

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☐ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

This work has taken advantage from the novel genome sequence of durum wheat cv. Svevo generated by the authors and of existing dataset downloaded from public databases as detailed in the Supplementary note (section 1: Additional Materials and Methods). Swiss-Prot (version 02-15-10), Arabidopsis Araport 11 (version 201606), a TrEMBL (version 02-15-10) and UniProt databases were the main sources.

Data analysis

This work employed a vast number of software and data analyses, each software is cited in the Supplementary note (section 1: Additional Materials and Methods) along with the corresponding reference or website information. In summary (numbers refer to the references listed in the Supplementary note file):

1- Genome assembly was carried out using the proprietary software package DenovoMAGIC2TM (NRGene, Nes Ziona, Israel) as described⁸.

2- Genome annotations was supported with the following tools:

- HISAT2 (version 2.0.4)⁹ to align multiple sets of RNA-seq data to the assemblies;
- Stringtie (version 1.2.3)⁹ to assemble mapped reads into transcript sequences for each dataset separately;
- GMAP11 (version 06/30/2016) to align all sequences to the assemblies;
- Cuffcompare from Cufflinks software suite¹² for transcript predictions;
- Transdecoder package (version 3.0.0) to extract the longest open reading frames for each transcript sequence and to translate them into predicted protein sequences;
- AHRD tool (Automated Assignment of Human Readable Descriptions, <https://github.com/groupschoof/AHRD>, version 3.3.3) to annotate gene functions;
- BUSCO (Benchmarking Universal Single-Copy Orthologs) tool (version 2, Embryophyta odb9)¹⁴ to determine the abundance of strongly conserved genes in the sets of all annotated genes;
- Tallymer¹⁵ to calculate the Basic k-mer defined repetitivity;
- Vmatch (www.vmatch.de) for transposons detection and classification by a homology search against the REdat_9.7_Triticeae and the

PGSB transposon library¹⁶;

- LTRharvest¹⁷ to identify full-length LTR retrotransposons (fl-LTRs);
- NLR-annotator version 0.7 pipeline (<https://github.com/steuernb/NLR-Annotator26>) to annotate the loci associated with Nucleotide-Binding Leucine-Rich Repeat domains (NLRs);
- CpGCluster²⁷ to detect CpG islands;
- Find Individual Motif Occurrences (FIMO)²⁸ to annotate Transcription factors binding sites (TFBS);
- Blast2GO PRO to assign GO terms;
- Bioconductor package GOstats²⁹ (version 2.42.0) for GO enrichment call;
- CD-HIT^{33,34} (version 4.6.5) and Tandem Repeats Finder³⁶ to identify domains with high copy number;
- OrthoMCL (version 2.0, www.orthomcl.org) to search for gene families.

3- Genetic maps were produced following a pipeline including: 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-map⁴² (<http://www.mstmap.org/>), MetaQTL analysis and projection of QTLs to DW assembly was calculated using the BiomeRCator version 4.2 software⁴⁶

4- For the analysis of genetic diversity the raw data from Illumina genotyping 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). Pairwise LD values were estimated by means of the `snpgdsLDMat` function in the R package `SNPRelate`⁵³, 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. Phylogenetic networks were computed using `SplitsTree4` version 4.14.665 while `Fst` statistics at each locus was computed by R package `pegas`⁷⁴. `fastPHASE` v1.4.876 and R package `imputeq`⁷⁷ were used to reconstruct the haplotypes from SNP data.

5- For construction of the refined interval for Cdu-B1 markers from the array were mapped to the genome of Svevo by GMAP¹¹. Markers uniquely mapping to chromosome 5B were used for Single Marker Regression analysis using Windows QTL Cartographer (<https://brwebportal.cos.ncsu.edu/qtlcart/index.php>). Sequence reads from exome capture experiment were processed by Trimmomatic version 0.3281 and aligned to the genome of Svevo using Novoalign version 3.02.05 (www.novocraft.com/products/novoalign). Variants were called using the SamTools version 1.2.139.

6- For comparative sequence analysis of HMA gene, protein sequences or translated CDS sequences for P1B-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 settings⁸⁵. Gaps and poorly aligned regions were removed from the multiple sequence alignment (MSA) by Gblocks version 0.91b86 using less stringent selection criteria⁸⁷; 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. Phylogenetic tree was reconstructed using the maximum-likelihood method with PhyML version 3.188

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

A dedicated durum wheat genome browser with all genome information presented in this manuscript (sequences, annotations, markers, passport data, etc.) is available at <http://www.interomics.eu/durum-wheat-genome>. Svevo genome and gene model were submitted to EBI-ENA under the study PRJEB22687. Svevo genome, gene model, SNP and QTL data and passport information of the GTC are available at GrainGenes (https://wheat.pw.usda.gov/GG3/jbrowse_Durum_Svevo). RNA-Seq and miRNA datasets can be downloaded at the SRA database under accessions SRP149116 (study: PRJNA473404). Durum wheat cv. Svevo GFF3 and VCF annotation files can be downloaded at: <https://figshare.com/>, accession: 10.6084/m9.figshare.6984035. The wild emmer wheat acc. Zavitan gene model version 2 is available at: <https://doi.org/10.5447/ipk/2019/0>. The accessions used in this study are available on request and based on a Material Transfer Agreement from the corresponding author (LC).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

The article refers to 3 main experiments.

