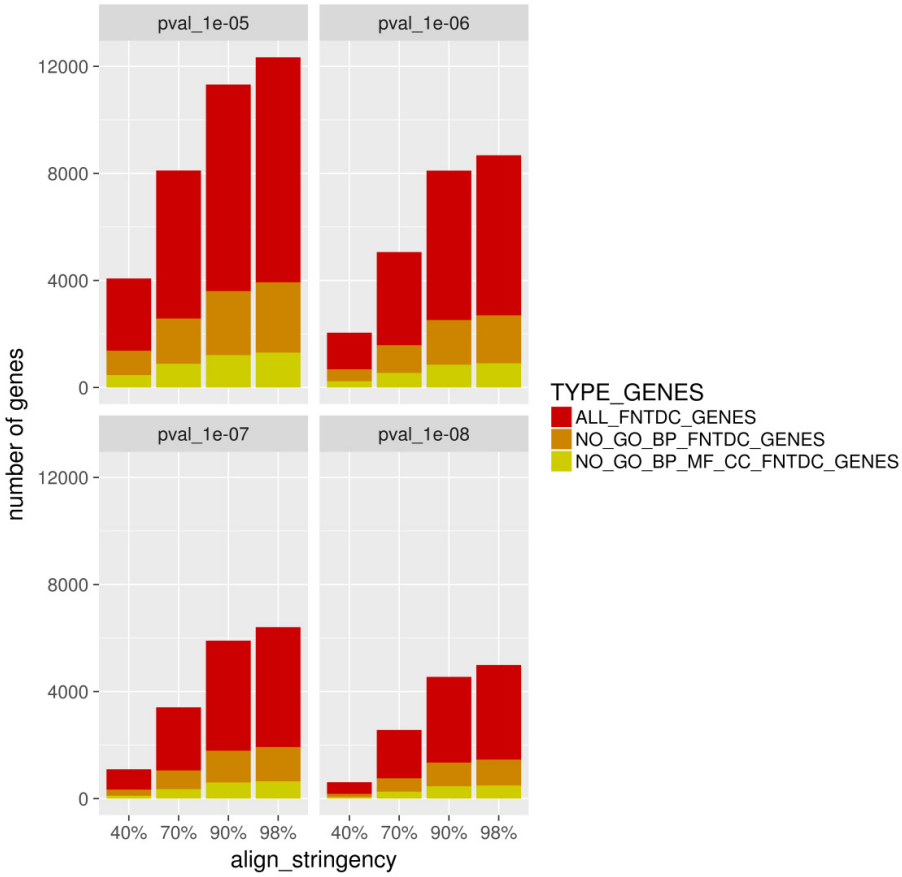
Supplementary Figure 32. Continued.



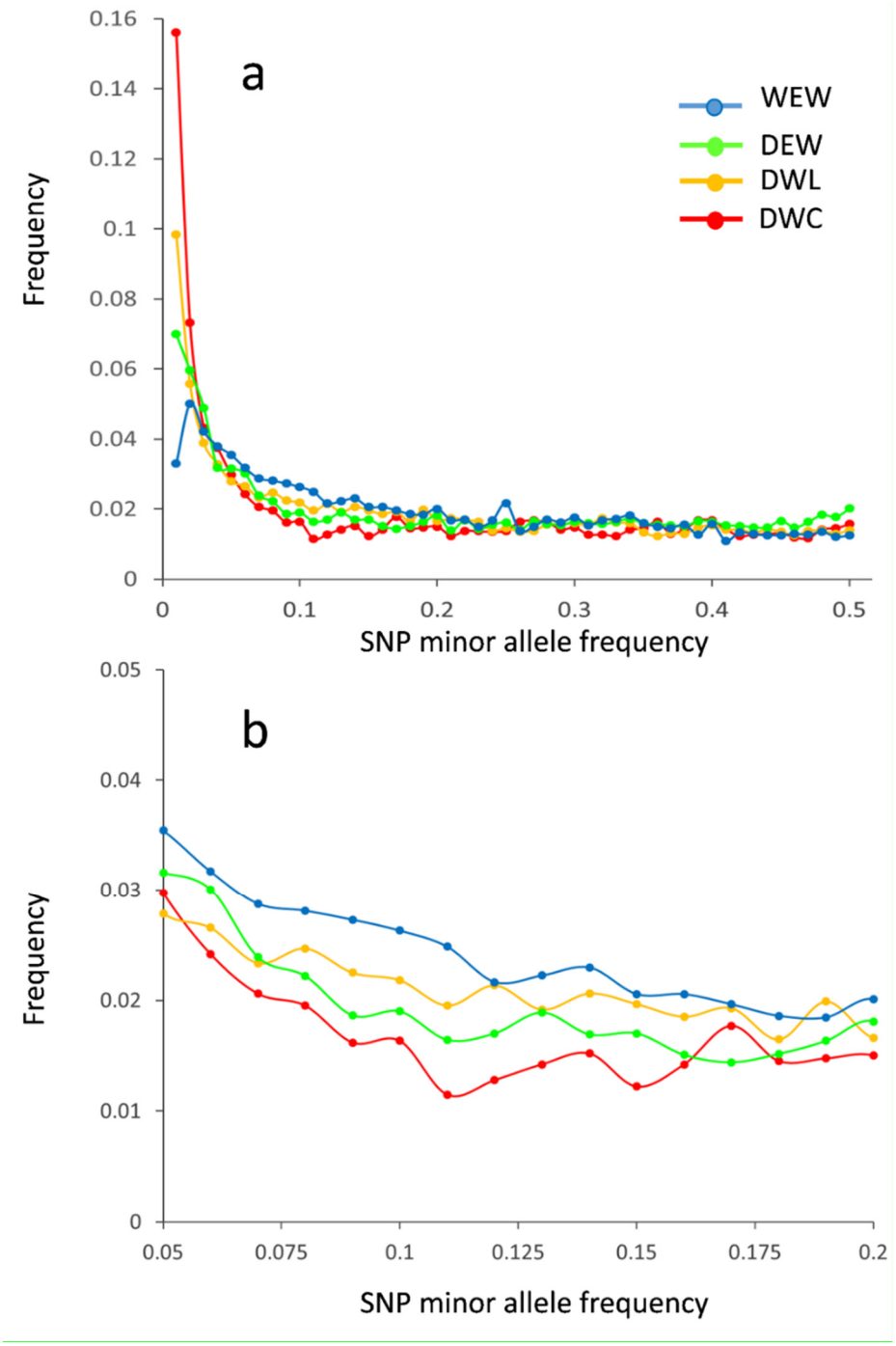
Supplementary Figure 32. Continued.



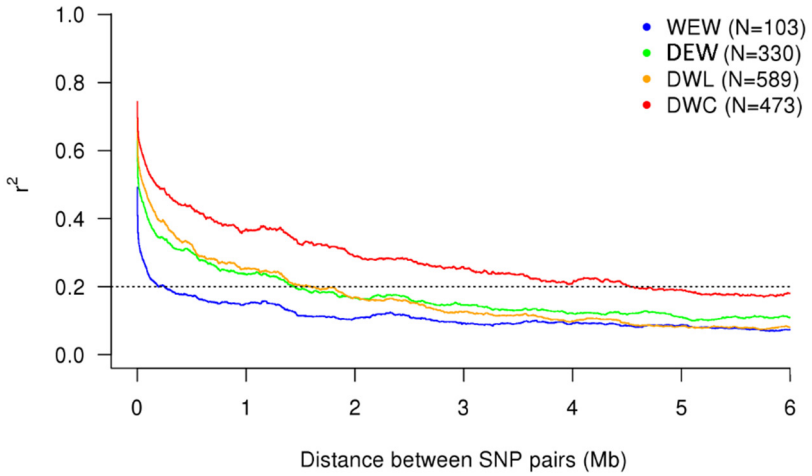
Supplementary Figure 33. Effect of homology detection stringency (tandem genes exclusion) and *p*-value threshold on the number of called plant functional non-tandem duplicated gene clusters (FNTDC) in durum wheat genome. Alignment stringency of 40% refers to a minimum 40% identity and ratio (both alignment length to query and alignment length to subject) of at least 0.4. as *p*-value threshold for GO-enrichment hypergeometric test.



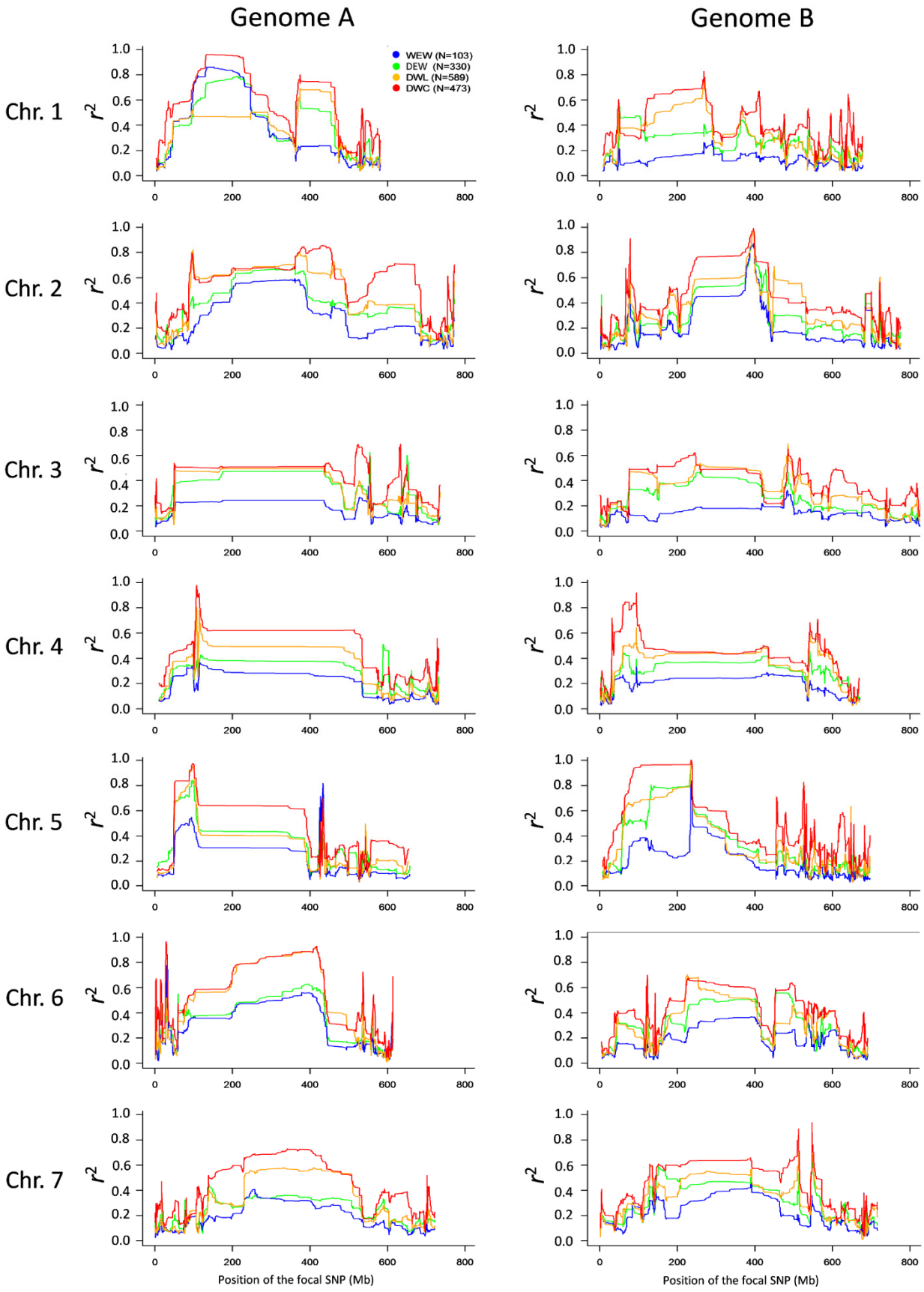
Supplementary Figure 34. Unknown genes embedded in candidate plant functional non-tandem duplicated gene clusters (FNTDC). The stacked barplots display, respectively, the total number of genes (red bars), the subset of the number of genes devoid of GO Biological Process (NO_GO_BP; orange bars) and the subset of genes devoid of any GO tag for all ontology domains (NO_GO_BP_MF_CC, yellow bars). Only genes embedded within called FNTDCs at the specified homology detection stringency (align stringency; x axis) and *p*-values as specified in sub-plot titles are considered. Cumulative values are shown (genes in subsets are not subtracted from supersets).



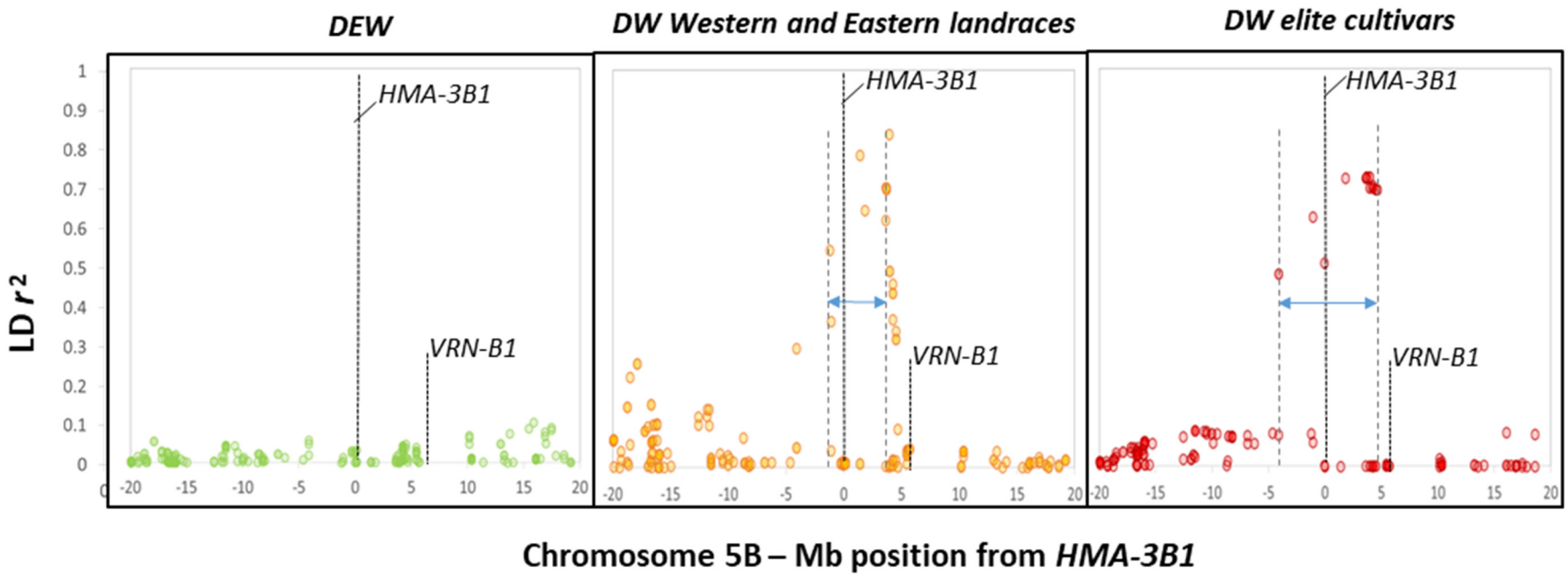
Supplementary Figure 35. Folded site frequency spectrum for 23,862 single locus genetically and physically mapped iSelect 90K wheat SNPs in four-main tetraploid wheat germplasm: WEW, DEW, DWL, DWC by 0.01 allele frequency steps. A: complete range; B: 0.05-0.20 minor allele frequency.