The first deals with the genome sequence assembly of a modern cultivar of durum wheat (Svevo), representing the first high quality cultivar released in Europe, for which a high resolution genetic map (Svevo x Zavitan) was already available. The Svevo genome was sequenced at 279X coverage (= 3.35 Tb of NGS data) using a combination of five size-selected genomic DNA libraries of insert size ranging from 450 bp (paired-end) up to 10 kb (mate pairs), in balanced relative amount. This sequencing data, analysed with the software package DenovoMAGIC2TM (available to NRGene, Nes Ziona, Israel), ensured to assemble 10.45 Gb of unique Svevo sequences, structured into contigs of N50-length = 56.20 Kb, scaffolds of N50-length = 5.972 Mb, and 14 Hi-C ordered pseudomolecules including gaps for 149.190 Mb only (1.42%). We provide a Svevo "golden-standard" assembly of quality comparable to that reported for the tetraploid wild emmer (Avni et al. 2017), Ae. tauschii (Luo et al. 2017), and Chinese Spring hexaploid wheat (IWGSC, 2018). The empirical and statistical experimental procedure that led to the choice of the sequencing depth is reported in Avni et al. (2017). The final assembly provides 66,559 high-confidence and 303,404 low-confidence genes. The completeness of gene content was then validated based on gene representativeness from independent public databases (98.1% of BUSCO genes, 97.7% of a dataset of experimentally determined genes, 92.3% of Triticeae reference proteins). It also provides 51,077 high-quality resolved full-length LTR retrotransposons.

The second experiment describes the diversity in tetraploid wheats using 1,854 accessions, this is the largest diversity panel reported so far for tetraploid wheat. It represents all known AABB-genome subspecies and samples all regions where tetraploid wheat (both wild and cultivated) is spread. In particular, the collection samples wild emmer wheat (WEW), domesticated emmer wheat (DEW), durum-related subspecies, durum wheat landraces (DWL) and durum wheat cultivars (DWC) in balanced counts. The collection has been assembled from world-wide researchers long time-involved in diversity and genomics of tetraploid wheat, and supplemented by additional 490 accessions specifically sampled from domestication areas. The whole-genome genetic diversity survey of the four main domestication-related sets (WEW, DEW, DWL, DWC) of tetraploid wheat was carried out based on the well-established and highly informative iSelect 90K wheat SNP assay. After filtering for uniqueness and information content, it provided a total of 17,416 transcript-associated SNPs (1.66/Mb or 16.6/10Mb). Based on the linkage disequilibrium decay rate in the four germplasm sets considered, the number of SNPs was considered adequate to assess the relative genetic diversity losses/gains among domestication-related sets.

The third experiments describes the cloning of a locus controlling Cd accumulation. This work was carried out using a large panel of genetic materials including 3 segregating populations: the high-density Svevo x Zavitan reference mapping population, a large F2 population of 5,081 individuals suitable for fine-mapping at 0.1 cM scale and a panel of cultivars and breeding lines. Further, a survey of the causal polymorphism was extended to the whole diversity panel (1,854 worldwide accessions).

Data exclusions

No specific data were excluded.

Replication

The analysis and assembly of genomic Illumina NGS data involved initial several standard quality-control and filtering steps as described in Supplemental Materials & Methods.

At the core of the analysis, the software package DenovoMAGIC2TM (available to NRGene, Nes Ziona, Israel) uses very stringent parameters and thresholds for all contig assembling and scaffolding steps as described in Supplemental Materials & Methods. In details, elongated or stitched reads were produced upon an overlap threshold of 10 bp minimum. Initial contig assembly was carried out based on De Bruijn graphs at very high kmer = 191 bp. The assembler has special features for the identification of unique non-repetitive sequences, for resolving repeats and, in the scaffolding step, to identify erroneous connections. Stringent parameters were also used for scaffold ordering and linking based on the mate-pairs data.

Both high-density genetic map and chromosome conformation capture sequencing (Hi-C) were used for scaffold validation, correction and re-ordering of final pseudomolecules. Only intra-chromosomal Hi-C links below certain distance-thresholds were used. Available IWGSC contigs and POPSEQ for hexaploid wheat were used as further validation/correction steps.

The annotation pipeline is robust as it has been developed and tested on the most advanced Triticeae genome assemblies, including barley, Aegilops tauschii and hexaploid wheat, combining in-silico bioinformatics and gene expression data. Several quality-thresholds were applied to assemble RNA-seq reads into transcripts. Finally the whole-assembly validation was carried out by considering its gene representativeness when tested against three independent and validated gene/protein sets (BUSCO, lab-validated gene set, Triticeae protein set).

Micro-RNAs, long non coding RNAs, transposons and LTR retrotransposons, prolamin and NB-LRR gene families were annotated based on homology searches against Triticeae and more-comprehensive databases. Functional non tandem duplicated gene clusters were found based on GO-enrichment specific tests.