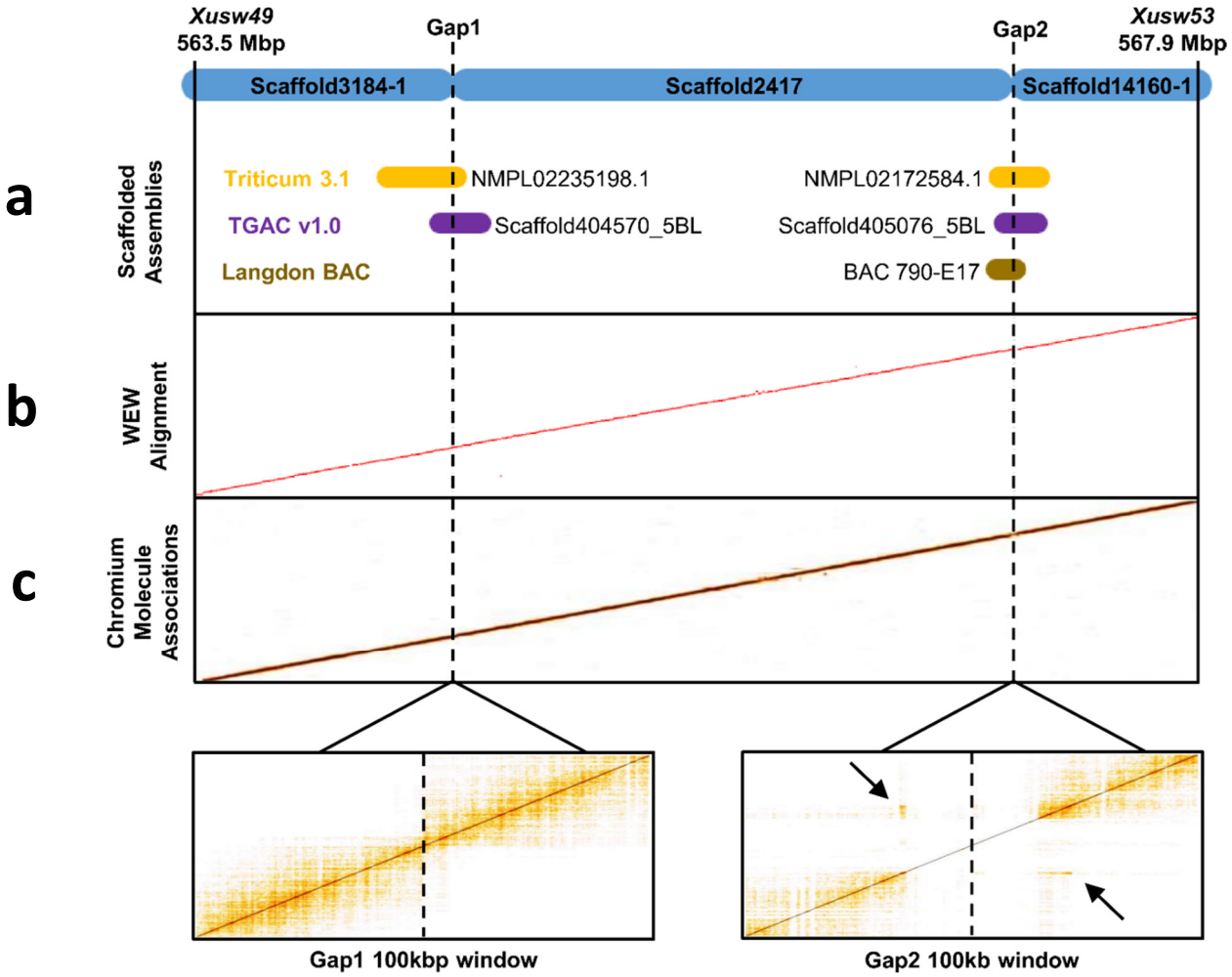
Supplementary Figure 36. LD decay over physical distance within germplasm groups. Considering an LD threshold value of $r^2 = 0.2$, the LD in WEW reached background level at 195 kb, DEW at 1.4 Mb, DWL at 1.6 Mb and DWC only at 4.5 Mb.



Supplementary Figure 37. Speed of LD decay estimated based on focal markers.



Supplementary Figure 38. LD decay in the *Cdu-B1* region. LD values (squared correlation coefficients, r^2) between *TdHMA3-B1* associated marker (*Xusw59*) and the other markers in the surrounding 40 Mb in DEW (330 accessions), DW landraces (589 accessions), and DW cultivars (473 accessions). The positions are indicated as Mb distance from *TdHMA-3B1*.



Supplementary Figure 39. Scaffold assembly and placement within *Cdu-B1* of the DW genome (cv. Svevo) is supported by collinearity to other wheat assemblies and by Chromium sequencing. *Cdu-B1* in the DW genome includes three scaffolds that were joined by POPSEQ and Hi-C; the two gaps between scaffolds are indicated by vertical hashed lines. **a,** Assembled scaffolds from the durum cultivar Langdon (brown), and two independent bread wheat assemblies for Chinese Spring, TGAC v1.0 (purple) and Triticum 3.1 (orange), align to scaffolds on both sides of the gaps. **b,** A NUCmer alignment of WEW and DW, represented as a dotplot, demonstrates strong collinearity between assemblies before and after gaps between scaffolds. **c,** Chromium sequencing of large DNA molecules from Svevo support the scaffold structure within *Cdu-B1*; orange dots indicate linked reads from the same DNA molecule (top). Increased resolution images show linked reads from the same molecule span both gaps (bottom). Though few molecule associations occur within Gap2, a large number of molecules span both sides of the gap (arrows).

Supplementary Tables: 1 to 24

Supplementary Table 1. Detailed sequencing data. PE: paired end, MP: mate paired.

#	Library type	Insert size	Sequencing Instrument	Read length	Minimal Coverage
1	PCR-free PE library	450-460 bp	HiSeq 2500	PE250 bp	X123
2	PCR-free PE library	700- 800 bp	HiSeqX	PE150-160 bp	X38
3	MP (Nextera MP Gel Plus)	2-4 kbp	HiSeqX	PE150-160 bp	X39
4	MP (Nextera MP Gel Plus)	5-7 kbp	HiSeqX	PE150-160 bp	X41
5	MP (Nextera MP Gel Plus)	8-10 kbp	HiSeqX	PE150-160 bp	X38

Supplementary Table 2. Detailed assembly results.

	Gap Stage: before filling		**Gap Stage: after filling**	
	Contigs	**Scaffolds**	**Contigs**	**Scaffolds**
Total number of sequences	3,477,456	260,924	474,837	129,464
Assembly size (bp)	9,715,852,855	10,377,602,212	10,299,203,836	10,450,113,204
Gaps size (bp)	0	579,420,725	0	149,190,797
Gaps %	0	5.58	0	1.42
N50-length (bp)	23,714	6,343,533	56,196	5,972,063
N50-number of sequences	236,188	458	54,378	493
N90-length (bp)	4,445	1,165,531	13,008	1,085,649
N90-number of sequences	923,879	1,870	194,738	2,019
Maximal length (bp)	301,754	54,143,938	475,246	54,262,061

Supplementary Table 3. Summary statistics of the durum wheat pseudomolecules.

Chromosome	Sequence length	Number of scaffolds
chr1A	585,266,722	126
chr1B	681,112,512	248
chr2A	775,448,786	167
chr2B	790,338,525	247
chr3A	746,673,839	190
chr3B	836,514,780	275
chr4A	736,872,137	241
chr4B	676,292,951	197
chr5A	669,155,517	155
chr5B	701,372,996	247
chr6A	615,672,275	140
chr6B	698,614,761	257
chr7A	728,031,845	198
chr7B	722,970,987	250
chrUn	498,719,471	126,526

Supplementary Table 4. Genomic features of pericentromeric, interstitial and distal chromosome regions for the 14 DW chromosomes: genetic and physical interval and size, physical-to-genetic ratio (recombination rate).

Region [(1)]	**Genetic interval**	**Physical interval**	**Genetic size**	**Physical size**	**Physical/genetic ratio**
	cM	Position (Mb)	(cM)	(Mb)	Mb/cM
1A-R1	0-36.33	0-27.8	36.33	27.8	0.765
1A-inter1	36.34-42.39	27.9-51.6	6.05	23.7	3.917
1A-C	42.4-45.38	51.7-360.7	2.98	309.0	103.691
1A-inter2	45.39-70.49	360.8-499.3	25.1	138.5	5.518
1A-R2	70.50-95.47	499.4-539.0	24.97	39.6	1.586
1A-inter3	95.48-95.82	539.1-566.5	0.34	27.4	80.588
1A-R3	95.83-116.48	566.6-585.3	20.65	18.6	0.901
1B-R1	0-27.43	0-34.0	27.43	34.0	1.240
1B-inter1	27.44-34.03	34.1-88.6	6.59	54.5	8.270
1B-C	34.04-38.09	88.7-314.3	4.05	225.6	55.704
1B-inter2	38.10-81.23	314.4-595.4	43.13	281	6.515
1B-R2	81.24-132.11	595.5-681.1	50.87	85.6	1.683
2A-R1	0-68.17	0-64.2	68.17	64.2	0.942
2A-inter1	68.18-87.15	64.3-201.1	18.97	136.8	7.211
2A-C	87.16-91.42	201.2-560.0	4.26	358.7	84.202
2A-inter2	91.43-108.17	561.0-666.9	16.74	105.9	6.326
2A-R2	108.17-167.17	667.0-775.5	59	108.4	1.837
2B-R1	0-67.59	0-120.0	67.59	119.9	1.774
2B-inter1	67.60-78.82	120.1-209.6	11.22	89.5	7.977
2B-C	78.83-82.18	209.7-441.5	3.35	231.9	69.224
2B-inter2	82.19-122.61	441.6-712.5	40.42	270.9	6.702
2B-R2	122.62-166.26	712.6-790.4	43.64	77.8	1.783
3A-R1	0-51.11	0-77.9	51.11	77.9	1.524
3A-inter1	51.12-53.70	78.0-109.4	2.58	31.4	12.171
3A-C	53.71-57.44	109.5-479.9	3.73	370.4	99.303
3A-inter2	57.45-83.92	480.0-596.6	26.47	116.6	4.405
3A-R2	83.93-148.48	596.7-746.7	64.55	149.9	2.322
3B-R1	0-48.09	0-61.1	48.09	61.1	1.271
3B-inter1	48.10-72.36	61.2-195.1	24.26	133.9	5.519
3B-C	72.37-78.46	195.2-504.8	6.09	309.6	50.837
3B-inter2	78.47-112.33	504.9-742.5	33.86	237.6	7.017
3B-R2	112.34-175.68	742.6-836.5	63.34	93.9	1.482
4A-R1	0-42.4	0-61.8	42.4	61.8	1.458
4A-inter1	42.50-51.73	61.9-133.1	9.23	71.2	7.714
4A-C	51.74-53.61	133.2-532.5	1.87	399.3	213.529
4A-inter2	53.62-67.60	532.6-592.5	13.98	59.9	4.285
4A-R2	67.7-114.46	592.6-648.5	46.76	56.0	1.198
4A-inter3	114.47-117.51	648.6-660.8	3.04	12.2	4.013
4A-R3	117.52-146.92	660.9-736.9	29.4	76.0	2.585
4B-R1	0-37.83	0-35.8	37.83	35.8	0.946
4B-inter1	37.84-42.68	35.9-71.2	4.84	35.3	7.293
4B-C	42.69-46.1	71.3-467.9	3.41	396.6	116.305
4B-inter2	46.1-70.67	468.0-642.3	24.57	174.3	7.094
4B-R2	70.68-108.25	642.4-676.3	37.57	33.9	0.902
5A-R1	0-38.19	0-39.7	38.19	39.7	1.040
5A-inter1	38.20-40.0	39.8-46.8	1.8	7	3.889
5A-C	40.1-44.70	46.9-385.4	4.6	338.5	73.587
5A-inter1	44.71-92.40	385.5-503.3	47.69	117.8	2.470
5A-R2	92.41-195.55	503.4-669.2	103.14	165.7	1.607
5B-R1	0-34.35	0-46.1	34.35	46.1	1.342

5B-inter1	34.36-40.53	46.2-89.9	6.17	43.7	7.083
5B-C	40.54-43.98	90.0-393.0	3.44	303.0	88.081
5B-inter1	43.97-76.78	393.2-536.0	32.81	142.8	4.352
5B-R2	76.79-148.95	536.1-701.4	72.16	165.3	2.291
6A-R1	0-41.32	0-38.9	41.32	38.9	0.941
6A-inter1	41.33-55.10	39.0-107.7	13.77	68.7	4.989
6A-C	55.2-58.24	107.8-480.6	3.04	372.9	122.664
6A-inter1	58.25-66.99	480.7-547.8	8.74	67.1	7.677
6A-R2	67-116.36	547.9-615.7	49.36	67.8	1.374
6B-R1	0-44.24	0-64.0	44.24	64.0	1.447
6B-inter1	44.25-65.48	64.1-161.7	21.23	97.6	4.597
6B-C	65.49-68.09	161.8-439.4	2.6	277.6	106.769
6B-inter1	69.00-101.25	439.5-649.0	32.25	209.5	6.496
6B-R2	101.26-129.59	646.0-698.6	28.33	52.6	1.857
7A-R1	0-67.68	0-82.9	67.68	83.0	1.226
7A-inter1	67.69-89.66	82.9-249.6	21.97	166.7	7.588
7A-C	89.65-90.77	249.6-510.4	1.12	260.8	232.857
7A-inter1	90.78-115.02	510.5-634.9	24.24	124.4	5.132
7A-R2	115.03-189.58	634.8-728.0	74.55	93.2	1.250
7B-R1	0-47.59	0-59.3	47.59	59.3	1.246
7B-inter1	47.60-69.96	59.4-135.4	22.36	76	3.399
7B-C	69.97-73.3	135.5-411.6	3.33	276.1	82.913
7B-inter1	73.4-101.90	411.7-613.7	28.5	202	7.088
7B-R2	101.91-162.73	613.8-723.0	60.82	109.1	1.794
Pericentromeric, low-recombinogenic regions (CHR-C)					
Total			*47.87*	*4,430.00*	*1,499.67*
Average			*3.42*	*316.43*	*107.12*
Interstitial regions (CHR-Inter)					
Total			*572.92*	*3,323.90*	*257.30*
Average			*19.10*	*110.80*	*8.57*
Highly-recombinogenic regions (CHR-C)					
Total			*572.92*	*2,206.90*	*43.61*
Average			*49.38*	*73.56*	*1.46*

(1) Chromosome regions are identified by chromosome and region identifiers, R1, highly-recombinogenic distal region in the short arm; R2 and R3, highly-recombinogenic distal regions in the long arm; C, recombination-depleted centromeric region.

Supplementary Table 5. RNASeq datasets used for gene prediction. A wide range of different RNA-Seq datasets were mapped to the genome assembly and provided (in addition to the mapping of reference genes) input data for the gene annotation pipeline. The data are available at the SRA database under accession SRP149116.

Origin	File in public repository	Tissue	Reads	Reference
Td cv. Svevo	This study	57 different treatments at seedling and adult plants organized in 9 pool of RNA samples (atlas gene expression)	2.8 billion reads	This study
Td cv. Svevo	This study	Grain at 5 developmental stages	1.7 billion reads	This study
Td cv. Senatore Cappelli	This study	Grain at 5 developmental stages	1.4 billion reads	This study
Ta cv. Chinese Spring	The Illumina and PacBio reads are available at study accession PRJEB15048 at EMBL-EBI European Nucleotide Archive	Leaf; root; seedling; seed; stem; spike	3 billion reads	10
Td cv. Kronos	Bioproject PRJNA191054 for *T. turgidum*. Raw data is available at the Short Read Archive (accession numbers: SRR769749, SRR769750, SRR863375, SRR863376, SRR863377, SRR863384, SRR863385, SRR863386, SRR863387, SRR863389, SRR863390, SRR863391, SRR863394)	Young roots; young shoots; spike; grain	0.5 billion reads	187
Ta cv. Chinese Spring	RNA-Seq data have been deposited under accession number ERP004714	Grain; leaf; root; spike; stem	2 billion reads	45
Td cv. Altar84	This study	Grain; root; leaf	0.18 billion reads	This study
Td cv. Capeiti8	This study	Grain; root; leaf	0.19 billion reads	This study
Td cv. Claudio	This study	Grain; root; leaf	0.16 billion reads	This study
Td cv. Creso	This study	Grain; root; leaf	0.17billion reads	This study
Td cv. Edmore	This study	Grain; root; leaf	0.20 billion reads	This study
Td. Cv. Kofa	This study	Grain; root; leaf	0.16 billion reads	This study
Td cv. Meridiano	This study	Grain; root; leaf	0.25 billion reads	This study
Td cv. Neodur	This study	Grain; root; leaf	0.21 billion reads	This study
Td cv. Saragolla	This study	Grain; root; leaf	0.18 billion reads	This study
Td cv. Strongfield	This study	Grain; root; leaf	0.16billion reads	This study
Td cv. Valnova	This study	Grain; root; leaf	0.2 billion reads	This study
Td. cv. Yavaros79	This study	Grain; root; leaf	0.133 billion reads	This study

Td cv. Svevo	This study	Grain; root; leaf; anther+ovaries; seed_milk	0.35 billion reads	This study
T. turgidum dicoccoides Zavitan accession	WEW: GeneBank LSYQ00000000 BioProject PRJNA310175	Leaf; root; flag leaf; developing spikes; glumes; flowers; grain;	0.50 billion reads	1
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, and SRR2084160.	Glumes	0.15 billion reads	188

Supplementary Table 6. a, Publicly available wheat molecular marker sets anchored to the Svevo genome assembly by BLAST. **b**, MetaQTLs identified per trait category.

a: Marker set	**Queries**	**Unique hits in Svevo(1)**	**Total hits in Svevo(2)**	**Average hits/query**
GBS(3)	939,536	846,001	846,001	-
iSelect 90K SNP	81,585	78,888	205,405	2.60
Axiom® 820K SNP	819,571	788,004	1,627,699	2.07
TaBW280K SNP	280,226	245,219	497,659	2.03
Axiom® 35K SNP	35,143	32,876	61,393	1.87
DArT	2,000	1,867	6,934	3.71
SSR	3,015	920	1,116	1.21
b: Trait class	**Linkage mapping QTL no.**		**GWAS-QTL no.**	
Domestication	47		-	
Phenology	86		114	
Plant height	51		53	
Disease response	132		275	
Biomass	98		35	
Quality-related traits	94		91	
Root apparatus	111		122	
Grain yield – related	526		249	
Other	17		4	

(1) Number of queries (markers) positioned on the Svevo genome assembly based on at least one BLAST hit according to the BLASTN setting parameters used (first hit).
(2) Total number of hits retrieved for each marker dataset according to the BLAST setting parameter used.
(3) For GBS markers, only BLAST best hits were retrieved.

Supplementary Table 7. Genomic features of pericentromeric, interstitial and distal chromosome regions for the 14 durum wheat chromosomes: physical interval, gene and QTL content, Gene and QTL density.

Region	Physical interval	Mb	Gene content (high confidence)	Gene density		QTL content	QTL density	QTL density
	Position (Mb)		No.	No./Mb	No./cM	No.	No./Mb	QTL /HCgenes
1A-R1	0-27.8	27.8	375	13.489	10.322	30	1.08	0.080
1A-inter1	27.9-51.6	23.7	195	8.228	31.736	5	0.21	0.026
1A-C	51.7-360.7	309	1,243	4.023	418.121	5	0.02	0.004
1A-inter2	360.8-499.3	138.5	960	6.931	38.247	13	0.09	0.014
1A-R2	499.4-539.0	39.6	483	12.197	19.263	8	0.20	0.017
1A-inter3	539.1-566.5	27.4	396	14.453	1185.294	8	0.29	0.020
1A-R3	566.6-585.3	18.7	256	13.763	12.300	9	0.48	0.035
1B-R1	0-34.0	34	434	12.765	15.749	23	0.68	0.053
1B-inter1	34.1-88.6	54.5	313	5.743	47.800	17	0.31	0.054
1B-C	88.7-314.3	225.6	796	3.528	196.296	15	0.07	0.019
1B-inter2	314.4-595.4	281	1,708	6.078	39.601	66	0.23	0.039
1B-R2	595.5-681.1	85.6	896	10.467	17.614	41	0.48	0.046
2A-R1	0-64.2	64.2	942	14.673	14.009	47	0.73	0.050
2A-inter1	64.3-201.1	136.8	924	6.754	48.708	20	0.15	0.022
2A-C	201.2-560.0	358.8	1,077	3.003	252.817	24	0.07	0.022
2A-inter2	561.0-666.9	105.9	730	6.893	43.668	16	0.15	0.022
2A-R2	667.0-775.5	108.5	1,528	14.096	25.881	63	0.58	0.041
2B-R1	0-120.0	120	1,235	10.300	18.272	89	0.74	0.072
2B-inter1	120.1-209.6	89.5	621	6.939	55.169	15	0.17	0.024
2B-C	209.7-441.5	231.8	858	3.700	256.716	10	0.04	0.012
2B-inter2	441.6-712.5	270.9	1,856	6.851	45.868	39	0.14	0.021
2B-R2	712.6-790.4	77.8	897	11.530	20.738	39	0.50	0.043
3A-R1	0-77.9	77.9	1,017	13.055	19.898	29	0.37	0.029
3A-inter1	78.0-109.4	31.4	206	6.561	79.845	3	0.10	0.015
3A-C	109.5-479.9	370.4	1,191	3.215	319.303	14	0.04	0.012
3A-inter2	480.0-596.6	116.6	855	7.333	32.301	31	0.27	0.036
3A-R2	596.7-746.7	150	1,702	11.354	26.398	53	0.35	0.031
3B-R1	0-61.1	61.1	853	13.961	17.758	75	1.23	0.088
3B-inter1	61.2-195.1	133.9	887	6.624	36.562	18	0.13	0.020
3B-C	195.2-504.8	309.6	1,216	3.928	199.836	14	0.05	0.012
3B-inter2	504.9-742.5	237.6	1,692	7.121	49.970	26	0.11	0.015
3B-R2	742.6-836.5	93.9	1,193	12.705	18.819	55	0.59	0.046
4A-R1	0-61.8	61.8	587	9.498	13.821	15	0.24	0.026
4A-inter1	61.9-133.1	71.2	421	5.913	45.720	9	0.13	0.021
4A-C	133.2-532.5	399.3	1,134	2.840	606.417	2	0.01	0.002
4A-inter2	532.6-592.5	59.9	648	10.818	47.425	27	0.45	0.042

Region	Physical interval	Mb	Gene content (high confidence)	Gene density		QTL content	QTL density	QTL density
	Position (Mb)		No.	No./Mb	No./cM	No.	No./Mb	QTL /HCgenes
4A-R2	592.6-648.5	55.9	726	12.964	15.590	43	0.77	0.059
4A-inter3	648.6-660.8	12.2	101	8.279	33.224	3	0.25	0.030
4A-R3	660.9-736.9	76	781	10.276	26.667	43	0.57	0.055
4B-R1	0-35.8	35.8	422	11.788	11.155	50	1.40	0.118
4B-inter1	35.9-71.2	35.3	247	6.997	51.033	14	0.40	0.057
4B-C	71.3-467.9	396.6	1,329	3.351	389.443	27	0.07	0.020
4B-inter2	468.0-642.3	174.3	1,079	6.190	43.793	46	0.26	0.043
4B-R2	642.4-676.3	33.9	419	12.360	11.179	24	0.71	0.057
5A-R1	0-39.7	39.7	370	9.320	46.504	22	0.55	0.059
5A-inter1	39.8-46.8	7	72	10.286	40.000	1	0.14	0.014
5A-C	46.9-385.4	338.5	746	2.204	162.174	13	0.04	0.017
5A-inter2	385.5-503.3	117.8	1,025	8.701	21.493	50	0.42	0.049
5A-R2	503.4-669.2	165.8	1,955	11.798	18.955	83	0.50	0.042
5B-R1	0-46.1	46.1	369	8.004	10.742	16	0.35	0.043
5B-inter1	46.2-89.9	43.7	293	6.705	47.488	6	0.14	0.020
5B-C	90.0-393.0	303	1,359	4.485	395.058	15	0.05	0.011
5B-inter2	393.2-536.0	142.8	1,238	8.669	37.732	34	0.24	0.027
5B-R2	536.1-701.4	165.3	1,794	10.853	24.861	54	0.33	0.030
6A-R1	0-38.9	38.9	590	15.167	14.279	26	0.67	0.044
6A-inter1	39.0-107.7	68.7	565	8.224	41.031	9	0.13	0.016
6A-C	107.8-480.6	372.8	1,110	2.977	365.132	10	0.03	0.009
6A-inter2	480.7-547.8	67.1	490	7.303	56.064	10	0.15	0.020
6A-R2	547.9-615.7	67.8	906	13.363	18.355	68	1.00	0.075
6B-R1	0-64.0	64	723	11.297	16.343	42	0.66	0.058
6B-inter1	64.1-161.7	97.6	613	6.281	28.874	21	0.22	0.034
6B-C	161.8-439.4	277.6	869	3.130	334.231	12	0.04	0.014
6B-inter2	439.5-649.0	209.5	1227	5.857	38.047	45	0.21	0.037
6B-R2	646.0-698.6	52.6	610	11.597	21.532	18	0.34	0.030
7A-R1	0-82.9	82.9	1,042	12.554	15.396	56	0.68	0.054
7A-inter1	82.9-249.6	166.7	1133	6.797	51.570	27	0.16	0.024
7A-C	249.6-510.4	260.8	849	3.255	758.036	10	0.04	0.012
7A-inter2	510.5-634.9	124.4	848	6.817	34.983	25	0.20	0.029
7A-R2	634.8-728.0	93.2	1,029	11.041	13.803	36	0.39	0.035
7B-R1	0-59.3	59.3	513	8.651	10.780	33	0.56	0.064
7B-inter1	59.4-135.4	76	501	6.592	22.406	17	0.22	0.034
7B-C	135.5-411.6	276.1	1,024	3.709	307.508	7	0.03	0.007
7B-inter2	411.7-613.7	202	1,187	5.876	41.649	36	0.18	0.030
7B-R2	613.8-723.0	109.2	1,035	9.487	17.017	82	0.75	0.079

Low-recombinogenic pericentromeric regions (CHR-C)								
Total		***4429.9***	***14801 (23.3%)***			***177 (8.4%)***	***-***	***-***
Average per region		***316.4***	***1057.2***			***12.7***	***0.041***	***0.012***
Interstitial regions (CHR-inter)								
Total		***3323.9***	***23031 (36.3%)***			***657(31.2%)***	***-***	***-***
Average per region		***110.8***	***767.7***			***21.9***	***0.209***	***0.029***
Highly-recombinogenic regions (CHR-R)								
Total		***2207.3***	***25682 (40.4%)***			***1,271 (60.4%)***	***-***	***-***
Average per region		***73.6***	***856.1***			***42.4***	***0.615***	***0.052***

Supplementary Table 8. Annotation statistics comparing durum wheat (DW) and wild emmer wheat (WEW) high confidence (HC) and low confidence (LC) genes.

Value	DW HC	WEW HC	DW LC	WEW LC
Number of genes	66,559	67,182	303,404	271,179
Number of genes on chrUn	2,566	2,149	12,145	7,947
Mean loci size (bp)	6,681	6,767	1,089	1,194
Median loci size (bp)	2,091	2,091	428	434
Number of single transcript genes	31,283	30,797	282,546	251,766
Number of multi transcript genes	35,276	36,385	20,858	19,413
Number of transcripts	196,153	205,916	341,975	307,880
Mean transcripts per gene	2.95	3.07	1.13	1.14
Mean CDS size (bp) *	1,241	1,241	520	522
Median CDS size (bp) *	1,056	1,062	414	417
Mean exons per transcript *	4.6	4.6	1.2	1.2
Median exons per transcript *	3	3	1	1
Number of single exon transcripts *	17,250	17,833	273,063	241,070
Number of multi exon transcripts *	49,309	49,349	30,341	30,109

* Numbers are for one representative transcript per gene.

Supplementary Table 9. Transposon composition of durum wheat (DW) and wild emmer wheat (WEW) expressed as a percentage of the entire genome and number of bases. The 3-letter code given in brackets refers to official transposon classification[187].

	DW		WEW	
	% of genome	**Mb**	**% of genome**	**Mb**
Mobile Element (TXX)	**82.2**	**8,596**	**82.2**	**8,639**
Class I: Retroelement (RXX)	**70.3**	**7,358**	**70.3**	**7,388**
LTR Retrotransposon (RLX)	70.0	7,320	69.9	7,350
Ty1/copia (RLC)	16.4	1,713	16.5	1,730
Ty3/gypsy (RLG)	32.5	3,396	32.4	3,409
unclassified LTR (RLX)	21.1	2,210	21.0	2,211
non-LTR Retrotransposon (RXX)	0.36	37.49	0.36	38.09
LINE (RIX)	0.34	36.08	0.35	36.64
SINE (RSX)	0.01	1.41	0.01	1.44
Class II: DNA Transposon (DXX)	**11.4**	**1,194**	**11.5**	**1,206**
DNA Transposon Superfamily (DTX)	11.3	1,179	11.3	1,192
CACTA superfamily (DTC)	10.9	1,143	11.0	1,156
hAT superfamily (DTA)	0.01	0.55	0.00	0.50
Mutator superfamily (DTM)	0.15	15.44	0.15	15.44
Tc1/Mariner superfamily (DTT)	0.04	3.68	0.03	3.67
PIF/Harbinger (DTH)	0.10	10.94	0.10	11.03
Unclassified (DTX)	0.05	5.59	0.05	5.66
MITE (DXX)	0.12	12.07	0.11	12.08
Helitron (DHH)	0.01	1.12	0.01	1.20
Unclassified DNA transposon (DXX)	0.01	1.23	0.01	1.27
Unclassified Element (TXX)	0.42	44.32	0.42	44.30
Retro-TE/DNA-TE ratio	6.20		6.10	
Gypsy/Copia ratio	2.00		2.00	

Supplementary Table 10. Comparison of unigene copy numbers between Svevo and Zavitan. **a,** The 36,434 unigene groups where divided into two main categories: (i) unaltered with identical member numbers for Svevo and Zavitan and (ii) changed with asymmetric unigene number between Svevo and Zavitan. Each category is subdivided again into different major scenarios. The table depicts the most common or typical cases for sequence variation scenarios, many more possible combinations exist. **b,** Nature of lineage specific genes. The ~4.800 lineage specific genes were mapped to the genome sequence of the other lineage. Around 2/3 of them are not completely lost, they still exist as degenerated shorter copies.

a Unigenes	**Unigene groups**	**% of unigene groups**	**Svevo HC genes**	**% of Svevo genes**	**Zavitan HC genes**	**% of Zavitan genes**	**Average group size**	**Median CDS length**
High confidence (HC) genes	36,434	100.0	66,559	100.0	67,182	100.0	3.7	1,059
Unigene group with balanced copy numbers (s=z)	**21,774**	**59.8**	**41,777**	**62.8**	**41,777**	**62.2**	**3.8**	**1,152**
2 copies each (s=z=2)	12,842	35.2	25,684	38.6	25,684	38.2	4.0	1,242
1 copy each (s=z=1)	6,793	18.6	6,793	10.2	6,793	10.1	2.0	756
>=3 copies each (s=z>2)	2,139	5.9	9,300	14.0	9,300	13.8	8.7	1,152
Unigene group with asymmetric numbers (s!=z)	**14,660**	**40.2**	**24,782**	**37.2**	**25,405**	**37.8**	**3.4**	**879**
Structural variants (s>0, z>0)	6,120	16.8	19,971	30.0	20,596	30.7	6.6	918
more in Svevo (s>z)	2,846	7.8	11,522	17.3	7,634	11.4	6.7	894
1 Zavitan loss* (s=2,z=1)	852	2.3	1,704	2.6	852	1.3	3.0	1,034
Svevo gains* (s>2,z=2)	427	1.2	1,427	2.1	854	1.3	5.3	1,083
more in Zavitan (z>s)	3,274	9.0	8,449	12.7	12,962	19.3	6.5	936
1 Svevo loss* (d=1,w=2)	1,121	3.1	1,121	1.7	2,242	3.3	3.0	1,092
Zavitan gains* (s=2,z>2)	503	1.4	1,006	1.5	1,693	2.5	5.4	1,008
only in Svevo (z=0)	4,313	11.8	4,811	7.2	0	0.0	1.1	735
only in Zavitan (s=0)	4,227	11.6	0	0.0	4,809	7.2	1.1	768
Clusters	28,794	79.0	62,639	94.1	63,462	94.5	4.4	1,074
Singletons	7,640	21.0	3,920	5.9	3,720	5.5	1.0	744
** loss and gain interpretation assumes the most likely initial state of 2 copies each*								

b	number for Svevo	Svevo %	% of all Svevo genes	number for Zavitan	Zavitan %	% of all Zavitan genes	classification
Lineage specific genes	**4,811**	**100.0**	**7.2**	**4,809**	**100.0**	**7.2**	
without hit in the other genome	1,493	31.0	2.2	1,237	25.7	1.8	
>= 1 hit on the other genome	**3,318**	**69.0**	**5.0**	**3,572**	**74.3**	**5.3**	
>90% overlap to HC gene*	1,225	25.5	1.8	1,057	22.0	1.6	Fragment of a longer gene, not clustered
>90% overlap to LC or pseudogene*	1,095	22.8	1.6	1,539	32.0	2.3	Structurally modified, not a HC gene any more
not annotated*	965	20.1	1.4	920	19.1	1.4	Candidates for genes missed in the annotation
>90% overlap to TEs*	33	0.7	0.0	56	1.2	0.1	Transposon
* value for best hit, if >1 hit							

Supplementary Table 11. Composition of the Global Tetraploid wheat Collection, including 1,856 tetraploid accessions (AABB genome) from 11 taxa and five hexaploid wheat (AABBDD genome) lines used for whole-genome SNP diversity analysis.

Wheat species or subspecies	**Common name**	**Genome**	**No.**
Triticum aestivum L. em Thell. subsp. *aestivum*	Common wheat	AABBDD	3
Triticum petropavlovskyi Udacz. et Migusch.	Xinjiang rice wheat	AABBDD	2
Triticum karamyschevii Nevski	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	AABB	364
Triticum turgidum L. subsp. *durum* (Desf.) Husn.	Durum wheat or pasta wheat (landraces)	AABB	806
Triticum turgidum L. subsp. *durum* (Desf.) Husn. (registered cultivars or breeding lines)	Durum wheat or pasta wheat (registered cultivars or breeding lines)	AABB	427
Triticum isphahanicum Heslot	Domesticated emmer wheat (Isfahan wheat)	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

Supplementary Table 12. Cloned genes relevant to durum wheat breeding and/or known to be under selection during emmer domestication, durum wheat evolution under domestication or breeding based on literature and their position on to the Svevo genome. The same genes are reported in Figs. 5 and 6 for comparison with reduced diversity and selection signal clusters.

locus_acronym	Locus_name	Chromosome	Sequence.start	Sequence.end	Mb	Ref.
Glu-A3	Glutenins	chr1A	5047553	5048723	5,05	189
TaSUT1A	Sucrose transporter	chr1A	194456769	194515110	194,46	190
Glu-A1	Glutenins	chr1A	500859392	501060448	500,86	153
T6P	Trehalose-6-phosphate synthase	chr1A	520686209	520692550	520,69	191
ELF3-A1	Early flowering 3	chr1A	582981365	582985067	582,98	192
TaSUT1B	Sucrose transporter	chr1B	230604772	230650472	230,60	190
ELF3-B1	Early flowering 3	chr1B	676974612	676978342	676,97	193
Ppd-A1	Photoperiod respose	chr2A	36577899	36565231	36,58	157
TaSus2-2A	Sucrose synthase	chr2A	120335255	120340200	120,34	194
TaSdr-A1	Seed dormancy	chr2A	156408483	156409475	156,41	195
TaCwi-A1	Cell wall invertase	chr2A	501893554	501897261	501,89	196
Ppd-B1	Photoperiod respose	chr2B	56297789	56294941	56,30	197
TaSus2-2B	Sucrose synthase	chr2B	169016255	169020790	169,02	194
TaSdr-B1	Seed dormancy	chr2B	198376152	198377132	198,38	195
TaCwi-B1	Cell wall invertase	chr2B	439343754	439347239	439,34	198
BRT-3A	Brittle Rachis	chr3A	61344533	61345121	61,34	1
BRT-3B	Brittle Rachis	chr3B	96155280	95381784	96,16	1
Rht-A1	Reduced height	chr4A	575088221	575090083	575,09	199
Phs-A1	Seed dormancy	chr4A	598755842	598762987	598,76	200
Rht-B1	Reduced height	chr4B	29292990	29294855	29,29	199
HMA3-A1	Heavy metal ATPase	chr5A	542961581	542964488	542,96	*This study*
VRN-A1	Vernalization	chr5A	549152139	549156384	549,15	150
Q-5A	Domestication	chr5A	608796291	608792747	608,80	154
HMA3-B1	Heavy metal ATPase	chr5B	563900691	563903585	563,90	*This study*
VRN-B1	Vernalization	chr5B	570831391	570844281	570,83	201
Q-5B	Domestication	chr5B	650078209	650075235	650,08	154
Phs-B1	Seed dormancy	chr5B	698826783	698832510	698,83	200
Gli	Alpha-gliadins	chr6A	24341990	24342853	24,34	202
NAC-A1	NAC domain-containing protein	chr6A	75453416	75454973	75,45	203
TaGW2-A	Grain weight	chr6A	235270703	235295537	235,27	155
Sr13-6A	Stem rust resistance	chr6A	611710263	611713775	611,71	204
NAC-B1	NAC domain-containing protein	chr6B	130826078	130826755	130,83	203
TaGW2-B	Grain weight	chr6B	300791272	300808374	300,79	155
Sr13-6B	Stem rust resistance	chr6B	689235987	689239462	689,24	204
VRN-A3	Vernalization	chr7A	69364420	69367738	69,36	205
TaTGW-7A	Grain weight	chr7A	204055853	204061744	204,06	158
TaCML20	Calmodulin 20	chr7A	686342874	686348391	686,34	206
VRN-B3	Vernalization	chr7B	9128364	9124817	9,13	205
TaTGW-7B	Grain weight	chr7B	168949495	168955358	168,95	158
TaCML20	Calmodulin 20	chr7B	663786935	663788698	663,79	206
Psy-B1	Phytoene synthase	chr7B	714361446	714362281	714,36	159

Supplementary Table 13. Putative functional genes within *Cdu-B1* region.

Gene[(1)]	Description[(2)]
TRITD5Bv1G197220	Kinase
TRITD5Bv1G197240	General transcription factor 3C polypeptide 3
TRITD5Bv1G197250	Stem-specific protein TSJT1, putative, expressed
TRITD5Bv1G197320	WD repeat and FYVE domain-containing protein 3
TRITD5Bv1G197370	**Zinc-transporting ATPase**
TRITD5Bv1G197440	Orotidine 5'-phosphate decarboxylase
TRITD5Bv1G197450	Transmembrane emp24 domain-containing protein
TRITD5Bv1G197460	ATP synthase subunit alpha
TRITD5Bv1G197470	Transport membrane protein
TRITD5Bv1G197480	NADH-ubiquinone oxidoreductase chain 6
TRITD5Bv1G197490	NADH-ubiquinone oxidoreductase chain 1
TRITD5Bv1G197500	ATP synthase subunit alpha
TRITD5Bv1G197520	NADH-ubiquinone oxidoreductase chain 6
TRITD5Bv1G197530	NADH-ubiquinone oxidoreductase chain 6
TRITD5Bv1G197540	Ribosomal protein S4
TRITD5Bv1G197650	Heat shock transcription factor
TRITD5Bv1G197710	Glycosyltransferase
TRITD5Bv1G197840	HXXXD-type acyl-transferase family protein, putative
TRITD5Bv1G197900	Bax inhibitor-1 family protein
TRITD5Bv1G198000	Kinase-like protein
TRITD5Bv1G198010	Histone H2B
TRITD5Bv1G198100	Pentatricopeptide repeat-containing protein, putative
TRITD5Bv1G198110	ATP-dependent RNA helicase SUV3
TRITD5Bv1G198120	WD-repeat protein, putative
TRITD5Bv1G198150	Pre-mRNA cleavage complex 2 protein Pcf11, putative isoform 2
TRITD5Bv1G198340	tRNA (guanine-N(7)-)-methyltransferase non-catalytic subunit
TRITD5Bv1G198350	DUF1677 family protein
TRITD5Bv1G198410	DUF1677 family protein
TRITD5Bv1G198420	DUF1677 family protein
TRITD5Bv1G198450	DUF1677 family protein
TRITD5Bv1G198460	DUF1677 family protein
TRITD5Bv1G198510	Sigma non-opioid intracellular receptor 1
TRITD5Bv1G198520	Protein OBERON 1
TRITD5Bv1G198650	Protein of unknown function (DUF642)
TRITD5Bv1G198720	Protein OBERON 1
TRITD5Bv1G198770	Tryptophan RNA-binding attenuator protein-like
TRITD5Bv1G198780	Seed maturation protein/Late embryogenesis abundant protein
TRITD5Bv1G198800	Late embryogenesis abundant D-like protein
TRITD5Bv1G198820	Dirigent protein
TRITD5Bv1G198830	Dirigent protein
TRITD5Bv1G198860	Dehydration-responsive element binding factor protein
TRITD5Bv1G198930	Acyl-CoA N-acyltransferase with RING/FYVE/PHD-type zinc finger protein, putative
TRITD5Bv1G198940	Photosystem II protein
TRITD5Bv1G198960	Myosin family protein, putative, expressed
TRITD5Bv1G199060	Protein MIZU-KUSSEI 1
TRITD5Bv1G199110	Evolutionarily conserved C-terminal region 2
TRITD5Bv1G199120	Short-chain dehydrogenase/reductase family protein
TRITD5Bv1G199130	PHD finger protein ING

(1) Filtered for genes with high confidence AHRD Quality Codes not annotated as transposable elements; *TdHMA3-B1* (*TRITD5Bv1G197370*) is indicated in red bold font.

(2) Genes were annotated with the AHRD tool (Automated Assignment of Human Readable Descriptions, https://github.com/groupschoof/AHRD, version 3.3.3).

Supplementary Table 14. Concentrations of Cd and mineral nutrients in mature grain of field-grown low Cd (8982-TL-L) and high Cd (8982-TL-H) DW near-isogenic lines (NILs).

Element (mg kg^{-1})	Low Cd NIL	High Cd NIL	*t*-test[1]	High/Low
N	26550 ± 269	27313 ± 557	$P = 0.221$	1.03
P	3218 ± 230	3193 ± 183	$P = 0.830$	0.99
K	4048 ± 102	4005 ± 230	$P = 0.895$	0.99
S	1813 ± 28	1890 ± 27	$P = 0.058$	1.04
Ca	300 ± 32	318 ± 58	$P = 0.844$	1.06
Mg	1340 ± 78	1325 ± 26	$P = 0.850$	0.99
Cd	0.052 ± 0.003	0.166 ± 0.010	$P = 6.5\times10^{-4}$	3.19
Cu	4.83 ± 0.36	4.90 ± 0.33	$P = 0.736$	1.01
Fe	49.6 ± 2.1	55.2 ± 2.6	$P = 0.125$	1.11
Mn	46.8 ± 3.5	52.2 ± 4.4	$P = 0.315$	1.11
Zn	31.1 ± 1.0	35.7 ± 2.2	$P = 0.078$	1.15

All concentration data are means ± s.e.m ($n = 4$ plots).
[1] Contrasts between low and high Cd NILs calculated by two-tailed, paired *t*-tests ($df = 6$).

Supplementary Table 15. *TdHMA3-B1a/b* allelic distribution of the Global Tetraploid wheat Collection (GTC) by country of origin, based on passport data. Accessions have been categorized by subspecies and by the geographical aggregates according to the United Nations M-49 list, except for the Fertile Crescent territories (Turkey, Southern Levant).

Tetraploid wheat taxa / geographical area	**Accessions**	***TdHMA3-B1a***	***TdHMA3-B1b***	***TdHMA3-B1a***	***TdHMA3-B1b***
	no.	no.	no.	freq.	freq.
Wild Emmer Wheat (*T. turgidum* ssp. *dicoccoides*)					
Fertile_Crescent_Southern_Levant (Lebanon, Syria, Jordan, Israel)	67	67	-	1.00	-
Fertile_Crescent_North-East (Turkey Karacadg, etc, etc)	36	36	-	1.00	-
Domesticated Emmer Wheat (*T. turgidum* ssp. *Dicocccum* and ssp. *isphahanicum*)					
Fertile Crescent (Turkey)	17	8	8	0.50	0.50
Fertile Crescent (Southern Levant, Lebanon-Syria-Jordan-Israel-Palestine)	25	16	1	0.94	0.06
Fertile_Crescent (general, not detailed)	11	8	0	1.00	0.00
Eastern Africa (Ethiopia-Kenia)	46	44	2	0.96	0.04
Southern Asia (Iran-Afghanistan)	39	33	4	0.89	0.11
Southern Asia (India)	18	15	3	0.83	0.17
Western Asia-Transcaucasia (Armenia-Georgia-Daghestan-Azerbaijan)	31	21	7	0.75	0.25
Western Asia- (Oman-Yemen-Kuwait-Saudi Arabia)	12	10	1	0.91	0.09
Northern Africa (Morocco-Tunisia)	8	5	3	0.63	0.38
Southern Europe (Greece-Albania-Serbia-Bosnia-Montenegro-Italy-Spain-Portugal)	66	56	10	0.85	0.15
Western Europe (Austria-Switzerland-Germany)	14	10	3	0.77	0.23
Eastern Europe (RussianFederation-Belarus-Poland-Ukraine)	22	19	3	0.86	0.14
Eastern Europe (Romania-Slovenia-Hungary-Czech Republic-Bulgaria)	21	19	2	0.90	0.10
Northern Europe (UK)	15	15	0	1.00	0.00
Central Asia (Kazakhstan-Uzbekistan)	2				
Unknown origin	**7**				
Durum wheat landraces (*T. turgidum* ssp. *Durum* and ssp. *aethiopicum*)					
Fertile Crescent (Turkey)	93	72	17	0.81	0.19
Fertile Crescent (Southern Levant, Lebanon-Syria-Jordan-Israel-Palestine-Iraq)	83	60	22	0.73	0.27
Fertile Crescent (Cyprus)	17	14	2	0.88	0.13
Fertile_Crescent (general, not detailed)					

Northern Africa (Egypt-Lybia-Tunisia-Algeria-Morocco)	137	77	53	0.59	0.41
Eastern Africa (Ethiopia-durum landraces)	172	111	50	0.69	0.31
Eastern Africa (Ethiopia-*T. aethiopicum*)	14	8	6	0.57	0.43
Eastern Europe (Romania-Bulgaria)	7	5	0	1.00	0.00
Eastern Europe (Russian Federation-Ukraine)	53	42	10	0.81	0.19
Central Asia (Kazhakstan-Uzbekistan)	7	5	2	0.71	0.29
Southern Europe (Greece-Albania-Croatia-Macedonia-Malta-Serbia-Italy-Spain-Portugal-	157	114	41	0.74	0.26
Western Asia-Transcaucasia (Armenia-Georgia-Azerbaijan)	20	17	2	0.89	0.11
Southern Asia (Iran-India)	29	25	4	0.86	0.14
Eastern Asia (China)	3	3	0	1.00	0.00
North America (USA-Canada)	10	7	3	0.70	0.30
Unknown origin	17	16	1	0.94	0.06
Durum wheat cultivars (*T. turgidum* ssp. *durum*)					
CIMMYT	48	29	16	0.64	0.36
ICARDA	83	30	51	0.37	0.63
Southern Europe (Italy-Spain)	140	35	91	0.28	0.72
Northern Africa (Morocco-Algeria)	17	7	9	0.44	0.56
Northern America (Canada-North Dakota)	46	8	36	0.18	0.82
Northern America (Desert Durum®, California-Arizona)	10	9	0	1.00	0.00
Western Europe (Austria-France)	45	10	26	0.28	0.72
South America (Argentina)	5	1	3	0.25	0.75
Ethiopia	24	7	9	0.44	0.56
Australia-New Zealand	6	1	5	0.17	0.83
Unknown origin	1	1	0	1.00	0.00
T. turgidum* subsp. *turgidum	8	6	2	0.75	0.25
T. turgidum* subsp. *turanicum	74	31	19	0.62	0.38
T. turgidum* subsp. *polonicum	22	15	5	0.75	0.25

Supplementary Table 16. *TdHMA3-B1a/b* allelic distribution of the Global Tetraploid wheat Collection (GTC) by genetic population structure as assessed based on *Fine STRUCTURE*/*ADMIXTURE* analysis.

***FineSTRUCTURE/ADMIXTURE* subpopulations**	***TdHMA 3-B1a (No.)***	***TdHMA 3-B1b (No.)***	***TdHMA3-B1a (freq.)***	***TdHMA 3-B1b (freq.)***
Wild Emmer Wheat				
*Q1*_Fertile_Crescent_Southern_Levant (Lebanon, Syria, Jordan, Israel)	67	0	1.000	0.000
*Q2*_Fertile_Crescent_North-East (Turkey Karacadg, etc, etc)	36	0	1.000	0.000
Wild Emmer Wheat *total*	103	0	1.000	0.000
Domesticated Emmer Wheat				
Q3 West Fertile Crescent/SouthernLevant Europe I	77	5	0.939	0.061
Q4 East Iran Transcaucasia Russia Asia	66	12	0.846	0.154
Q5 East Ethiopia India	62	10	0.861	0.139
Q6 West FertileCrescent/Turkey West-Balkans Russia	28	17	0.622	0.378
Q7 T. carthlicum East Transcaucasia Russia	15	2	0.882	0.118
Q8 West Fertile Crescent/SouthernLevant to Europe II	34	0	1.000	0.000
Domesticated Emmer Wheat *total*	282	46	0.860	0.140
Durum wheat landraces				
Q9 Ethiopia I	59	21	0.738	0.263
Q10 Ethiopia II	71	34	0.676	0.324
Q11 West Greece Western-Balkans	53	6	0.898	0.102
Q12 West Cyprus/Southern Levant NorthAfrica Spain Portugal	140	49	0.741	0.259
Q13 West Egypt Morocco Spain (including *T. turanicum* accessions)	22	64	0.256	0.744
Q14 T. turanicum	29	0	1.000	0.000
Q15 Fertile Crescent (Turkey, Syria, Cyprous, Iran, Iraq)	69	2	0.972	0.028
Q16 East Russian Federation	50	16	0.758	0.242
Q17 East Fertile Crescent Turkey Transcaucasia Russia Asia	95	17	0.848	0.152
Durum_Wheat_Landraces *total*	588	209	0.738	0.262
Durum wheat cultivars				
*Q18*_CIMMYT/CIMMYT-related Mediterranean_Germplasm (semi-dwarf, PPD insensitive)	100	122	0.450	0.550
Q19 NorthAmerica France Germplasm (*PPD* sensitive)	22	91	0.195	0.805
Q20 Italy/ICARDA NorthAfrican/Syrian founders' bred Germplasm	9	78	0.103	0.897
Durum_Wheat_Cultivars *total*	131	291	0.310	0.690

Supplementary Table 17. List of prolamin genes identified in durum wheat Svevo and wild emmer wheat Zavitan.

Protein Family	Family type	Chromosomal location	Number of genes	
			Svevo	**Zavitan**
Gliadin	α	6AS	35	16
		6BS	24	16
		U(1)	10	21
	γ	1AS	5	4
		1BS	6	6
		U	-	-
	ω	1AS	6	10
		1BS	7	4
		U	1	5
	δ	1AS	2	-
		1BS	1	1
		U	-	-
Glutenin	HMW	1AL	2	2
		1BL	2	2
		U	-	-
	LMW	1AS	3	4
		1BS	1	5
		U	6	1
Gliadin-like avenin		4AL	3	6
		7AS	5	6
		U	5	1
Total	124	107		

(1) indicates unmapped sequences.

Supplementary Table 18. Pseudogene basic metrics. A homology search with a combined query set of durum wheat (DW) and wild emmer wheat (WEW) canonical HC genes (131,023 transposon cleaned genes) resulted in the annotation of ~280,000 gene like sequences for DW and ~300,000 for WEW. About 90% of them are short gene fragments. Taking only those that cover at least 80% of their parent reduces the numbers to ~28,000 and ~27,000 respectively. The different pseudogene types are explained in the method section 1.2.2.

	Pseudogene features						**Pseudogene types (%)**				
	No.	Mean length (no.)	Mean identity to parent (%)	Mean coverage of parent (%)	Parent genes (no.)		Duplicated	Processed	Mono exon parent	Frag-mented	Chimeric
	All pseudogenes and gene fragments										
DW	279,773	289	89.2	27.7	36,314		24.1	2.6	17.4	54.8	1.0
WEW	299,528	280	89.2	26.3	37,903		24.4	2.7	16.0	55.9	1.0
	Pseudogenes with >= 80% coverage of their parent gene										
DW	28,106	756	92.7	94.7	9,952		47.2	1.6	45.8	3.8	1.5
WEW	27,388	742	92.7	94.6	9,559		47.2	1.6	45.9	3.8	1.4

Supplementary Table 19. Mapping populations used to support the wheat iSelect 90K SNP diversity analysis.

Cross	***T. turgidum* taxa**	**Population no. lines/type**	**Marker type**	**Mapped markers (no.)**
Tetraploid consensus map-2015[43]			SNP, DArT, SSR, STS	30,144
Svevo × Zavitan[1]	*T. durum* × *T. dicoccoides*	150 RIL	90 K array	14,086
Colosseo × Lloyd	*T. durum* × *T. durum*	84 RIL	90 K array	7,629
Meridiano × Claudio	*T. durum* × *T. durum*	90 RIL	90 K array	4,978
Simeto × Levante	*T. durum* × *T. durum*	89 RIL	90 K array	5,324
Mohawk × Cocorit C69	*T. durum* × *T. durum*	154 RIL	90 K array	6,387
Mohawk × Ardente	*T. durum* × *T. durum*	66 RIL	15 K array	1,563
Svevo × Ciccio	*T. durum* × *T. durum*	93 RIL	90 K array	7,622
Kofa × W9262-260D3	*T. durum* × *T. durum*	143 DH	90 K array	5,290
Svevo × Russello SG7	*T. durum* × *T. durum*	324 RIL	15 K array	2,051
Simeto × Molise Colli	*T. durum* × *T. dicoccum*	107 RIL	90 K array	13,237
Rusty × PI193883	*T. durum* × *T. dicoccum*	186 RIL	90 K array	14,908
Rusty × PI387696	*T. durum* × *T. carthlicum*	181 RIL	90 K array	12,515
Ben × PI41025	*T. durum* × *T. dicoccum*	194 RIL	9 K array	2,456
Svevo × Zavitan	*T. durum* × *T. dicoccoides*	150 RIL	90 K array	16,372
Latino × MG5323	*T. durum* × *T. dicoccum*	94 RIL	90 K array	12,478
Kofa × Svevo	*T. durum* × *T. durum*	249 RIL	SSR	247
Kofa × UC1113	*T. durum* × *T. durum*	93 RIL	SSR	207
Langdon × G18-16	*T. durum* × *T. dicoccoides*	152 RIL	SSR, DArT®	268

Supplementary Table 20. Pericentromeric chromosomal region sites of extended genetic diversity among the four-main tetraploid wheat germplasm pools (WEW: wild emmer wheat, DEW: domesticated emmer wheat, DWL: durum wheat landraces, DWC: durum wheat cultivars), as highlighted by the patterns of genome-wide genetic diversity (D) scan based on the wheat iSelect 90K SNP assay. In most cases, the regions correspond to depletion in genetic diversity passing from an ancestral to a more recently derived tetraploid germplasm.

a. Pericentromeric region sites of extended genetic diversity depletion (> 50 Mb).

Transition	**Chr.**	**Position** [1]	**Width**	**Diversity depletion**	**Gene content (high conf.)**
Germplasm transition [2]		**Mb** [3]	**Mb.**	**Direction** [4]	**no.**
DEW-DWL	1A	50-280	230	(-)	846
DEW-DWL	1B	90-230	140	(-)	544
DEW-DWL	2A	190-440	260	(-)	592
DWL-DWC	2A	190-440	260	(+)	592
DWL-DWC	3A	130-200	70	(-)	294
DEW-DWL	3B	80-210	130	(-)	741
DEW-DWL	3B	680-740	60	(-)	482
WEW-DEW	4A	200-250	50	(-)	73
WEW-DEW	4A	260-500	240	(-)	619
DWL-DWC	4B	50-110	60	(+)	416
WEW-DEW	4B	190-350	160	(-)	216
DEW-DWL	5A	120-200	80	(-)	202
DEW-DWL	5A	210-270	60	(-)	131
WEW-DEW	5B	90-270	180	(-)	644
WEW-DEW	6A	160-460	300	(-)	773
DEW-DWL	7A	310-460	150	(+)	362
DEW-DWL	7B	230-370	130	(-)	358

(1) Selection of chromosal regions based on (i) presence of a depletion of genetic diversity falling to or below D_{10Mb} 0.1 in at least one of the two germplasm pools; (ii) difference in diversity (ΔD_{10Mb}) ≥0.2; (iii) regions consistently extending for ≥20 Mb.

(2) WEW: wild emmer wheat, DEW: domesticated emmer wheat, DWL: durum wheat landraces, DWC: durum wheat cultivars.

(3) Diversity plots has been carried out with averaged SNP diversity computed on a 10 Mb step (D_{10Mb}).

(4) (-) decrease in genetic diversity from the ancestral to the derived subspecies/group; (+) increase in diversity.

b. Chromosomal region site of genetic diversity depletion from 20 to 50 Mb.

Transition	Chr.	Position [1]	Width	Diversity depletion	Gene content (high conf.)
Germplasm [2]		**Mb [3]**	**Mb.**	**Direction [4]**	**no.**
DWL-DWC	1A	310-340	30	(-)	142
DEW-DWL	1A	390-420	30	(-)	122
WEW-DEW	1B	240-260	20	(-)	49
DWL-DWC	1B	320-340	20	(-)	94
DEW-DWL	1B	400-440	40	(+)	239
WEW-DEW	2A	140-180	30	(-)	216
DWL-DWC	2A	510-550	40	(-)	229
DWL-DWC	2A	600-650	40	(-)	349
DEW-DWL	2A	660-690	30	(-)	348
DEW-DWL	2B	30-50	20	(-)	295
DEW-DWL	2B	150-180	30	(-)	269
DWL-DWC	2B	240-410	50	(-)	552
DEW-DWL	2B	410-460	50	(-)	267
DEW-DWL	2B	490-510	20	(-)	122
WEW-DEW	3A	210-240	30	(-)	111
WEW-DEW	3A	260-290	30	(-)	31
WEW-DEW	3A	410-430	20	(-)	68
DEW-DWL	3A	460-510	40	(-)	322
DWL-DWC	3A	580-600	20	(-)	186
DEW-DWL	3B	30-50	20	(-)	207
WEW-DEW	3B	80-120	40	(-)	255
DEW-DWL	3B	260-290	20	(-)	85
DWL-DWC	3B	290-330	40	(-)	42
DEW-DW	3B	350-380	30	(-)	78
WEW-DEW	3B	430-450	30	(-)	113
DEW-DWL	3B	510-530	20	(-)	145
DEW-DWL	3B	540-570	30	(-)	248
DWL-DWC	3B	660-740	80	(-)	644
DEW-DWL	4A	80-110	40	(-)	189
DEW-DW	4A	580-600	20	(-)	333
DWL-DWC	4B	50-110	60	(+)	416
DEW/DWL	4B	370-390	20	(-)	78
DEW-DWL	4B	440-480	40	(-)	205
WEW-DEW	4B	500-520	30	(-)	98
DEW-DWL	5A	20-40	20	(-)	170
DWL-DWC	5A	80-110	30	(+)	122
DEW-DWL	5A	360-400	40	(-)	285
DEW-DWL	5A	430-450	20	(-)	225
DEW-DWL	5A	460-490	30	(-)	222
DEW-DWL	5B	20-40	20	(-)	128
WEW-DEW	5B	290-330	40	(-)	235
DEW-DWL	5B	350-370	20	(-)	95
DWL-DWC	5B	490-530	40	(-)	405
DEW-DWL	5B	560-590	30	(-)	366
DEW-DWL	5B	660-680	20	(-)	252
DEW-DWL	6A	20-50	30	(-)	330
WEW-DEW	6A	120-150	30	(-)	154
Transition	**Chr.**	**Position [1]**	**Width**	**Diversity depletion**	**Gene content (high conf.)**
Germplasm [2]		**Mb [3]**	**Mb.**	**Direction [4]**	**no.**
DEW-DWL	6A	490-520	30	(-)	223
WEW-DEW	6B	60-80	20	(-)	164
WEW-DEW	6B	240-260	20	(-)	61
WEW-DEW	6B	280-310	30	(-)	68

WEW-DEW	6B	340-370	30	(-)	58
DEW-DWL	6B	480-520	40	(-)	233
DEW-DWL	7A	210-240	30	(-)	149
DEW-DWL	7A	460-530	70	(-)	368
DEW-DWL	7A	540-560	20	(-)	182
DWL-DWC	7B	90-110	20	(-)	122
DEW-DWL	7B	530-560	30	(-)	218

(1) Selection of chromosomal regions based on (i) presence of a depletion of genetic diversity falling to or below D_{10Mb} 0.1 in at least one of the two germplasm pools; (ii) difference in diversity (ΔD_{10Mb}) ≥0.2; (iii) regions consistently extending for ≥20 Mb.

(2) WEW: wild emmer wheat, DEW: domesticated emmer wheat, DWL: durum wheat landraces, DWC: durum wheat cultivars.

(3) Diversity plots has been carried out with averaged SNP diversity computed on a 10 Mb step (D_{10Mb}).

(4) (-) decrease in genetic diversity from the ancestral to the derived subspecies/group; (+) increase in diversity.

Supplementary Table 21. Haplotypes of homozygous $F_{2:3}$ lines derived from recombinant F_2 lines from the 8982-TL/H mapping population. Between 1 and 4 $F_{2:3}$ lines were classified into each haplotype group. For each haplotype group, the mean concentration of Cd in grain (ng g^{-1} ± standard error of the mean) and the classification of each haplotype as high or low grain Cd concentration genotypes are indicated. For each marker "a" represents the molecular variant from the low Cd parent, and "b" represents for the high Cd parent.

	***Cdu-B1* Markers**												
Number of homozygous $F_{2:3}$ families	***ScOPC20***	***Xusw49***	***Xusw59***	***Xusw50***	***Xusw51***	***Xusw52***	***Xusw15b***	***Xusw17***	***Xusw47***	***Xusw53***	***Xusw14***	**Grain Cd (ng g^{-1})**	**Cd Class**
Haplotype 1 (*n* = 1)	b	b	**B**	**b**	**b**	**b**	**b**	**b**	**b**	b	a	680	High
Haplotype 2 (*n* = 4)	b	b	**B**	**b**	**b**	**b**	**b**	**b**	**b**	a	a	623 ± 34	High
Haplotype 3 (*n* = 2)	b	b	**a**	**a**	**a**	**a**	**a**	**a**	**a**	a	a	220 ± 10	Low
Haplotype 4 (*n* = 4)	b	a	**a**	**a**	**a**	**a**	**a**	**a**	**a**	a	a	213 ± 21	Low
Haplotype 5 (*n* = 1)	a	a	**a**	**a**	**a**	**a**	**a**	**a**	**a**	a	b	230	Low
Haplotype 6 (*n* = 4)	a	a	**a**	**a**	**a**	**a**	**a**	**a**	**a**	b	b	195 ± 23	Low
Haplotype 7 (*n* = 4)	a	b	**B**	**b**	**b**	**b**	**b**	**b**	**b**	b	b	595 ± 89	High
8982-TL-H (*n* = 10)	b	b	**B**	**b**	**b**	**b**	**b**	**b**	**b**	b	b	690 ± 68	High
8982-TL-L (*n* = 10)	a	a	**a**	**a**	**a**	**a**	**a**	**a**	**a**	a	a	198 ± 23	Low

Supplementary Table 22. Primers used for localization of Cdu-B1, BTB/POZ-domain containing protein gene, cloning of HMA3 genes, and generation of yeast expression constructs.

Primer name	Primer (5'-3')	Notes
***Cdu-B1* fine mapping markers**		
Xusw49-F	CACCGAGCTGTCCTAATGAAG	STS-HRM Marker for LOC_Os03g53250
Xusw49-R	CTGCAGAAGTACTCTGGATCC	STS-HRM Marker for LOC_Os03g53250
Xusw50-F	TTCAGTGATAACTTACACCAG	STS-HRM Marker for LOC_Os03g53490
Xusw50-R	AGCTTCTTGCGTTCTTCCATC	STS-HRM Marker for LOC_Os03g53490
Xusw51-F	ATGGTTGGCTGTAGAACAAGG	STS-HRM Marker for LOC_Os03g53500
Xusw51-R	CTCACGCCGTGAGAACGTTAC	STS-HRM Marker for LOC_Os03g53500
Xusw52-F	TTCATTGTCAGATGATTCTGG	STS-HRM Marker for LOC_Os03g53530
Xusw52-R	CTTCCAGATCTTCACAAGCTT	STS-HRM Marker for LOC_Os03g53530
Xusw53-F	GATGAACCGCATATCCTTCCT	STS-HRM Marker for LOC_Os03g53700
Xusw53-R	CTCATTGTCACAAGCAATCAT	STS-HRM Marker for LOC_Os03g53700
CAPS markers		
Xusw14-F	TACAGCCGCTCAGTTGCTC	ESM Marker: *XBF474164*. Restriction site: *Bso*B1
Xusw14-R	CAACATATGTCTGGCCTACTACTCT	ESM Marker: *XBF474164*. Restriction site: *Bso*B1
Xusw17-F	TCCACCCCCTTCCATCCCTAT	ESM Marker: *XBF293297*. Restriction site: *Sbf*I
Xusw17-R	TTGTCCTGCGGCTTCACCATC	ESM Marker: *XBF293297*. Restriction site: *Sbf*I
Xusw15b-F	TATGTGTTGTGATTTGCTGAG	STS Marker: *Xusw15*. Restriction site: *Taq*1
Xusw15b-R	GAACCTTGGACGATTGCTAAC	STS Marker: *Xusw15*. Restriction site: *Taq*1
Xusw47-F	GCTAGGACTTGATTCATTGAT	ESM Marker: *XBF474090*. Restriction site: *Hpy*188I
Xusw47-R	AGTGATCTAAACGTTCTTATA	ESM Marker: *XBF474090*. Restriction site: *Hpy*188I
***HMA3-B1* marker**		
Xusw59-F or HMA3-B1-F	TTCTTGCTGTTCATCCGCCTG	297 bp amplicon (high-Cd), 280 bp amplicon (low-Cd)
Xusw59-R or HMA3-B1-R	AATACGGGACTGCGAGACGGC	297 bp amplicon (high-Cd), 280 bp amplicon (low-Cd)
***HMA3* cDNA cloning**		
HMA3-F1	CTCGTCGTGCTCAACAGC	3'-RACE Primary
3'-RACE-QO	CCAGTGAGCAGAGTGACG	3'-RACE Primary
HMA3-F2	TCGCTGGAGATGAGAAGG	3'-RACE Nested
3'-RACE-QI	GAGGACTCGAGCTCAAGC	3'-RACE Nested
HMA3-F3	GGCTCTGTCGTTGACTTATTTGC	FL-CDS *HMA3-A1* and *HMA3-B1*
HMA3-R3	TGCAAGCTTCCCTTGCTACC	FL-CDS *HMA3-A1* and *HMA3-B1*
HMA3-D1-R1	CGGCACAAAATATACAAAGAGGAC	FL-CDS *HMA3-D1* (common wheat)
***HMA3-B1* gDNA cloning/sequencing**		
HMA3-F1	ATGGGCGGCGGCGAGTCGTAC	Amplicon HMA3-F1-R1
HMA3-R1	GTGGTGAAGAGGAAGACGATG	Amplicon HMA3-F1-R1
HMA3-F2	GACATCAACATCCTCATGCTT	Amplicon HMA3-F2-R2
HMA3-R2	CCATTGTCCTCACGGCGATGT	Amplicon HMA3-F2-R2
HMA3-F3	ACATCGCCGTGAGGACAATGG	Amplicon HMA3-F3-R3
HMA3-R3	TTTGCTCTCGATGCTTGAGAT	Amplicon HMA3-F3-R3
HMA3-F4	ATCTCAAGCATCGAGAGCAAA	Amplicon HMA3-F4-R4
HMA3-R4	TGAGGATGTCGCTGGACATGA	Amplicon HMA3-F4-R4
HMA3-F5	TCATGTCCAGCGACATCCTCA	Amplicon HMA3-F5-R5
HMA3-R5	GCCGACACGCAGCTCGATGAA	Amplicon HMA3-F5-R5
HMA3-F6	CGTGCTCAACAGCATGCTGCT	Amplicon HMA3-F6-R6
HMA3-R6	AAGATCGAACGGCCATTCTTC	Amplicon HMA3-F6-R6
Yeast expression constructs (restriction sites underlined, linker sequence lowercase)		
y*TtHMA3*-ORF2-*Bam*HI	ACAGGATCCAAAAATGTTGTTACGTGGTATCGCTG	HMA3-B1b ORF2
Linker-EcoRI	AAAGAATTCtaaaccagcaccgtcacc	HMA3-B1b ORF2
YCF1-BamHI-S	ACAGGATCCAGAAAATGGCTGGTAATCTTGTTTC	YCF1 ORF
YCF1-XhoI-AS	AAACTCGAGTGTAAGGGGTATGTGGTGAGG	YCF1 ORF
yEGFP-BamHI- F1	ATTGGATCCTTAggtgacggtgctggttta	yEGFP ORF
yEGFP-EcoRI-D1	AAAGAATTCAGTGGCGCGCCTTATTTG	yEGFP ORF
HMA3-GFP constructs by Overlap Extension (OE; restriction sites underlined, linker sequence lowercase)		
yTtHMA3-BamHI-A1	ACAGGATCCAAAAATGATGGGTGGTG	HMA3-B1a, HMA3-B1b (ORF1) OE fragment A-B
yTtHMA3-B1t-link	taaaccagcaccgtcaccTAATGCGGTA	HMA3-B1b (ORF1) OE fragment A-B
yTtHMA3-ORF2-BamHI	ACAGGATCCAAAAATGTTGTTACGTGGTATCGCTG	HMA3-B1b (ORF2) OE fragment A-B
yTtHMA3-B1-link	taaaccagcaccgtcaccTAATGGAGAAC	HMA3-B1a, HMA3-B1b (ORF2) OE fragment A-B
yEGFP-link-C1	TTAggtgacggtgctggttta	HMA3-GFP OE fragment C-D
yEGFP-EcoRI-D1	AAAGAATTCAGTGGCGCGCCTTATTTG	HMA3-GFP OE fragment C-D
BTB/POZ-domain containing protein markers targeting a loss-of-function variant present in Svevo (high resolution melting detection technique or CAPS using *DdeI* enzyme)		
TRIDC5BG059880-F1	CGTTGATGTTGAATTCCGAGT	
TRIDC5BG059880-R1	ATGCTGAGCCATTGCAGATT	
TRIDC5BG059880-F2	CAGGAGACACCATCCATCCT	
TRIDC5BG059880-R2	GGAGCTCGTATCGACTTGCT	

Supplementary Table 23. Average genetic diversity in the distal, highly-recombining and centromeric, recombination-depleted chromosome regions for the 14 DW chromosomes. WEW: wild emmer wheat, DEW: domesticated emmer wheat, DWL: durum wheat landraces, DWC: durum wheat cultivars.

Region	Physical interval	Gene content (high confidence)	Gene density		D_{10Mb}			
	Mb	No.	No./Mb	No./cM	DWC	DWL	DEW	WEW
1A-R1	0-27.7	375	13.507	10.322	0.218	0.289	0.336	0.230
1A-R2	499.4-538.9	481	12.145	19.263	0.183	0.251	0.316	0.253
1A-R3	566.6-585.3	254	13.637	12.300	0.239	0.275	0.251	0.263
1B-R1	0-33.9	432	12.716	15.749	0.307	0.384	0.265	0.307
1B-R2	595.5-681.1	896	10.466	17.613	0.350	0.223	0.319	0.329
2A-R1	0-64.2	955	14.875	14.009	0.240	0.246	0.267	0.273
2A-R2	667.1-775.4	1,527	14.085	25.881	0.237	0.259	0.324	0.263
2B-R1	0-119.9	1,235	10.297	18.272	0.274	0.267	0.301	0.329
2B-R2	712.6-790.4	905	11.636	20.738	0.232	0.286	0.271	0.276
3A-R1	0-77.9	1,016	13.034	19.879	0.265	0.279	0.257	0.321
3A-R2	596.7-746.7	1,704	11.364	26.398	0.230	0.280	0.314	0.305
3B-R1	0-61.1	854	13.974	17.758	0.251	0.283	0.321	0.279
3B-R2	742.6-836.5	1,192	12.696	18.819	0.225	0.279	0.318	0.300
4A-R1	0-61.8	586	9.4885	13.820	0.173	0.215	0.291	0.322
4A-R2	592.6-648.5	729	13.026	15.590	0.211	0.241	0.280	0.246
4A-R3	660.9-736.9	784	10.316	26.667	0.312	0.294	0.279	0.302
4B-R1	0-35.8	422	11.779	11.155	0.334	0.301	0.278	0.290
4B-R2	642.4-676.3	420	12.383	11.179	0.220	0.224	0.285	0.297
5A-R1	0-39.7	1,776	44.701	46.504	0.164	0.195	0.366	0.286
5A-R2	503.4-669.2	1,955	11.797	18.954	0.247	0.263	0.279	0.277
5B-R1	0-46.1	369	8.0042	10.743	0.170	0.219	0.314	0.321
5B-R2	536.0-701.4	1,794	10.851	24.861	0.252	0.254	0.284	0.318
6A-R1	0-38.9	590	15.162	14.279	0.328	0.208	0.245	0.158
6A-R2	547.9-615.7	906	13.368	18.354	0.256	0.301	0.256	0.251
6B-R1	0-64.0	723	11.293	16.342	0.275	0.327	0.268	0.339
6B-R2	646.0-698.6	610	11.592	21.532	0.231	0.239	0.319	0.296
7A-R1	0-82.9	1,042	12.559	15.396	0.280	0.287	0.285	0.265
7A-R2	634.8-728.0	1,029	11.041	13.803	0.247	0.258	0.313	0.291
7B-R1	0-59.3	513	8.649	10.779	0.224	0.298	0.310	0.311
7B-R2	613.8 -722.9	1,035	9.484	17.017	0.309	0.313	0.329	0.313
Average					***0.250***	***0.268***	***0.295***	***0.287***
1A-C	51.7-360.6	1,246	4.032	418.121	0.088	0.109	0.346	0.301
1B-C	88.7-314.3	795	3.524	196.296	0.139	0.156	0.242	0.235
2A-C	201.2-559.9	1,077	3.002	252.817	0.206	0.077	0.133	0.165
2B-C	209.7-441.5	860	3.709	256.716	0.08	0.266	0.380	0.237
3A-C	109.5-479.9	1,191	3.215	319.303	0.183	0.173	0.112	0.182
3B-C	195.2-504.8	1,217	3.931	199.836	0.150	0.192	0.250	0.224
4A-C	133.2-532.5	1,134	2.840	606.417	0.056	0.079	0.081	0.220
4B-C	71.2-467.9	1,328	3.349	389.443	0.210	0.199	0.220	0.364
5A-C	46.8-385.4	746	2.203	162.174	0.165	0.153	0.252	0.351
5B-C	90.0-393.0	1,359	4.485	395.058	0.082	0.115	0.146	0.313
6A-C	107.8-480.6	1,110	2.976	365.132	0.128	0.094	0.126	0.384
6B-C	161.8-439.4	869	3.130	334.231	0.196	0.210	0.187	0.272
7A-C	249.6-510.4	849	3.255	758.036	0.301	0.241	0.243	0.218
7B-C	135.5-411.6	1,024	3.708	307.507	0.135	0.187	0.360	0.291
Average					***0.151***	***0.161***	***0.220***	***0.269***

Supplementary Table 24. Chromosome location of *Cdu-B1* markers and *HMA3* on chromosome 5B of durum wheat cv. Svevo and wild emmer wheat accession Zavitan.

Query Sequence[1]	**Position in Chromosome 5B (start - end)**[2]	
	Wild emmer wheat (Zavitan)	Durum wheat (cv. Svevo)
Xusw49	572890476 – 572890758	563586136 – 563586418
Xusw59	**573244560 – 573244842**	**563900630 – 563900929**
OsHMA3	**573244658 – 573247449**	**562900745 – 563903535**
BdHMA3	**573244662 – 573247307**	**563900749 – 563903393**
Xusw50	574645059 – 574645227	565279790 – 565279958
Xusw51	574653163 – 574654179	565287597 – 565288613
Xusw52	575064421 – 575064537	565697079 – 565697195
Xusw15b	575835057 – 575835318	566480780 – 566481041
Xusw47	577150868 – 577151318	567822312 – 567822762
Xusw53	577183551 – 577183949	567855129 – 567855527

(1) Query sequences in bold text overlap on the physical map of *Cdu-B1*.

(2) Positions were determined by BLASTN with an expect value < 1e-10.

Captions for Supplementary Data Sets S1 to S13

Supplementary Data Set 1. Inventory of linkage mapping and GWAS-QTLs mapped on tetraploid wheat and projected on the Svevo genome assembly.

Supplementary Data Set 2. Population structure of the Global Tetraploid wheat Collection (GTC, 1,856 accessions) assessed by ADMIXTURE/fineSTRUCTURE analysis and their allelic score at *HMA3-B1* assessed by *Xusw59* perfect marker.

Supplementary Data Set 3. Population structure of the Global Tetraploid wheat Collection (GTC) assessed with four model- and non model- based quantitative clustering methods: DAPC-K means, DAPC-Ward's, sNMF, ADMIXTURE.

Supplementary Data Set 4. Whole-genome diversity reduction and selection signal analysis in tetraploid wheat.

Supplementary Data Set 5. Detailed F_{st} analysis of the tetraploid diversity panel for chromosome 5B.

Supplementary Data Set 6. Gene content and functional analysis of the *Cdu-B1* region in chromosome 5B as defined by the linkage disequilibrium interval in the durum wheat landrace and cultivar panels ± 2 Mb.

Supplementary Data Set 7. Experimental conditions and tissue used for the RNA-Seq and smallRNA-Seq expression analysis.

Supplementary Data Set 8. Sequencing data on 216 experimentally validated wheat genes.

Supplementary Data Set 9. Unigene clusters made with both wild emmer wheat Zavitan and durum wheat Svevo HC genes.

Supplementary Data Set 10. Genome-wide atlas of 597 putative high impact variants differentiating between wild emmer wheat accession Zavitan and durum wheat cultivar Svevo by chromosome.

Supplementary Data Set 11. Detailed genetic maps used to project iSelect Illumina 9K, 15K, 90K SNP, DArT, SSR, EST-SSR, and STS markers on the Svevo genome assembly.

Supplementary Data Set 12. Locus identifiers and protein sequences used for the HMA phylogenetic analysis.

Supplementary Data Set 13. List of miRNAs detected in all or in one unique sequencing pool.

References

1. Avni, R. *et al.* Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* **357**, 93-97 (2017).
2. Gremme, G., Brendel, V., Sparks, M.E. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.* **47**, 965–978 (2005).
3. Bennetzen, J.L. *et al.* Reference genome sequence of the model plant Setaria. *Nat Biotechnol.* **30**, 555–561 (2012).
4. The International Brachypodium Initiative, Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763–768 (2010).
5. Ouyang, S. *et al.* The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.* **35**, D883–D887 (2007).
6. Paterson, A.H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
7. Krishnakumar, V. *et al.* Araport: the Arabidopsis Information Portal. *Nucleic Acids Res.* **43**, D1003–D1009 (2015).
8. The UniProt Consortium, UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204-212 (2015).
9. Pertea, M., Kim, D., Pertea, G.M., Leek, J.T. & Salzberg, S.L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protocols* **11**, 1650–1667 (2016).
10. Clavijo, B.J. *et al.* 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 (2017).
11. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
12. Trapnell, C. *et al.*, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–78 (2012).
13. Sabot, F. *et al.* Updating of transposable element annotations from large wheat genomic sequences reveals diverse activities and gene associations. *Mol. Genet. Genomics* **274**, 119–130 (2005).
14. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
15. Kurtz, S. *et al.* A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* **9**, 517 (2008).
16. Spannagl, M. *et al.* PGSB PlantsDB : updates to the database framework for comparative plant genome research. *Nucleic Acids Res.* **44**, 1141-1147 (2016).
17. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
18. Eddy, S.R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
19. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10-12 (2011).
20. Axtell, M.J. ShortStack: Comprehensive annotation and quantification of small RNA genes. *RNA* **19**, 740-751 (2013).
21. Cagrici, H.B., Alptekin, B. & Budak, H. RNA sequencing and co-expressed long non-coding RNA in modern and wild wheats. *Scientific Rep.* **7**, 10670 (2017).
22. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
23. Nussbaumer, T. *et al.* MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic Acids Res.* **41**, e166 (2013).

24. Sun, L. *et al.* Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* **41**, e166 (2013).
25. Zhang, Z. & Wood, W.I. A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics* **19**, 307–308 (2003).
26. Steuernagel, B., Jupe, F., Witek, K., Jones, J.D. & Wulff. B.B. NLR-parser: rapid annotation of plant NLR complements. *Bioinformatics* **31**, 1665–1667 (2015).
27. Hackenberg, M. *et al. CpGcluster*: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics* **7**, 446 (2006).
28. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018, (2011).
29. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257-258 (2007).
30. Wilderman, P.R. & Peters, R.J. A single residue switch converts abietadiene synthase into a pimaradiene specific cyclase. *J. Am. Chem. Soc.* **129**, 15736–15737 (2007).
31. Zerbe, P. & Bohlmann, J. Plant diterpene synthases: exploring modularity and metabolic diversity for bioengineering. *Trends Biotechnol.* **33**, 419–428 (2015).
32. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12**: 656-664 (2002).
33. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).
34. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).
35. Gish, W. (1996-2003) http://blast.wustl.edu
36. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573-580 (1999).
37. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Reviews Genet.* **8**, 973–982 (2007).
38. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595 (2010).
39. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
40. McLaren, W. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
41. Avni, R. *et al.* Ultra-dense genetic map of durum wheat × wild emmer wheat developed using the 90K iSelect SNP genotyping assay. *Mol. Breed.* **34**, 1549–1562 (2014).
42. Wu, Y., Bhat, P.R., Close, T.J. & Lonardi, S. Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *Genet.* **4**, e1000212 (2008).
43. Maccaferri, M. *et al.* 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**, 648–663 (2015).
44. Chen, J. & Gupta, A.K. Parametric statistical change point analysis. *Birkhauser, Boston* (2012).
45. Choulet, F. *et al.* Structural and functional partitioning of bread wheat chromosome 3B. *Science* **345**, 1249721 (2014).
46. Arcade, A. *et al.* BioMercator: integrating genetic maps and QTL towards discovery of candidate genes. *Bioinformatics* **20**, 2324-2326 (2004).
47. Darvasi, A. & Soller, M. A simple method to calculate resolving power and confidence interval of QTL map location. *Behav. Genet.* **27**, 125-132 (1997).

48. Wang, S. *et al.* Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array *Plant Biotechnol. J.* **12**, 787-796 (2014).
49. Liu, S. *et al.* Validation of Chromosomal Locations of 90K Array Single Nucleotide Polymorphisms in US Wheat *Crop Sci.* **56**, 364–373 (2016).
50. Nei, M. Analysis of gene diversity in subdivided populations. *Proc. Nat. Acad. Sci. USA* **70**, 3321-3323 (1973).
51. Excoffier, L. & Lischer, H.E.L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Res.* **10**, 564-567 (2010).
52. Goudet, J. Hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* **5**, 184–186 (2005).
53. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
54. Zeileis, A. & Grothendieck, G. zoo: S3 infrastructure for regular and irregular time series. *J. Statistic Software* **14**, 1–27 (2005).
55. Zheng, X., & Weir, B.S. Eigen analysis of SNP data with an identity by descent interpretation. *Theoret. Pop. Biol.* **107**, 65-76 (2016).
56. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
57. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
58. Lawson, D.J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
59. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-59 (2000).
60. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC genetics* **11**, 94 (2010).
61. Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G. & François, O. Fast and efficient estimation of individual ancestry coefficients. *Genetics* **196**, 973-983 (2014).
62. Sargolzaei, M., Chesnais, J.P. & Schenkel, F.S. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* **15**, 478 (2014).
63. Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M. & Franke, A. A comprehensive evaluation of SNP genotype imputation. *Human Genet.* **125**, 163–171 (2009).
64. Hancock D.B. *et al.* Assessment of genotype imputation performance using 1000 Genomes in African American studies. *PLoS One* **7**:e50610 (2012).
65. Huson, D.H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
66. Kilian, B. *et al.* Molecular diversity at 18 loci in 321 wild and 92 domesticate lines reveal no reduction of nucleotide diversity during Triticum monococcum (einkorn) domestication: Implications for the origin of agriculture. *Mol. Biol. Evol.* **24**, 2657–2668.
67. Özkan, H., Willcox, G., Graner, A., Salamini, F. & Kilian B. Geographic distribution and domestication of wild emmer wheat (*Triticum dicoccoides*). *Genet. Resour. Crop Evol.* **58**, 11–53 (2011).
68. Luo, M.C. *et al.* The structure of wild and domesticated emmer wheat populations, gene flow between them, and the site of emmer domestication. *Theor. Appl. Genet.* **114**, 947-59 (2007).

69. Weir, B.S. & Cockerham, C.C. Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).
70. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393-402 (2010).
71. Sabeti, P.C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
72. Fariello, M.I., Boitard, S., Naya, H., SanCristobal, M. & Servin, B. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* **193**, 929–941 (2013).
73. Hufford, M.B. *et al.* Comparative population genomics of maize domestication and improvement. *Nat Genet*. **44**, 808–811 (2012).
74. Paradis, E. Pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics* **26**, 419-20 (2010).
75. Szpiech, Z.A. & Hernandez, R.D. An efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.* **31**, 2824–2827 (2014).
76. Scheet, P. & Stephens, M. A fast and flexible method for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629-644 (2006).
77. Khvorykh, G. inzilico/imputeqc v1.0.0 (2018). GitHub repository, https://github.com/inzilico/imputeqc.
78. Wiebe, K. *et al.* Targeted mapping of *Cdu1*, a major locus regulating grain cadmium concentration in durum wheat (*Triticum turgidum* L. var *durum*). *Theor. Appl. Genet.* **121**, 1047-1058 (2010).
79. Clarke, J.M., Leisle, D., DePauw, R.M. & Thiessen, L.L. Registration of five pairs of durum wheat genetic stocks near-isogenic for cadmium concentration. *Crop Sci.* **37**, 297 (1997).
80. Jordan, K.W. *et al.* A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biol.* **16**, 48 (2015).
81. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
82. Zimin, A.V. The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *Gigascience* **6**, 1-7 (2017).
83. Harris, N.S. & Taylor, G.J. Cadmium uptake and partitioning in durum wheat during grain filling. *BMC Plant Biol.* **13**, 103 (2013).
84. Solovyev, V.V. Statistical approaches in eukaryotic gene prediction. *Handbook of statistical genetics, 3rd edn. Wiley-Interscience,* London (2007).
85. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).
86. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
87. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577 (2007).
88. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704 (2003).
89. Le, S.Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).

90. Lefort, V., Longueville, J.E. & Gascuel, O. SMS: Smart Model Selection in PhyML. *Mol. Biol. Evol.* **34**, 2422-2424 (2017).
91. Smith, S.A. & Dunn, C.W. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* **24**, 715-716 (2008).
92. Dräger, D.B. *et al.* Two genes encoding Arabidopsis halleri MTP1 metal transport proteins co-segregate with zinc tolerance and account for high MTP1 transcript levels. *Plant J.* **39**, 425–439 (2004).
93. Heckman, K.L. & Pease, L.R. Gene splicing and mutagenesis by PCR-driven overlap extension. *Nat. Protocols* **2**, 924-932 (2007).
94. Bublitz, M., Morth, J.P. & Nissen, P. P-type ATPases at a glance. *J. Cell Sci.* **124**, 2515–2519 (2011).
95. Mumberg, D., Muller, R. & Funk, M. Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. *Gene* **156**, 119–122 (1995).
96. Gietz, R.D. & Schiestl, R.H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protocols* **2**, 31-34 (2007).
97. Sherman, F. Getting started with yeast. *Methods Enzymol.* **350**, 3-41 (2002).
98. Pronk, J.T. Auxotrophic yeast strains in fundamental and applied research. *Appl. Environ. Microb.* **68**, 2095–2100 (2002).
99. Toussaint, M. &Conconi, A.High-throughput and sensitive assay to measure yeast cell growth: a bench protocol for testing genotoxic agents. *Nature Prot.* **1**, 1922-1928 (2006).
100. Sheff, M.A. & Thorn, K.S. Optimized cassettes for fluorescent protein tagging in *Saccharomyces cerevisiae*. *Yeast* **21**, 661–670 (2004).
101. Clarke, J.M., Norvell, W.A., Clarke, F.R. & Buckley, W.T. Concentration of cadmium and other elements in the grain of near-isogenic durum lines. *Can. J. Plant Sci.* **82**, 27–33 (2002).
102. Parker, D.R. & Norvell, W.A. Advances in solution culture methods for plant mineral nutrition research. *Advan. Agron.* **65**, 151-213 (1999).
103. Zhang, M. *et al.* Preparation of megabase-sized DNA from a variety of organisms using the nuclei method for advanced genomics research. *Nature Prot.* **7**, 467-78 (2012).
104. The International Wheat Genome Sequencing Consortium (IWGSC), A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* **345**, 1251788 (2014).
105. Mascher, M. *et al.* A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**, 427–433 (2017).
106. Mayer, K.F.X. Unlocking the barley genome by chromosomal and comparative genomics. Plant Cell **23**,1249–1263 (2011).
107. Dubcovsky, J. & Dvorak, J. Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* **316**,1862-1866 (2007).
108. Thibaud-Nissen, F., Ouyang, S. & Robin Buell, C. Identification and characterization of pseudogenes in the rice gene complement. *BMC Genomics* **10**, 317 (2009).
109. Wang, L. *et al.* Genome-wide survey of pseudogenes in 80 fully re-sequenced *Arabidopsis thaliana* accessions. *PLoS ONE* **7**, e51769 (2012).
110. Kan, Y.C. *et al.*, Transcriptome analysis reveals differentially expressed storage protein transcripts in seeds of Aegilops and wheat. *J. Cereal Sci.* **44**, 75–85 (2006).

111. Chen, X.Y. *et al.* Genetic characterization of cysteine-rich type-b avenin-like protein coding genes in common wheat. *Scientific Rep.* **6**, 30692 (2016).
112. Dong, L. *et al.* Rapid evolutionary dynamics in a 2.8-Mb chromosomal region containing multiple prolamin and resistance gene families in *Aegilops tauschii*. *Plant J.* **87**, 495-506 (2016).
113. Wang, D.W. *et al.* Genome-wide analysis of complex wheat gliadins, the dominant carriers of celiac disease epitopes. *Scientific Rep.* **7**, 44609 (2017).
114. Anderson, O.D., Litts, J.C., Gautier, M.F. & Greene, F.C. Nucleic acid sequence and chromosome assignment of a wheat storage protein gene. *Nucleic Acids Res.* **12**, 8129–8144 (1984).
115. Huo, N. *et al.* Dynamic Evolution of α-gliadin prolamin gene family in homeologous genomes of hexaploid wheat. *Scientific Reports* **8**, 5181(2018).
116. Huo, N. *et al.* New insights into structural organization and gene duplication in a 1.75-Mb genomic region harboring the α-gliadin gene family in *Aegilops tauschii*, the source of wheat D genome. *Plant J.* **92**, 571-583 (2017).
117. Marone, D., Russo, M.A., Laidò, G., De Leonardis, A.M. & Mastrangelo, A.M. Plant nucleotide binding site-leucine-rich repeat (NBS-LRR) genes: active guardians in host defense responses. *Int. J. Mol. Sci.* **14**, 7302–7326 (2013).
118. Lee, H.A. & Yeom, S.I. Plant NB-LRR proteins: tightly regulated sensors in a complex manner. *Brief. Funct. Genomics* **14**, 233-242 (2015).
119. Bouktila, D. *et al.* Full-genome identification and characterization of NBS-encoding disease resistance genes in wheat. *Mol. Genet. Genomics* **290**, 257–271 (2015).
120. Marchler-Bauer, A. *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures *Nucleic Acids Res.* **45**, D200-D203 (2017).
121. Wright, E.S. Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. *R Journal* **8**, 352-359 (2016).
122. Nützmann, H.E., Huang, A. & Osbourn, A. Plant metabolic clusters – from genetics to genomics. *New Phytol.* **211**, 771-789 (2016).
123. Nützmann, H.W. & Osbourn, A. Gene clustering in plant specialized metabolism. *Curr. Opin. Biotechnol.* **26**, 91–99 (2014).
124. Boycheva, S., Daviet, L., Wolfender, J.L. & Fitzpatrick, T.B. The rise of operon-like gene clusters in plants. *Trends Plant Sci.* **19**, 447–459 (2014).
125. Timmis, J.N., Ayliffe, M.A., Huang, C.Y. & Martin, W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* **5**, 123–35 (2004).
126. Jonczyk, R. *et al.* Elucidation of the final reactions of DIMBOA-glucoside biosynthesis in maize: characterization of Bx6 and Bx7. *Plant Physiol.* **146**, 1053–1063 (2008).
127. Stein, L.D. The generic genome browser: a building block for a model organism system database. *Genome Res.* **12.10**, 1599-1610 (2002).
128. Stein, L.D. Using GBrowse 2.0 to visualize and share next-generation sequence data. *Brief Bioinformatics* **14.2**, 162-171 (2013).
129. Winfield, M.O. *et al.* High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotech. J.* **14**, 1195-1206 (2016).
130. Rimbert, H. *et al.* High throughput SNP discovery and genotyping in hexaploid wheat. *PLoS One* **13**, e0186329 (2018).

131. Allen, A.M. *et al.* Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivum*). *Plant Biotech. J.* **15**, 390-401 (2017).
132. Malomane, D.K. *et al.* Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. *BMC Genomics* **5**, 19, 22 (2018).
133. Qanbari, S. & Simianer, H. Mapping signatures of positive selection in the genome of livestock. *Livestock Sci.* **166**, 133-143 (2014).
134. Haudry, A. *et al.* Grinding up wheat: a massive loss of nucleotide diversity since domestication. *Mol. Biol. Evol.* **24**, 1506-1517 (2007).
135. Akhunov, E.D. *et al.* Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. *BMC Genomics* **11**, 702 (2010).
136. Badaeva, E.D., Loskutov, I.G., Shelukhina, O.Y. & Pukhalsky, V.A. Cytogenetic analysis of diploid series of avena L. containing As genome. *Russ. J. Genet.* **41**, 1428–1433 (2005).
137. Maccaferri, M., Sanguineti, M.C., Donini, P. & Tuberosa, R. Microsatellite analysis reveals a progressive widening of the genetic basis in the elite durum wheat germplasm. *Theor. Appl. Genet.* **107**, 783–797 (2003).
138. Maccaferri, M., Sanguineti, M.C., Noli, E. & Tuberosa, R. Population structure and long-range linkage disequilibrium in a durum wheat elite collection. *Mol. Breed.* **15**, 271–289 (2005).
139. De Vita, P. *et al.* 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 (2007).
140. Soriano, J.M., Villegas, D., Aranzana, M.J., García del Moral, L.F. & Royo, C. Genetic structure of modern durum wheat cultivars and mediterranean landraces matches with their agronomic performance. *PLoS One* **11**, e0160983 (2016).
141. Kabbaj, H. *et al.* 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 (2017).
142. Oliveira, H.R. *et al.* Wheat in the Mediterranean revisited – tetraploid wheat landraces assessed with elite bread wheat Single Nucleotide Polymorphism markers. *BMC Genet.* **15**, 54 (2014).
143. Özkan, H. *et al.* A reconsideration of the domestication geography of tetraploid wheats. *Theor. Appl. Genet.* **110**, 1052–1060 (2005).
144. Badaeva, E.D. et al. Chromosomal passports provide new insights into diffusion of emmer wheat. *PLoS One* **10**, e0128556 (2015).
145. Jorgensen, C. *et al.* A high-density genetic map of wild emmer wheat from the Karaca Dağ region provides new evidence on the structure and evolution of wheat chromosomes. *Front. Plant Sci.* **8**, 1798 (2017).
146. Marone, D. *et al.* A high-density consensus map of A and B wheat genomes. *Theor. Appl. Genet.* **125**, 1619-1638 (2012).
147. Maccaferri, M. et al. A consensus framework map of durum wheat (*Triticum durum* Desf.) suitable for linkage disequilibrium analysis and genome-wide association mapping. *BMC Genomics* **15**, 873 (2014).
148. Sela, H. *et al.* Rapid linkage disequilibrium decay in the Lr10 gene in wild emmer wheat (*Triticum dicoccoides*) populations. *Theor. Appl. Genet.* **122**, 175-187 (2011).
149. Morrell, P.L., Toleno, D.M., Lundy, K.E. & Clegg, M.T. Low levels of linkage disequilibrium in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite high rates of self-fertilization. *Proc. Natl. Acad. Sci. USA* **102**, 2442-2447 (2005).

150. Yan, L. et al. Positional cloning of the wheat vernalization gene VRN1. *Proc. Natl. Acad. Sci. USA* **100**, 6263-6268 (2003).
151. Faris, J.D., Zhang, Q., Chao, S., Zhang, Z. & Xu S.S. 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 (2014).
152. Faris, J.D., Zhang, Z. & Chao, S. Map-based analysis of the tenacious glume gene *Tg-B1* of wild emmer and its role in wheat domestication. *Gene* **542**, 198-208 (2014).
153. Xu, Q. et al. PCR-based markers for identification of HMW-GS at *Glu-B1x* loci in common wheat. *J. Cereal Sci.* **47**, 394-398 (2008).
154. Zhang, Z. *et al.* Duplication and partitioning in evolution and function of homoeologous Q loci governing domestication characters in polyploid wheat. *Proc. Natl. Acad. Sci. USA* **108**, 18737-18742 (2011).
155. Zhang, Y.I. *et al.* Analysis of the functions of TaGW2 homoeologs in wheat grain weight and protein content traits. *Plant J.* **94**, 857-866 (2018).
156. Su, Z., Hao, C., Wang, L., Dong, Y. & Zhang, X. Identification and development of a functional marker of TaGW2 associated with grain weight in bread wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* **122**, 211-23 (2011).
157. Wilhelm, E.P., Turner, A.S. & Laurie, D.A. Photoperiod insensitive *Ppd-A1a* mutations in tetraploid wheat (*Triticum durum* Desf.). *Theor. Appl. Genet.* **118**, 285–294 (2009).
158. Hu, M.J. et al. Cloning and characterization of *TaTGW-7A* gene associated with grain weight in wheat via SLAF-seq-BSA. *Front. Plant Sci.* **7**, 1902 (2016).
159. Zhang, W. & Dubcovsky, J. Association between allelic variation at the *Phytoene synthase 1* gene and yellow pigment content in the wheat grain. *Theor. Appl. Genet.* **116**, 635–645 (2008).
160. Oladzad-Abbasabadi A. *et al.* Identification and validation of a new source of low grain cadmium accumulation in durum wheat. *G3: Genes, Genomes, Genetics* **8**, 923-932 (2018).
161. Argüello, J.M., Eren, E. & Gonzalez-Guerrero, M. The structure and function of heavy metal transport P_{1B}-ATPases. *BioMetals* **20**, 233–248 (2007).
162. Axelsen, K.B. & Palmgren, M.G. Inventory of the superfamily of P-type ion pumps in Arabidopsis *Plant Physiol.* **126**, 696–706 (2001).
163. Pedersen, B.P. *et al.* Large scale identification and categorization of protein sequences using structured logistic regression. *PLoS One* **9**, e85139 (2014).
164. Thever, M.D. & Saier, M.H. Bioinformatic characterization of P-type ATPases encoded within the fully sequenced genomes of 26 eukaryotes. *J. Membr. Biol.* **229**, 115-130 (2009).
165. Williams, L.E. & Mills, R.F. P_{1B}-ATPases--an ancient family of transition metal pumps with diverse functions in plants. *Trends Plant Sci.* **10**, 491-502 (2005).
166. Argüello, J.M. Identification of ion-selectivity determinants in heavy-metal transport P_{1B}-type ATPases. *J. Membr. Biol.* **195**, 93-108 (2003).
167. Baekgaard, L. *et al.* A combined Zinc/Cadmium sensor and Zinc/Cadmium export regulator in a heavy metal pump. *J. Biol. Chem.* **285**, 31243–31252 (2010).
168. Eren, E., Kennedy, D.C., Maroney, M.J. & Arguello, J. M. A novel regulatory metal binding domain is present in the C terminus of Arabidopsis Zn^{2+}-ATPase HMA2. *J. Biol. Chem.* **281**, 33881–33891 (2006).
169. Verret, F. *et al.* Heavy metal transport by AtHMA4 involves the N-terminal degenerated metal binding domain and the C-terminal His_{11} stretch. *FEBS Lett.* **579**, 1515–1522 (2005).

170. Morth, J.P. *et al.* A structural overview of the plasma membrane Na^+,K^+-ATPase and H^+-ATPase ion pumps. *Nat. Rev. Mol. Cell Biol.* **12**, 60-70 (2011).
171. Ueno, D. *et al.* Gene limiting cadmium accumulation in rice. *Proc. Nat. Acad. Sci. USA* **107**, 16500–16505 (2010).
172. Yan, J. *et al.* A loss-of-function allele of *OsHMA3* associated with high cadmium accumulation in shoots and grain of Japonica rice cultivars. *Plant Cell Environ.* **39**, 1941–1954 (2016).
173. Gardarin, A. *et al.* Endoplasmic reticulum is a major target of cadmium toxicity in yeast. *Mol. Microbiol.* **76**, 1034-1048 (2010).
174. Harris, S.N. & Taylor, J.G. Remobilization of cadmium in maturing shoots of near isogenic lines of durum wheat that differ in grain cadmium accumulation. *J. Exp. Bot.* **52**, 1473-1481 (2001).
175. Gravot, A. *et al.* AtHMA3, a plant P_{1B}-ATPase, functions as a Cd/Pb transporter in yeast. *FEBS Lett.* **561**, 22–28 (2004).
176. Morel, M. *et al.* AtHMA3, a P_{1B}-ATPase allowing Cd/Zn/Co/Pb vacuolar storage in Arabidopsis. *Plant Physiol.* **149**, 894–904 (2009).
177. Ueno, D. *et al.* Elevated expression of *TcHMA3* plays a key role in the extreme Cd tolerance in a Cd-hyperaccumulating ecotype of *Thlaspi caerulescens*. *Plant J.* **66**, 852–862 (2011).
178. Sasaki, A., Yamaji, N. & Ma, J.F. Overexpression of OsHMA3 enhances Cd tolerance and expression of Zn transporter genes in rice. *J. Exp. Bot.* **65**, 6013-6021 (2014).
179. Chao, D.Y. *et al.* Genome-wide association studies identify heavy metal ATPase3 as the primary determinant of natural variation in leaf cadmium in *Arabidopsis thaliana*. *PLoS Genet.* **8**, e1002923 (2012).
180. Satoh-Nagasawa, N. *et al.* Mutations in rice (*Oryza sativa*) heavy metal ATPase 2 (OsHMA2) restrict the translocation of zinc and cadmium. *Plant Cell Physiol.* **53**, 213-24 (2012).
181. Yamaji, N., *et al.* Preferential delivery of zinc to developing tissues in rice is mediated by P-type heavy metal ATPase OsHMA2. *Plant Physiol.* **162**, 927-939 (2013).
182. Khan, M.A., Castro-Guerrero, N. & Mendoza-Cozatl, D.G. Moving toward a precise nutrition: preferential loading of seeds with essential nutrients over non-essential toxic elements. *Front. Plant Sci.* **5**, 51 (2014).
183. Dong, Z. *et al.* Ideal crop plant architecture is mediated by *tassels replace upper ears1*, a BTB/POZ ankyrin repeat gene directly targeted by TEOSINTE BRANCHED1. *Proc. Nat. Acad. Sci*. USA **114**, E8656-E8664 (2017).
184. Schnable, P.S. *et al*. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112-1115 (2009).
185. International Barley Genome Sequencing Consortium *et al.* A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**, 711-716 (2012).
186. Bauer, E. *et al.* Towards a whole-genome sequence for rye (*Secale cereale* L.). *Plant J.* **89**, 853-869 (2017).
187. Krasileva, K.V. *et al.* Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biol.* **14**, R66 (2013).
188. Zou, H. *et al.* Transcriptome profiling of wheat glumes in wild emmer, hulled landraces and modern cultivars. *BMC Genomics* **16,** 777 (2015).
189. Zhang, W., Gianibelli, MC., Rampling, LR. & Gale, KR. Characterisation and marker development for low molecular weight glutenin genes from *Glu-A3* alleles of bread wheat (*Triticum aestivum*. L). *Theor. Appl. Genet.* **108**, 1409-1419 (2004).
190. Aoki, N. *et al.* Three sucrose transporter genes are expressed in the developing grain of hexaploid wheat. *Plant Mol. Biol.* **50**, 453–462 (2002).

191. Xie, D.W. *et al.* Identification of the trehalose-6-phosphate synthase gene family in winter wheat and expression analysis under conditions of freezing stress. *J. Genet.* **94**, 55-65 (2015).
192. Alvarez, M.A., Tranquilli, G., Lewis, S., Kippes, N. & Dubcovsky, J. Genetic and physical mapping of the earliness per se locus *Eps-Am1* in *Triticum monococcum* identifies *EARLY FLOWERING 3* (*ELF3*) as a candidate gene. *Funct. Integr. Genomics.* **16**, 365–382 (2016).
193. Zikhali, M., Wingen, L.U. & Griffiths, S. 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**, 287-299 (2016).
194. Jiang, Q. *et al.* The wheat (*T. aestivum*) *sucrose synthase 2* gene (*TaSus2*) active in endosperm development is associated with yield traits. *Funct. Integr. Genomics* **11**, 49-61 (2011).
195. Zhang, Y., Miao, X., Xia, X. & He, Z. 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**, 855-66 (2014).
196. Ma, D., Yan, J., He, Z., Wu, L. & Xia, X. Characterization of a cell wall invertase gene *TaCwi-A1* on common wheat chromosome 2A and development of functional markers. *Mol. Breed.* **29**, 43–52 (2012).
197. Takenaka, S. & Kawahara, T. 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 (2012).
198. Jiang, Y. *et al.* 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 (2014).
199. Pearce, S. *et al.* Molecular characterization of *Rht-1* dwarfing genes in hexaploid wheat. *Plant Physiol.* **157**, 1820-31 (2011).
200. Shorinola, O. *et al.* 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**, 4169–4178 (2016).
201. Chu, C.G. *et al.* 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 (2011).
202. Gu, Y.Q. *et al.* Genomic organization of the complex alpha-gliadin gene loci in wheat. *Theor. Appl. Genet.* **109**, 648–657 (2004).
203. Uauy, C., Distelfeld, A., Fahima, T., Blechl, A. and Dubcovsky, J. A *NAC* gene regulating senescence improves grain protein, zinc, and iron content in wheat. *Science* **314**, 1298-1301 (2006).
204. Zhang, W. *et al.* Identification and characterization of *Sr13*, a tetraploid wheat gene that confers resistance to the Ug99 stem rust race group. *Proc. Natl. Acad. Sci. USA* **114**, E9483-E9492 (2017).
205. Yan, L. *et al.* The wheat and barley vernalization gene *VRN3* is an orthologue of *FT*. *Proc. Natl. Acad. Sci. USA* **103**, 19581-19586 (2006).
206. Kalaipandian, S. *et al.* Overexpression of *TaCML20*, a calmodulin-like gene, enhances water soluble carbohydrate accumulation and yield in wheat. *Physiol. Plant.* doi: 10.1111/ppl.12786 (2018).