Genetic mapping was based on 17 interconnected maps, mostly based on the same iSelect 90K wheat SNP Illumina array platform, for which a common pipeline for genotype call and genetic map was adopted, starting from the original Illumina iSelect raw data files. The pipeline included several quality-check and filtering steps. Based on the available reference sequences of SNP and other markers, the genetic maps were anchored to the Svevo genome assembly using the most complete, gap-free high-density Svevo x Zavitan unique reference binned map. This genetic-to-physically anchored map was also used in the process of QTL projection on the final Svevo assembly.

The highly-robust and widely used iSelect 90K wheat SNP Infinium array-based genotyping platform was used for the diversity survey. The genotyping projects were all carried out in high-standard Illumina-certified genotyping labs. The raw iSelect Illumina SNP genotyping data obtained for the wide tetraploid wheat diversity panel from various genotyping projects were processed with the same pipeline described above for the genetic mapping data. Only unique Mendelian-segregating loci, whose genetic map location was coincident with the corresponding Svevo physical address, were used for further analysis.

In all the Illumina SNP genotyping projects including both genetic mapping and diversity accession panels, common reference genotypes were used as internal quality-standard checks. For this purpose, the well-known "Cappelli", "Langdon" and "Svevo" reference DNA, together with those from several parents of mapping populations were used.

After passport analysis and clustering of genetic profiles, the 2,558 tetraploid diversity panel accessions that were initially assembled from various genotyping projects were reduced to 1,854 non-redundant tetraploid accessions. Passport info were mostly useful to predict cases of duplicated samples or high genetic relationships, thus representing an additional quality check control.

The population genetic structure of the tetraploid diversity panel was assessed with two independent well-known softwares (ADMIXTURE and fineSTRUCTURE), based on replicated analysis runs as recommended by software analysis instruction manuals. The two softwares gave bi-directional comparable, highly-complementary results.

The whole-genome genetic diversity survey based on the four main domestication-related sets of tetraploid wheat was carried out using a

SNP dataset filtered for uniqueness of map location (dingle-locus Mendelian loci), absence of sigletons and double-singletons, and, in order 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).

All the fine-mapping and positional cloning steps of the Cadmium transporter locus, including several genotyping and phenotyping steps, were carried out using standard replications. Same applies to the candidate gene expression analysis, including both in-planta and in yeast expression and complementation experiments. Importantly, near-isogenic line stocks, collections of diverse germplasms of adequate counts, and yeast mutant stocks were used as to most adequate materials to guarantee highly-repeatable and reliable results.

In conclusion, the experimental design, methods of replication and validation, and statistical methods used meet the standards outlined by Nature Genetics. All attempts at replication were successful. Several experiments required multiple replications and in each case the number of replications is described.

Randomization

When generating the Illumina NGS Svevo genome sequencing data, library construction was carried out assuring the high quality and integrity of the original genomic DNA used. DNA random fragmentation (shearing/sonication) and tagmentation for Nextera libraries were carried out by following carefully optimized Illumina protocols that ensured the random representativeness of genomic sequences for all paired-end and mate-pairs libraries. Concerning the distribution of sequencing depth among the different types of libraries, 123X coverage was dedicated to PE libraries of 450bp-insert size (half of total coverage), this was required to obtain highly-accurate contig assemblies. Coverage dedicated to each of the 750bp-PE and to the three Nextera Mate Pairs libraries (3kb, 5kb, 10kb) was equally partitioned (38-41X each) in order to ensure that balanced sequence information of different insert size was conveyed to the scaffold assembler.

During the selection of the accessions of the tetraploid diversity panel, great attention was given to sample accessions from each of the four main wheat germplasm groups (WEW, DEW, DWL, DWC) to ensure an accurate germplasm representativeness. Within each group, care in uniform sub-regional sampling and further germplasm bank's accession availability inspection was considered to represent all main sub-areas related to diffusion and domestication. Within sub-areas, random sampling in balanced numbers was carried out, after passport inspections of seedbank available accessions, in order to limit as much as possible sampling of duplicated or highly related accessions. Prior to the final diversity analysis, great care was taken in the joint inspection of passport and molecular information available in order to filter out clearly duplicated, highly similar and redundant accessions. Germplasm structure was assessed in great details, based on two independent dedicated software.

As to the whole-genome scan for differential gene diversity among the four main germplasm subgroups, given the predominantly descriptive objective of this analysis, namely describing the diversity present among germplasm collections, selective sweep tests and corrections for population structure were not applied, except for the initial withdrawal of the Ethiopian DEW and DWL accessions, two groups clearly distinct from the main bulks of European, Mediterranean and Central Asian germplasm. A similar approach was applied for the Cadmium-transporter allele survey distribution (carried out for all 1,854 accessions available).

As to the fine-mapping and positional cloning steps of the Cadmium transporter locus, including several genotyping and phenotyping steps, standard randomization best practices were used across all experiments.

Blinding

Not applicable. This research did not include experiments with observer biases.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging