Article

Einkorn genomics sheds light on history of the oldest domesticated wheat

https://doi.org/10.1038/s41586-023-06389-7

Received: 16 October 2022

Accepted: 29 June 2023

Published online: 2 August 2023

Open access

Check for updates

Hanin Ibrahim Ahmed[1,2,11], Matthias Heuberger[3,11], Adam Schoen[4,11], Dal-Hoe Koo[5], Jesus Quiroz-Chavez[6], Laxman Adhikari[1,2], John Raupp[5], Stéphane Cauet[7], Nathalie Rodde[7], Charlotte Cravero[7], Caroline Callot[7], Gerard R. Lazo[8], Nagarajan Kathiresan[9], Parva K. Sharma[4], Ian Moot[4], Inderjit Singh Yadav[4], Lovepreet Singh[4], Gautam Saripalli[4], Nidhi Rawat[4], Raju Datla[10], Naveenkumar Athiyannan[1,2], Ricardo H. Ramirez-Gonzalez[6], Cristobal Uauy[6], Thomas Wicker[3], Vijay K. Tiwari[4]✉, Michael Abrouk[1,2]✉, Jesse Poland[1,2]✉ & Simon G. Krattinger[1,2]✉

Einkorn (*Triticum monococcum*) was the first domesticated wheat species, and was central to the birth of agriculture and the Neolithic Revolution in the Fertile Crescent around 10,000 years ago[1,2]. Here we generate and analyse 5.2-Gb genome assemblies for wild and domesticated einkorn, including completely assembled centromeres. Einkorn centromeres are highly dynamic, showing evidence of ancient and recent centromere shifts caused by structural rearrangements. Whole-genome sequencing analysis of a diversity panel uncovered the population structure and evolutionary history of einkorn, revealing complex patterns of hybridizations and introgressions after the dispersal of domesticated einkorn from the Fertile Crescent. We also show that around 1% of the modern bread wheat (*Triticum aestivum*) A subgenome originates from einkorn. These resources and findings highlight the history of einkorn evolution and provide a basis to accelerate the genomics-assisted improvement of einkorn and bread wheat.

Einkorn (*T. monococcum*) was the first wheat species that humans domesticated around 10,000 years ago in the Fertile Crescent, a region in the Near East that is often referred to as the Cradle of Civilization[1,2]. Wild einkorn was an ingredient of the oldest known bread-like products, baked by hunter-gatherers in modern-day Jordan four millennia before the dawn of agriculture[3]. Einkorn had a pivotal role in the establishment of agriculture in the Fertile Crescent and it is the only diploid wheat species ($2n = 2x = 14$, A^mA^m genome) of which both wild and domesticated forms exist. A noticeable morphological difference between wild and domesticated einkorn is the grain dispersal system. Wild einkorn has a fragile rachis that facilitates seed dispersal, whereas the rachis in domesticated einkorn is non-brittle[4]. Einkorn is closely related to *Triticum urartu*, the A genome donor of tetraploid durum (*Triticum durum*) and hexaploid bread wheats (*T. aestivum*)[5]. In contrast to *T. urartu*, wild and domesticated einkorn have a long history of cultivation and human selection in diverse environmental conditions, which makes einkorn a valuable source of genetic variation for wheat breeding. Multiple natural and artificial einkorn introgressions into bread wheat containing agriculturally important genes have been described[6–10]. Population genetic analyses indicate that wild einkorn clusters into three distinct groups (races α, β and γ) and point to a region around the Karacadağ mountains in Southeastern Turkey as the site of einkorn domestication[11–17].

Here we establish and analyse a comprehensive set of genomic resources for einkorn, including de novo annotated chromosome-scale reference assemblies of one wild and one domesticated einkorn accession, as well as whole-genome sequencing of an einkorn diversity panel. Our results unravel the complex evolutionary history of einkorn and offer insights into the genome dynamics of Triticeae, including the centromere structure, while establishing valuable resources that augment the genomic toolbox for wheat improvement.

Chromosome-scale einkorn assemblies

We generated reference assemblies of two einkorn accessions using a combination of PacBio circular consensus sequencing[18], optical mapping[19] and chromosome conformation capture[20] (Extended Data Table 1, Supplementary Table 1 and Supplementary Fig. 1). TA10622 is a domesticated einkorn landrace (*T. monococcum* L. subsp. *monococcum*) with non-brittle rachis that was collected in Albania at the beginning of the twentieth century. Wild einkorn accession TA299 (*T. monococcum* L. subsp. *aegilopoides*; race α) was collected during an

[1]Plant Science Program, Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. [2]Center for Desert Agriculture, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. [3]Department of Plant and Microbial Biology, University of Zurich, Zurich, Switzerland. [4]Department of Plant Science and Landscape Architecture, University of Maryland, College Park, MD, USA. [5]Wheat Genetics Resource Center and Department of Plant Pathology, Kansas State University, Manhattan, KS, USA. [6]John Innes Centre, Norwich Research Park, Norwich, UK. [7]INRAE, CNRGV French Plant Genomic Resource Center, Castanet-Tolosan, France. [8]Crop Improvement and Genetics Research Unit, Western Regional Research Center, Agricultural Research Service, United States Department of Agriculture, Albany, CA, USA. [9]KAUST Supercomputing Core Lab (KSL), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. [10]Global Institute for Food Security, University of Saskatchewan, Saskatoon, Saskatchewan, Canada. [11]These authors contributed equally: Hanin Ibrahim Ahmed, Matthias Heuberger, Adam Schoen. ✉e-mail: vktiwari@umd.edu; michael.abrouk@kaust.edu.sa; jesse.poland@kaust.edu.sa; simon.krattinger@kaust.edu.sa

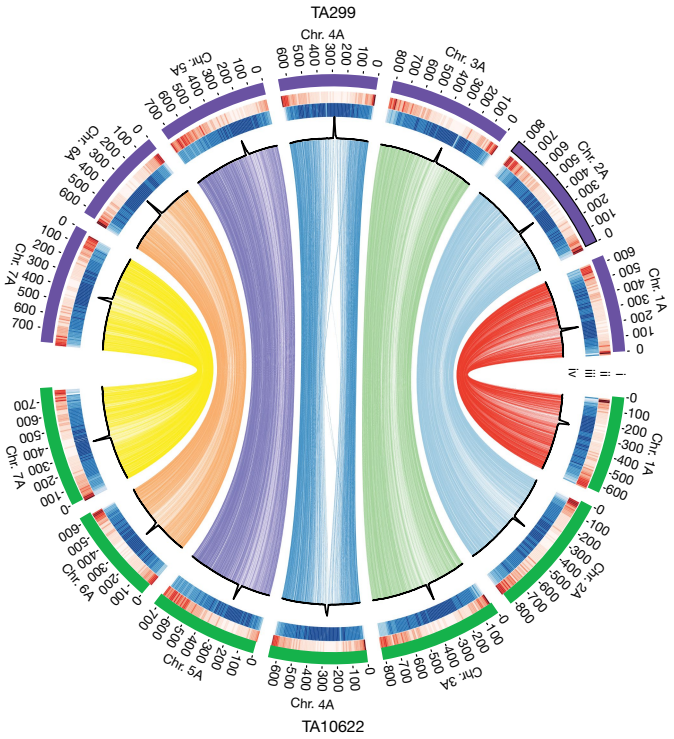


Fig. 1 | Einkorn genome structure and functional features. Circos plot showing synteny between the assemblies of wild einkorn (TA299) and domesticated einkorn (TA10622). The tracks depict structural and functional features of the two einkorn reference assemblies. The number and length of pseudomolecules (i), gene density along pseudomolecules (ii), repeat density along pseudomolecules (iii) and CENH3 ChIP–seq read coverage along pseudomolecules (iv) are shown. Peaks in each pseudomolecule define the centromeres (iv). The lines in the inner circle represent 17,586 orthologous high-confidence genes between TA299 and TA10622. Only relationships between the same chromosomes are shown.

expedition in 1972 in northern Iraq[21] and has a brittle rachis. Assembly integrities were verified using an einkorn genetic map (Supplementary Tables 2 and 3). We observed a high degree of collinearity across the two sets of pseudomolecules (Fig. 1 and Supplementary Fig. 2) and between the two einkorn assemblies and the bread wheat A subgenome (Supplementary Fig. 3). The most obvious exceptions were the well-described rearrangements of bread wheat chromosome 4A, which experienced inversions and translocations in polyploid wheat[22]. We annotated 32,230 and 32,090 high-confidence gene models on the 7 pseudomolecules of TA299 and TA10622, respectively (BUSCO scores of 99.2% for TA299 and 99.4% for TA10622) (Supplementary Tables 4 and 5).

Previous short-read-based wheat assemblies often did not resolve large tandem and segmental duplications[23,24]. On chromosome 4A of TA10622, we identified an approximately 1 Mb tandem duplication (Extended Data Fig. 1a,b). The two segments were 1,058,744 bp and 1,040,693 bp in length, with 98.3% sequence identity on average (Extended Data Fig. 1b,c). The most plausible explanation for the origin of this large tandem duplication is an unequal recombination between retrotransposons[25] (Extended Data Fig. 1d). Each of the duplicated segments contained one high-confidence gene encoding a MADS-box transcription factor (*Tm.TA10622.r1.4AG0101640* and *Tm.TA10622.r1.4AG0101820*) (Extended Data Fig. 1b), which have important roles in the regulation of plant growth and development[26–29]. The corresponding segment in the wild TA299 accession was not duplicated, carrying instead a single copy of the MADS-box gene. We estimated the presence/absence of the tandem duplication across a diversity panel comprising 218 wild and cultivated einkorn accessions (Extended Data Fig. 1e). All but one wild einkorn accession belonging to races α and γ had one copy of the corresponding segment, whereas most domesticated einkorn accessions carried the tandem duplication. Wild einkorn accessions belonging to race β, the proposed progenitor of domesticated einkorn, had two copies for this segment, indicating that the tandem duplication predated domestication (Extended Data Fig. 1e). The tandem duplication was specific to einkorn and was not found in bread wheat.

In addition to the megabase-sized tandem duplication on chromosome 4A, we successfully resolved large duplications in the highly repetitive centromere regions. The centromere of chromosome 2A in TA299 contained two large duplications of around 700 kb each, one of which was followed by an inversion. Both events can be traced to unequal recombination between *RLG_Cereba* retrotransposons (Extended Data Fig. 2). These results highlight the superiority of long-read-based genome assemblies to resolve and study the dynamics of large tandem duplications.

Analysis of complete einkorn centromeres

Centromeres are critical regions of eukaryotic chromosomes for the assembly of the spindle apparatus and cell division[30]. Although centromere function is conserved across eukaryotes, there is considerable variation in the underlying genomic sequences. Centromeres often remain as persistent gaps in genome assemblies owing to their complex, highly repetitive sequences. Many centromeres are colonized by centromeric satellite repeats of 130–180 bp in size (corresponding to the length of DNA wrapped around a nucleosome) and/or centromere-specific transposable elements (TEs). Centromere identity is defined epigenetically by the presence of centromere-specific CENH3 histone variants (CENP-A in mammals)[31].

We performed chromatin immunoprecipitation with sequencing (ChIP–seq) analysis of CENH3, which identified one distinct region per chromosome in TA10622 and TA299 with high read coverage, indicating the positions of functional centromeres (Fig. 2a, Extended Data Fig. 3, Supplementary Fig. 4 and Supplementary Tables 6 and 7). In addition to the high CENH3 coverage, we found that einkorn centromeres are local minima of CpG methylation and H3K4me3 histone modification (Supplementary Figs. 5 and 6). Crucial for our analysis was that centromeric regions in both accessions were assembled contiguously without sequence gaps (Supplementary Fig. 7) and validated by optical map data. The only exception was the chromosome 2A centromere of TA10622, which carried two small gaps. In contrast to previous studies that investigated small portions of individual wheat centromeres[32,33] or highly fragmented assemblies[34], the assembly of complete einkorn centromeres enabled us to perform a detailed analysis of the structure and dynamics across whole Triticeae centromeres. Functional einkorn centromeres ranged from 4 to 5.8 Mb in size, with an average of 5.46 Mb (TA299) and 5.24 Mb (TA10622). In both einkorn accessions, the functional centromeres contained only one to nine genes, except for chromosome 7A of accession TA299, which contained 39 annotated genes. Most genes residing in CENH3-enriched domains were not expressed, whereas genes located in functional centromeres, but outside of CENH3-enriched domains, showed varying expression levels (Supplementary Fig. 8).

In *Arabidopsis thaliana*, centromeres contain megabase-scale arrays of approximately 178 bp tandem satellite repeats[35]. Similar centromeric satellite repeat arrays have been reported in different grass species[36–38]. By contrast, we did not detect high-copy, centromere-specific satellite repeats in einkorn (Supplementary Fig. 9 and Supplementary Note 1). The majority of the functional einkorn centromeres comprised TEs (91.25–97.35%). Although the TE proportions in functional einkorn centromeres were comparable to the chromosome arms, the exact TE composition differed markedly.

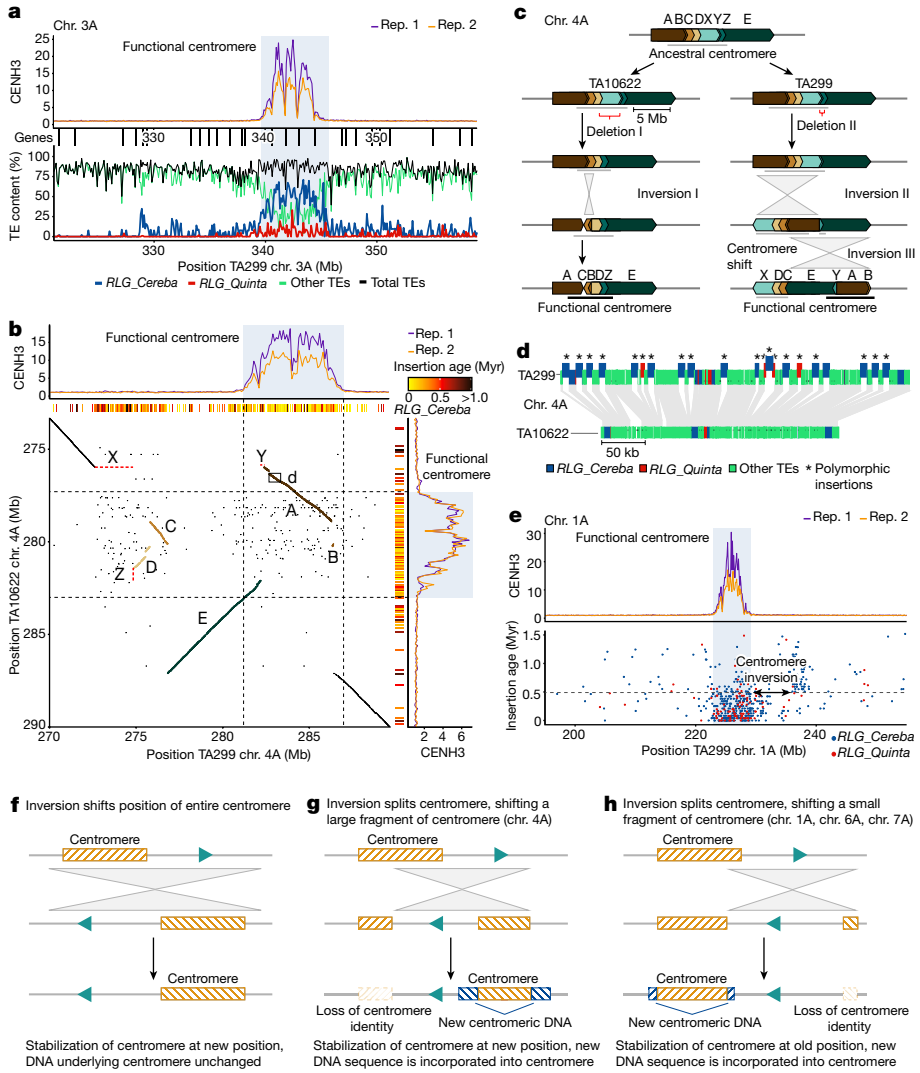


Fig. 2 | Dynamics of einkorn centromeres. a, The composition of the TA299 chromosome 3A centromere. The top track shows CENH3 ChIP–seq coverage. The vertical lines underneath the track indicate genes. The bottom track shows TE composition. The *x* axis indicates chromosomal positions in megabases. The functional centromere is highlighted (blue shading). **b**, Dot plot alignment of chromosome 4A centromeric regions of TA299 (horizontal) and TA10622 (vertical). CENH3 ChIP–seq coverage and positions of *RLG_Cereba* insertions are aligned with the dot plot. *RLG_Cereba* insertion age is colour-coded in million of years (Myr). Rearranged chromosomal segments are shown in colours that correspond to those in **c**. The small rectangle indicates an approximately 400 kb region that is shown in detail in **d**. **c**, Evolutionary model explaining the organization of chromosome 4A centromeres in TA10622 and TA299. A–E indicate segments that experienced inversions compared with the ancestral centromere. X–Z represent segments that were deleted in one of the two accessions. **d**, Comparison of the shifted TA299 chromosome 4A centromere with its counterpart in TA10622. Conserved sequences are connected by the shaded grey areas. New TE insertions are shown partially raised. All new TE insertions are of the *RLG_Cereba* and *RLG_Quinta* families. **e**, Evidence of an additional inversion of around 10 Mb in chromosome 1A that moved part of the functional centromere (indicated by the two-headed arrow). Top, CENH3 ChIP–seq coverage. Bottom, the chromosomal positions of *RLG_Cereba* and *RLG_Quinta* retrotransposons (*x* axis) and their insertion age (*y* axis). The distribution and insertion ages of retrotransposons indicate that the inversion occurred around 500,000 years ago (grey dashed line) in a common ancestor of TA10622 and TA299. **f**,**g**, Examples of how inversions can cause centromere shifts. **h**, Example of how a centromere remains at or near to its original location after a segment is moved by an inversion.

The centromere-associated retrotransposon family *RLG_Cereba* was predominant in the einkorn functional centromeres (around 70%), followed by *RLG_Quinta* elements (around 20%). Outside the functional centromeres, these two retrotransposon families were rare and other TE families dominated[39] (Fig. 2a and Extended Data Fig. 3). We identified approximately 2,500 full-length *RLG_Cereba* and around 1,000 full-length *RLG_Quinta* retrotransposons in each of the two einkorn assemblies. The youngest *RLG_Cereba* and *RLG_Quinta* insertions mapped inside the currently functional centromeres, supporting a model that the integrase enzyme encoded by *RLG_Cereba* elements recognize CENH3-containing nucleosomes[40]. The evolutionary youngest *RLG_Cereba* subpopulation was almost exclusively found in functional centromeres (Extended Data Fig. 4 and Supplementary Fig. 10). Older elements were 'pushed away' from functional centromeres by the more recent insertions (Extended Data Fig. 4). These results indicate that *RLG_Cereba* and *RLG_Quinta* retrotransposons are suitable markers for the location of current and past centromeres (Extended Data Fig. 5).

We found that functional einkorn centromeres are highly dynamic and evolving rapidly (Extended Data Figs. 5 and 6). This is in agreement with observations made in other grass species, indicating that centromeres can shift positions over time[23,41]. Sequence collinearity

between TA299 and TA10622 was low or completely absent across functional centromeres, while chromosome segments adjacent to the functional centromeres aligned well (Extended Data Fig. 6). We found multiple inversions inside or in the immediate vicinity of centromeres (Fig. 2 and Extended Data Fig. 6), and we hypothesize that inversions are major drivers of centromere evolution. Inversions can displace parts of functional centromeres, resulting in centromere shifts.

The most notable example of a centromere shift was found on chromosome 4A, which was highly rearranged between TA299 and TA10622 (Fig. 2b, Extended Data Fig. 6 and Supplementary Fig. 11). TA10622 had one distinct region with a high density of *RLG_Cereba* insertions. The youngest of these insertions coincided with the highest CENH3 signals. By contrast, TA299 showed two regions with high *RLG_Cereba* insertion densities separated by around 10 Mb. The region with the youngest insertions overlapped with the current functional centromere (as identified by high CENH3 signals), whereas the second region contained a fragment of the ancestral centromere (Fig. 2b). We deduced a model to explain the differences in this centromere between TA299 and TA10622 (Fig. 2c). We propose that a series of inversions divided the ancestral centromere into two fragments, after which the centromere was re-established at the site of the larger ancestral centromere fragment in TA299 (Fig. 2c). A detailed comparison of around 400 kb of the new TA299 centromere with its (ancestral) counterpart in TA10622 showed that all new TE insertions in the new TA299 centromere were of the *RLG_Cereba* and *RLG_Quinta* family (Fig. 2d). On the basis of *RLG_Cereba* insertion ages, we estimated that this centromere shift in TA299 occurred between 20,000 and 100,000 generations ago (Extended Data Fig. 5). Moreover, an independent deletion of around 2 Mb in TA10622 was followed by a centromere shift of about 1 Mb (Fig. 2c). We found additional evidence for inversions in the centromeric regions of chromosomes 1A, 6A and 7A (Fig. 2e and Extended Data Fig. 5). In chromosomes 1A and 6A, *RLG_Cereba* and *RLG_Quinta* insertion sites and ages indicated that parts of functional centromeres had moved in both accessions around 500,000 and 300,000 generations ago, respectively (Extended Data Fig. 5). On chromosome 7A, an approximately 13 Mb segment was inverted, thereby removing about 1 Mb of the functional centromere in TA10622. We estimated that this event took place around 100,000 generations ago (Extended Data Figs. 5 and 6). In chromosomes 1A, 6A and 7A, centromeres appeared to be only partially impacted, resulting in a re-establishment of functional centromeres in the same chromosomal region (Extended Data Fig. 6).

Our data revealed a substantial number of centromere rearrangements. We propose that inversions can cause major centromere shifts if more than half of the functional centromere is displaced by an inversion (Fig. 2f,g). In such cases, the functional centromere is re-established in a new location (Fig. 2g), as illustrated with chromosome 4A. By contrast, when only small portions of a centromere are moved, the functional centromere appears to remain at or near its original position and its original size is re-established over time (Fig. 2h). This also supports a model of optimal centromere size as compromised centromeres appear to settle at a consistent size across chromosomes, despite significant disruption over time.

Einkorn population genomics

To investigate einkorn genetic diversity and evolutionary history, we generated whole-genome sequencing data (around tenfold coverage) for a diversity panel comprising 219 einkorn accessions. We selected the constituent accessions of the panel to optimally represent diversity on the basis of geographical origin and genotyping data[42], with 158 wild (124 α race, nine β race and 25 γ race) and 61 domesticated einkorn accessions (Supplementary Table 8). In total, 121,459,674 high-quality single-nucleotide polymorphisms (SNPs) were retained and we observed a low false-positive error rate of variant calling[43–46] (Methods). Nucleotide diversity (π) was highest in the γ races ($\pi = 0.0023$) and similar across the other three groups (α, $\pi = 0.0011$; β, $\pi = 0.0018$; domesticated, $\pi = 0.0014$) (Supplementary Fig. 12a). Notably, but consistent with previous observations[13], we did not observe a large reduction in nucleotide diversity in domesticated einkorn.

Phylogeny and principal component analysis (PCA) confirmed that wild einkorn clusters into α, β and γ races[13,42] (Fig. 3a and Supplementary Fig. 12b). The domesticated einkorn accessions clustered together with race β, most of which were collected in the Karacadağ area in southeastern Turkey (Supplementary Table 8). This supports the hypothesis that einkorn was domesticated from a small and restricted wild population closely related to present-day β accessions.

We estimated the ancestry coefficient for each accession to examine the evolutionary history of einkorn. Consistent with the PCA, the three groups of wild einkorn races separated into three distinct clusters at $K = 3$ (where K is the number of putative ancestral populations), and domesticated einkorn grouped with the β accessions (Fig. 3b). At $K = 4$, domesticated einkorn split into two separate groups (Fig. 3b and Supplementary Fig. 13), which was not observed in a previous analysis using genotyping-by-sequencing data[42]. The cross-entropy values reached a plateau starting at $K = 6$ (Supplementary Fig. 14). Compared with $K = 4$, race β separated from domesticated einkorn, whereas wild α race accessions formed two distinct groups at $K = 6$ (Fig. 3b and Supplementary Table 9).

To test for admixture and whether specific genomic segments contributed to the split of domesticated einkorn, we estimated genetic differentiation (F_{ST}) between the two domesticated einkorn groups in sliding windows across the seven chromosomes, revealing two large segments that were highly differentiated between the two groups. These two blocks spanned the centromeric and pericentromeric regions of chromosomes 2A (around 266 Mb) and 5A (around 329 Mb) (Fig. 3c). Divergence analyses across these two segments confirmed a strong separation of domesticated einkorn accessions (Extended Data Fig. 7a,b). We performed PCA considering only variants that were located within these two diverged segments, revealing the clustering of some domesticated einkorn accessions with wild γ rather than β accessions (Fig. 3d and Extended Data Fig. 7c). These results suggest an introgression of genetic material from race γ into the domesticated einkorn gene pool. Pericentromeric regions show low recombination frequency in wheat[24,47,48], explaining why they can persist as large blocks.

To obtain a more complete estimate of the proportions of γ introgressions in domesticated einkorn, we performed pairwise comparisons of nucleotide diversity across chromosomes between one γ accession and each of the domesticated einkorn accessions. In addition to the two large segments on chromosomes 2A and 5A, we determined that most (92%) domesticated einkorn accessions carry an approximately 150 Mb γ genomic segment in the pericentromeric region of chromosome 7A (Extended Data Fig. 7d,e, Supplementary Fig. 15 and Supplementary Tables 10 and 11). Owing to its high frequency in the domesticated einkorn gene pool, we did not detect this segment by F_{ST} analysis. TreeMix[49] analysis supported the influx of genetic material from wild einkorn race γ into the domesticated einkorn gene pool (Supplementary Fig. 16). Overall, we estimate that the introgressions from race γ accounted for an average of 6.7% (range 0.3–13.1%) of the domesticated einkorn genome (Supplementary Table 10), resulting in an increased nucleotide diversity within the domesticated gene pool (Supplementary Fig. 17).

Wild einkorn accessions found in the Fertile Crescent mainly belong to race α, with race β being restricted to an area around the Karacadağ mountains. Race γ is not present in the Fertile Crescent[13,42] and is mainly found in central and northwestern Turkey (Fig. 3e). We hypothesize that the probable geographical site of the γ introgression can be inferred by comparing the genetic relatedness of the introgressed segments in

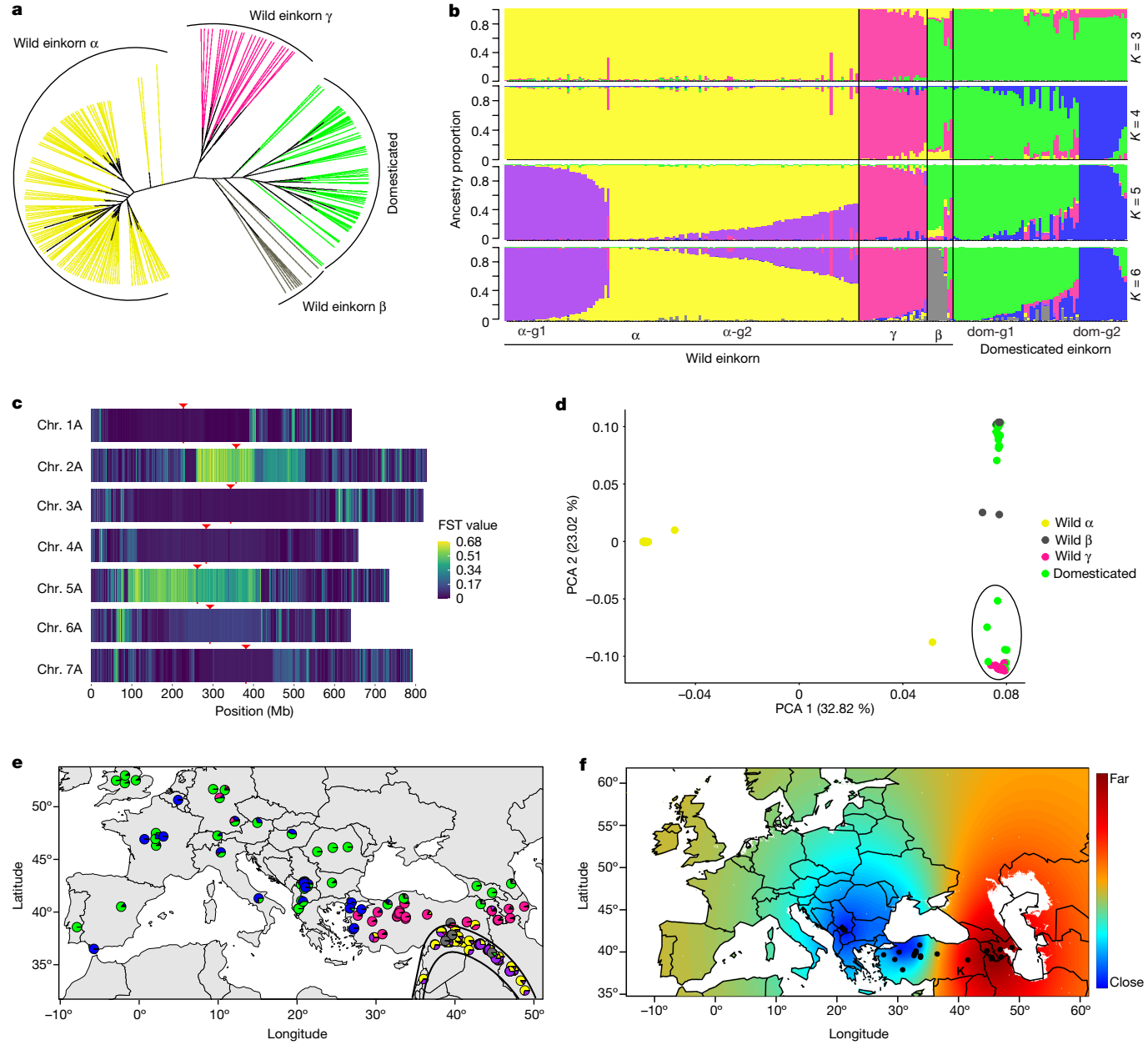


Fig. 3 | Einkorn population genomics. a, Unrooted neighbour-joining tree. **b**, Population structure (from $K = 3$ to $K = 6$). Each vertical bar represents one accession, and the bars are filled with colours representing the proportion of each ancestry. Einkorn groups were assigned considering $K = 6$ (on the basis of the cross-entropy value) based on the maximal local contribution of ancestry except for β (all β accessions were assigned as one group, regardless of the contribution of an ancestry). α group 1 (α-g1, $n = 37$) is shown in purple, α group 2 (α-g2, $n = 87$) is shown in yellow, γ ($n = 24$), β ($n = 9$), domesticated einkorn group 1 (dom-g1, $n = 44$) is shown in green and domesticated einkorn group 2 (dom-g2, $n = 17$) is shown in blue. A detailed list of accessions is provided in Supplementary Table 9. **c**, The mean fixation index (F_{ST}) between the two domesticated einkorn groups calculated in 1 Mb sliding windows. Only accessions with 80% ancestry threshold at $K = 4$ were considered. Centromere midpoints are indicated by red arrowheads. **d**, PCA using only variants that are present on the introgressed segment on chromosome 5A. Accessions were coloured according to the structure analysis in **b**. Circled accessions include wild γ accessions and some domesticated einkorn accessions. **e**, The geographical location of einkorn collection sites. The colours in pie charts correspond to the ancestry at $K = 6$. The Fertile Crescent is indicated by black lines. Only accessions with known collection sites are shown. **f**, Geographical projection of the first PCA axis for γ accessions on the basis of the introgressed segment on chromosome 2A (this analysis was performed excluding α and β accessions). The black dots represent the location of γ accessions. Blue colour represents the collection sites of γ accessions that were genetically the least diverged from the γ introgression found in domesticated einkorn. The Karacadağ region (K) is indicated on the map.

domesticated einkorn to wild γ accessions. The geographical projection of the first and the second PCA axes using the introgressed segments identified one group of γ accessions from central and northwestern Turkey that showed the closest genetic relatedness to the introgressed segments in domesticated einkorn. We propose that this geographical region, which is several hundred kilometres away from the site of einkorn domestication, is the most likely region where the hybridization between ancestral domesticated einkorn and a wild γ accession might have occurred (Fig. 3f and Extended Data Fig. 8a,b).

From the Fertile Crescent, cultivation of domesticated crops rapidly expanded following two main migration routes: one to the west into central Turkey (and later Europe) and a second to the east into

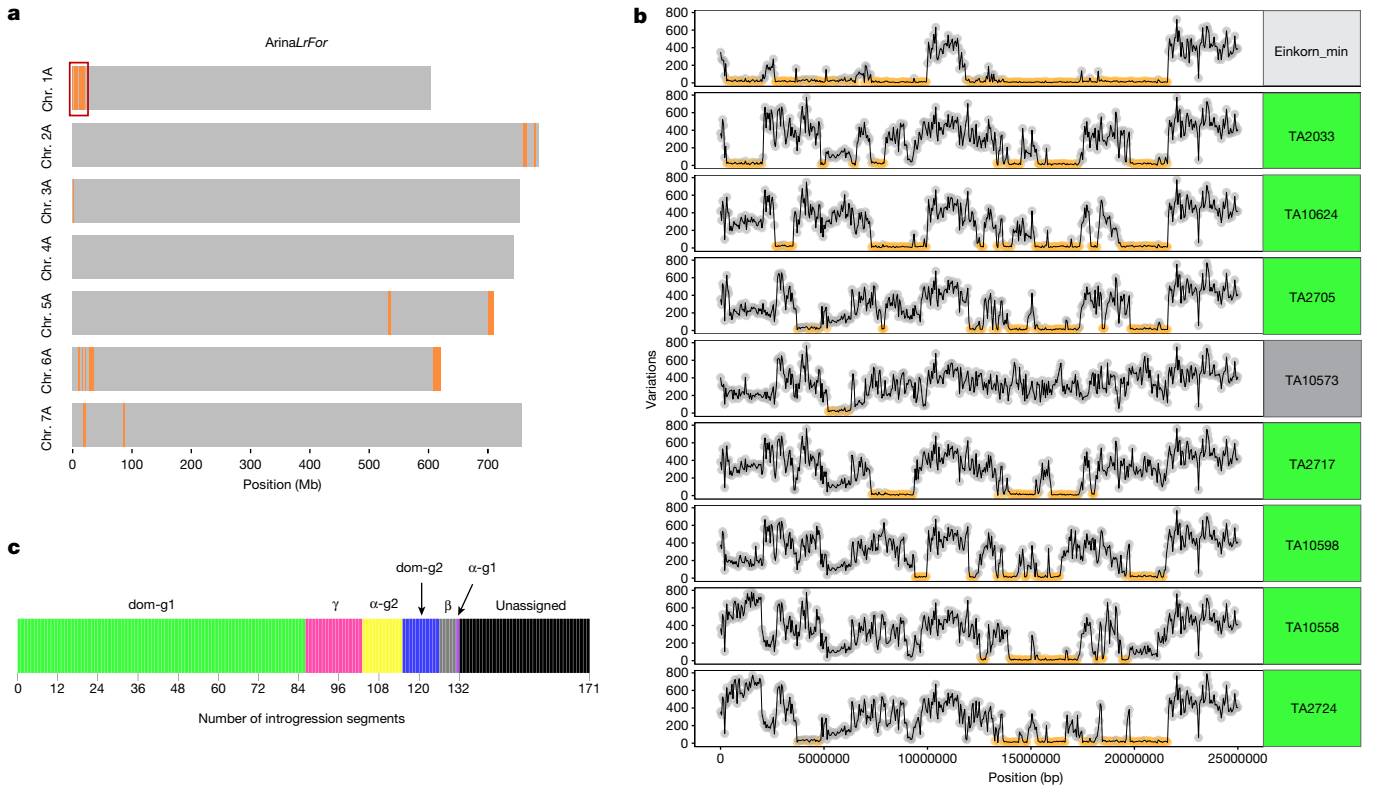


Fig. 4 | Einkorn introgression into bread wheat. a, Einkorn introgressions (highlighted in orange) into Arina*LrFor* identified using the *k*-mer variation approach (IBSpy; Supplementary Note 2). The red square on chromosome arm 1AS corresponds to the region shown in detail in **b**. **b**, IBSpy variations between Arina*LrFor* (chromosome 1A, position 0–25 Mb) and einkorn. Regions with variation scores of ≤30 (identical by state) are indicated in orange, corresponding to einkorn introgressions. Einkorn_min represents a consensus that shows the lowest variation scores across all resequenced einkorn accessions. The remaining plots illustrate the variation scores between Arina*LrFor* and eight selected einkorn accessions. Accession names highlighted in green and grey belong to domesticated groups 1 (dom-g1) and β, respectively. **c**, The number of introgression segments that could be assigned to a particular einkorn group.

Transcaucasia[50]. Notably, the domesticated einkorn accessions with the highest proportions of γ introgression were collected from western Turkey (coinciding with the most likely site of hybridization), whereas domesticated einkorn accessions collected in eastern Turkey and Georgia had a very low proportion of γ introgressions (Extended Data Fig. 8c and Supplementary Table 10). This pattern may reflect the different migration routes of domesticated einkorn. After domestication in the Karacadağ region, einkorn was most likely cultivated in close proximity to wild γ populations in central and northwestern Turkey, leading to the influx of genetic material from race γ. By contrast, the domesticated einkorn populations that were moved to the east were not impacted by this hybridization, which would explain the low proportion of γ introgressions in domesticated einkorn accessions from eastern Turkey and Georgia. The apparent lack of a strong domestication bottleneck in domesticated einkorn was explained with a 'dispersed-specific' model of einkorn domestication, including multiple domestication events from geographically dispersed wild β populations[13]. A hallmark of domesticated einkorn is the non-fragile rachis. In our diversity panel, all domesticated einkorn accessions had the same haplotype in the *non-brittle rachis1* (*btr1*) gene, including a critical alanine to threonine amino acid substitution[4], indicating that this key domestication gene has a single origin in domesticated einkorn. The lack of a strong diversity reduction in domesticated einkorn could therefore also be the result of gene flow after domestication, as demonstrated for the introgressions from wild γ accessions. Recent population and pan-genome analyses confirmed that hybridizations had an important role in increasing genetic diversity in wheat after domestication[23,51]. The introgression of genetic material from wild γ accessions may have had an important role in the adaptation of domesticated einkorn to new climatic conditions outside the Fertile Crescent.

Einkorn introgressions into bread wheat

Although einkorn is not the donor of the hexaploid bread wheat A subgenome, several einkorn introgressions have been described in hexaploid wheat[6,9,52–54]. To gain a comprehensive estimate of the proportions of einkorn introgressions in the modern bread wheat gene pool, we used *k*-mer-based approaches to detect einkorn introgressions in ten chromosome-scale bread wheat assemblies[23,24] (Supplementary Note 2). For a positive control, we tested a known einkorn translocation carrying the *Yr34* stripe-rust-resistance gene located at the distal end of chromosome arm 5AL in bread wheat lines Arina*LrFor* and SY Mattis[6]. We detected an approximately 8–10 Mb einkorn segment at the expected position in the two wheat lines (Fig. 4a, Extended Data Fig. 9 and Supplementary Table 12), supporting the idea that our *k*-mer-based approaches are suitable to detect einkorn introgressions. On average, we determined that the bread wheat A subgenomes contain around 1% of einkorn introgressions, ranging from 0.7% in cultivar CDC Landmark to 1.9% in Arina*LrFor* (Fig. 4a and Supplementary Table 12).

We identified 171 einkorn segments with a cumulative size of 472.8 Mb across the 10 bread wheat cultivars. The average segment size was 2.8 Mb, ranging from 50 kb to 16.7 Mb. In most cases, the introgressed einkorn segment could not be assigned to a single einkorn accession from our diversity panel. Instead, different regions of an

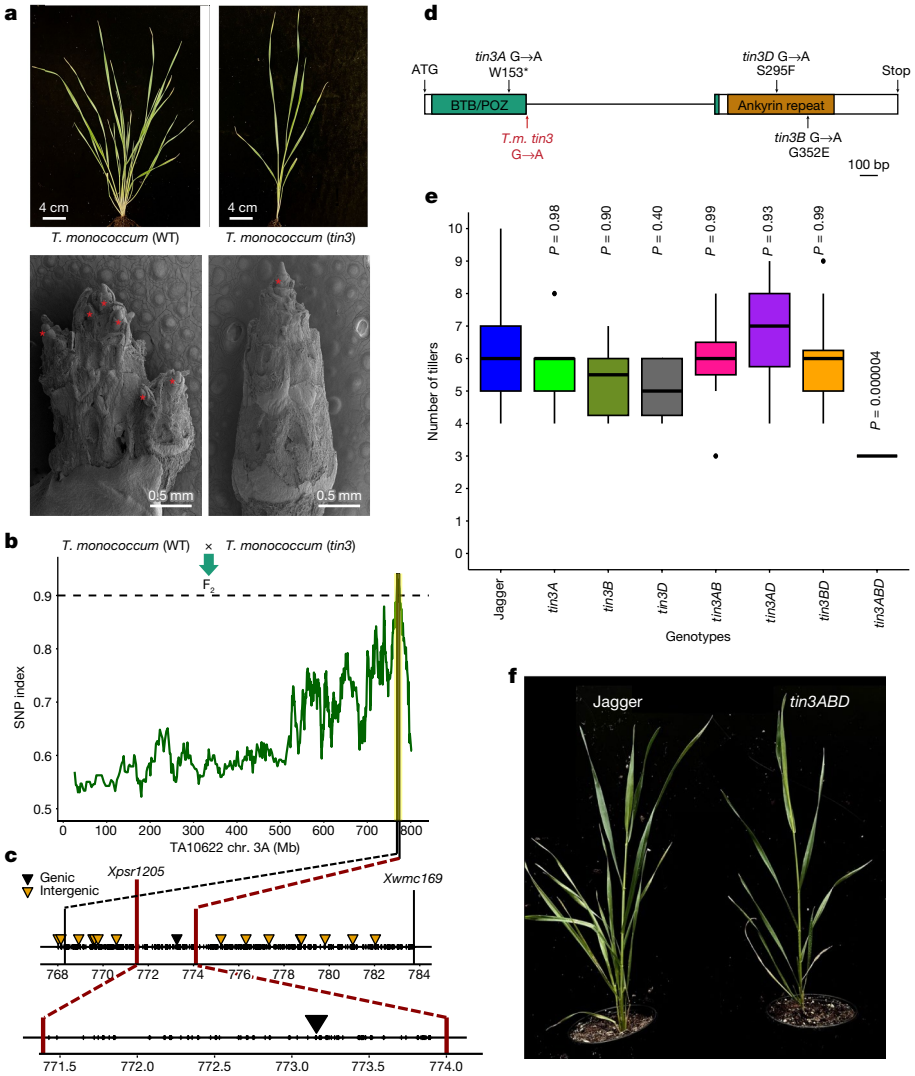


Fig. 5 | Positional cloning of *tin3* and translational research in hexaploid bread wheat. a, Phenotypes of wild-type *T. monococcum* accession TA4342-L96 (top left) and *tin3* (top right) at the tillering stage. Bottom left, scanning electron microscopy (SEM) images of seedlings showing primary shoot bud and axillary tiller bud formation (indicated by asterisks) in TA4342-L96 after leaf removal at the eight-leaf stage. Bottom right, SEM image of a seedling showing only the primary bud (indicated by an asterisk) at the shoot apex with no axillary buds in the *tin3* mutant after leaf removal at the six-leaf stage. The SEM experiment was repeated three times. **b**, The SNP index in a mutant *tin3* bulk ($n = 30$ F_2 plants) across einkorn chromosome 3A. The TA10622 reference assembly was used for read mapping. **c**, The *tin3* target interval in TA10622. *Xpsr1205* and *Xwmc169* indicate the positions of previously identified *tin3*-flanking markers. The triangles indicate the positions of EMS-induced point mutations. **d**, *Tin3* (*Tm.TA10622.r1.3AG0164370*) gene structure. The boxes represent exons and the line represents the intron. The G to A point mutation in *tin3* is indicated by a red arrow. The locations of SNPs found within the Jagger TILLING population for all three homeologous *tin3* copies is indicated in black. *T.m.*, *T. monococcum*. **e**, Tiller numbers in bread wheat cultivar Jagger ($n = 20$), *tin3A* ($n = 12$), *tin3B* ($n = 6$), *tin3D* ($n = 14$), *tin3AB* ($n = 7$), *tin3AD* ($n = 8$), *tin3BD* ($n = 12$) and *tin3ABD* ($n = 8$). All eight *tin3ABD* triple mutants developed exactly three tillers. The box boundaries indicate the first and third quartiles. The lines extending from the boxes (whiskers) indicate the variability outside the lower and upper quartiles. The lines in the middle of the boxes represent the median values of π. Outliers are plotted as individual points. *P* values were calculated using two-sided Tukey's honest significant difference tests, comparing with Jagger. **f**, Representative images showing the tillering phenotypes of Jagger (left) and *tin3* triple mutants (right).

introgressed segment showed close relatedness to various einkorn accessions of the diversity panel (Fig. 4b), suggesting that the direct donor of the introgressions was not present in the panel. However, 132 out of the 171 introgressed segments (with a cumulative size of 431.8 Mb) could be assigned to a specific group of einkorn accessions. The majority of the 132 segments with a clear origin (86 segments with a cumulative size of 287.1 Mb) were assigned to domesticated einkorn group 1 (dom-g1) (Fig. 4c and Supplementary Table 12). The remaining segments originated from the other domesticated and wild einkorn groups (Fig. 4c). These results indicate that most of the einkorn introgressions in the elite bread wheat genome originated from hybridizations between ancient tetraploid or hexaploid wheats and domesticated einkorn.

Mapping of a plant architecture gene

Diploid einkorn can be used as a model to map and clone agriculturally important genes and to translate this knowledge into polyploid bread wheat breeding. A diploid model species is particularly useful to clone recessive genes, of which the phenotypic effects are masked in a polyploid. Here we used the einkorn genomic resources to map the *tiller inhibition* (*tin3*) gene. Tillering is a key shoot architecture trait in

cereals, contributing to spike number and grain yield. *tin3* was originally identified as a recessive ethyl-methanesulfonate (EMS)-induced mutation in the domesticated einkorn accession TA4342-L96. The *tin3* mutant showed a reduced tiller number (Fig. 5a) and the causal gene was mapped to chromosome arm 3AL[55,56]. To further map *tin3*, we used a MutMap-based approach[57] with the TA10622 reference assembly to identify EMS-induced point mutations associated with *tin3*. MutMap and the physical positioning of previously identified *tin3*-flanking markers[56] revealed an approximately 2.5 Mb target interval, in which only one EMS-type (G to A) point mutation was found (Fig. 5b,c). This SNP disrupted the exon–intron junction of the gene *Tm.TA10622r1.3AG0164370*, resulting in intron retention and the formation of an aberrant protein (Fig. 5d). A kompetitive allele-specific PCR (KASP) marker derived on this SNP co-segregated with the *tin3* phenotype in a 'TA4342-L96 × *tin3*' mapping population of 750 F_2 gametes. *Tm.TA10622r1.3AG0164370* encodes a putative co-transcription factor with an N-terminal BTB/POZ domain and an ankyrin-repeat domain at the C terminus. The *tin3* candidate is orthologous to the *Uniculme4* (*Cul4*) gene that controls tillering in barley[58]. We next identified mutations in the *tin3* bread wheat orthologues using a hexaploid wheat TILLING population[59] (Fig. 5d). Although bread wheat mutants containing point mutations in only one or two of the *tin3* homeologues showed normal tillering (Fig. 5e,f), triple mutants affecting all three bread wheat subgenomes showed a significant decrease in tiller number (Fig. 5e,f). In summary, the einkorn genomic resources facilitated rapid identification of the gene underlying the *tin3* mutant in both diploid and polyploid wheat. We demonstrate that the knowledge gained in genetics from diploid wheat can be rapidly transferred to hexaploid bread wheat biology and improvement.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-06389-7.

1. Levy, A. A. & Feldman, M. Evolution and origin of bread wheat. *Plant Cell* **34**, 2549–2567 (2022).
2. Salamini, F., Ozkan, H., Brandolini, A., Schafer-Pregl, R. & Martin, W. Genetics and geography of wild cereal domestication in the Near East. *Nat. Rev. Genet.* **3**, 429–441 (2002).
3. Arranz-Otaegui, A., Gonzalez Carretero, L., Ramsey, M. N., Fuller, D. Q. & Richter, T. Archaeobotanical evidence reveals the origins of bread 14,400 years ago in northeastern Jordan. *Proc. Natl Acad. Sci. USA* **115**, 7925–7930 (2018).
4. Pourkheirandish, M. et al. On the origin of the non-brittle rachis trait of domesticated einkorn wheat. *Front. Plant Sci.* **8**, 2031 (2018).
5. Marcussen, T. et al. Ancient hybridizations among the ancestral genomes of bread wheat. *Science* **345**, 1250092 (2014).
6. Chen, S. et al. Stripe rust resistance gene *Yr34* (synonym *Yr48*) is located within a distal translocation of *Triticum monococcum* chromosome $5A^{m}L$ into common wheat. *Theor. Appl. Genet.* **134**, 2197–2211 (2021).
7. Kerber, E. & Dyck, P. Inheritance of stem rust resistance transferred from diploid wheat (*Triticum monococcum*) to tetraploid and hexaploid wheat and chromosome location of the gene involved. *Can. J. Genet. Cytol.* **15**, 397–409 (1973).
8. Saintenac, C. et al. Identification of wheat gene *Sr35* that confers resistance to Ug99 stem rust race group. *Science* **341**, 783–786 (2013).
9. Kolmer, J., Anderson, J. & Flor, J. Chromosome location, linkage with simple sequence repeat markers, and leaf rust resistance conditioned by gene *Lr63* in wheat. *Crop Sci.* **50**, 2392–2395 (2010).
10. The, T. Chromosome location of genes conditioning stem rust resistance transferred from diploid to hexaploid wheat. *Nat. New Biol.* **241**, 256 (1973).
11. Heun, M. et al. Site of einkorn wheat domestication identified by DNA fingerprinting. *Science* **278**, 1312–1314 (1997).
12. Heun, M., Haldorsen, S. & Vollan, K. Reassessing domestication events in the Near East: einkorn and *Triticum urartu*. *Genome* **51**, 444–451 (2008).
13. 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 (2007).
14. Brandolini, A., Volante, A. & Heun, M. Geographic differentiation of domesticated einkorn wheat and possible Neolithic migration routes. *Heredity* **117**, 135–141 (2016).
15. Behre, K. E., Wasylikowa, K. & van Zeist, W. *Progress in Old World Palaeoethnobotany* (Taylor & Francis, 1991).
16. Harlan, J. R. & Zohary, D. Distribution of wild wheats and barley: the present distribution of wild forms may provide clues to the regions of early cereal domestication. *Science* **153**, 1074–1080 (1966).
17. Badr, A. et al. On the origin and domestication history of barley (*Hordeum vulgare*). *Mol. Biol. Evol.* **17**, 499–510 (2000).
18. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
19. Lam, E. T. et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30**, 771–776 (2012).
20. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
21. Johnson, B. L. & Waines, J. G. Use of wild-wheat resources. *Hilgardia* **31**, 8–9 (1977).
22. Dvorak, J. et al. Reassessment of the evolution of wheat chromosomes 4A, 5A, and 7B. *Theor. Appl. Genet.* **131**, 2451–2462 (2018).
23. Walkowiak, S. et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature* **588**, 277–283 (2020).
24. International Wheat Genome Sequencing Consortium. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191 (2018).
25. Cossu, R. M. et al. LTR retrotransposons show low levels of unequal recombination and high rates of intraelement gene conversion in large plant genomes. *Genome Biol. Evol.* **9**, 3449–3462 (2017).
26. Backhaus, A. E. et al. High expression of the MADS-box gene *VRT2* increases the number of rudimentary basal spikelets in wheat. *Plant Physiol.* **189**, 1536–1552 (2022).
27. Li, K. et al. Interactions between SQUAMOSA and SHORT VEGETATIVE PHASE MADS-box proteins regulate meristem transitions during wheat spike development. *Plant Cell* **33**, 3621–3644 (2021).
28. Prasad, K., Parameswaran, S. & Vijayraghavan, U. *OsMADS1*, a rice MADS-box factor, controls differentiation of specific cell types in the lemma and palea and is an early-acting regulator of inner floral organs. *Plant J.* **43**, 915–928 (2005).
29. Huang, Y. et al. *Wide Grain 7* increases grain width by enhancing H3K4me3 enrichment in the Os*MADS1* promoter in rice (*Oryza sativa* L.). *Plant J.* **102**, 517–528 (2020).
30. McKinley, K. L. & Cheeseman, I. M. The molecular basis for centromere identity and function. *Nat. Rev. Mol. Cell Biol.* **17**, 16–29 (2016).
31. Earnshaw, W. C. Discovering centromere proteins: from cold white hands to the A, B, C of CENPs. *Nat. Rev. Mol. Cell Biol.* **16**, 443–449 (2015).
32. Liu, Z. et al. Structure and dynamics of retrotransposons at wheat centromeres and pericentromeres. *Chromosoma* **117**, 445–456 (2008).
33. Li, B. C. et al. Wheat centromeric retrotransposons: the new ones take a major role in centromeric structure. *Plant J.* **73**, 952–965 (2013).
34. Su, H. D. et al. Centromere satellite repeats have undergone rapid changes in polyploid wheat subgenomes. *Plant Cell* **31**, 2035–2051 (2019).
35. Naish, M. et al. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. *Science* **374**, eabi7489 (2021).
36. International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763–768 (2010).
37. Schnable, P. S. et al. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
38. Cheng, Z. et al. Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* **14**, 1691–1704 (2002).
39. Wicker, T. et al. Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biol.* **19**, 103 (2018).
40. Neumann, P. et al. Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mob. DNA* **2**, 4 (2011).
41. Wolfgruber, T. K. et al. Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons. *PLoS Genet.* **5**, e1000743 (2009).
42. Adhikari, L. et al. Genetic characterization and curation of diploid A-genome wheat species. *Plant Physiol.* **188**, 2101–2114 (2022).
43. Zhao, X. et al. Population genomics unravels the Holocene history of bread wheat and its relatives. *Nat. Plants* **9**, 403–419 (2023).
44. Zhou, Y. et al. Triticum population sequencing provides insights into wheat adaptation. *Nat. Genet.* **52**, 1412–1422 (2020).
45. Ramu, P. et al. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat. Genet.* **49**, 959–963 (2017).
46. Abrouk, M. et al. Fonio millet genome unlocks African orphan crop diversity for agriculture in a changing climate. *Nat. Commun.* **11**, 4488 (2020).
47. Jordan, K. W. et al. The genetic architecture of genome-wide recombination rate variation in allopolyploid wheat revealed by nested association mapping. *Plant J.* **95**, 1039–1054 (2018).
48. Sidhu, D. & Gill, K. S. Distribution of genes and recombination in wheat and other eukaryotes. *Plant Cell Tiss. Org. Cult.* **79**, 257–270 (2005).
49. Pickrell, J. & Pritchard, J. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
50. Narasimhan, V. M. et al. The formation of human populations in South and Central Asia. *Science* **365**, eaat7487 (2019).
51. Keilwagen, J. et al. Detecting major introgressions in wheat and their putative origins using coverage analysis. *Sci. Rep.* **12**, 1908 (2022).
52. Chhuneja, P. et al. Mapping of adult plant stripe rust resistance genes in diploid A genome wheat species and their transfer to bread wheat. *Theor. Appl. Genet.* **116**, 313–324 (2008).
53. Shi, A., Leath, S. & Murphy, J. A major gene for powdery mildew resistance transferred to common wheat from wild einkorn wheat. *Phytopathology* **88**, 144–147 (1998).
54. Bonafede, M., Kong, L., Tranquilli, G., Ohm, H. & Dubcovsky, J. Reduction of a *Triticum monococcum* chromosome segment carrying the softness genes *Pina* and *Pinb* translocated to bread wheat. *Crop Sci.* **47**, 821–828 (2007).

55. Kuraparthy, V., Sood, S., Dhaliwal, H. S., Chhuneja, P. & Gill, B. S. Identification and mapping of a tiller inhibition gene (*tin3*) in wheat. *Theor. Appl. Genet.* **114**, 285–294 (2007).
56. Kuraparthy, V., Sood, S. & Gill, B. S. Genomic targeting and mapping of tiller inhibition gene (*tin3*) of wheat using ESTs and synteny with rice. *Funct. Integr. Genom.* **8**, 33–42 (2008).
57. Abe, A. et al. Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat. Biotechnol.* **30**, 174–178 (2012).
58. Tavakol, E. et al. The barley *Uniculme4* gene encodes a BLADE-ON-PETIOLE-like protein that controls tillering and leaf patterning. *Plant Physiol.* **168**, 164–174 (2015).
59. Rawat, N. et al. A TILLING resource for hard red winter wheat variety Jagger. *Crop Sci.* **59**, 1666–1671 (2019).

Methods

Plant materials

The plant material used in this study was selected from a collection of 733 accessions of wild and domesticated einkorn held at the Wheat Genetics Resource Center (WGRC) at Kansas State University[42]. To generate reference assemblies, we selected one domesticated einkorn accession (*T. monococcum* L. subsp. *monococcum*; TA10622) collected in Albania and one wild accession (*T. monococcum* L. subsp. *aegilopoides*; TA299) collected in northern Iraq. These two accessions were selected on the basis of genotyping-by-sequencing data[42] as representative accessions within the respective clade (domesticated and wild race α), had low heterozygosity and available passport information. For whole-genome sequencing, a panel of 219 accessions (Supplementary Table 8) was selected on the basis of diversity and geographical origin[42].

Reference genome sequencing, assembly, and validation

PacBio HiFi library preparation and sequencing. High-molecular-mass (HMM) genomic DNA was isolated from young leaves obtained from 3-week-old seedlings after dark treatment for 48 h. DNA was extracted according to a HMM DNA extraction protocol for long-read sequencing[60]. DNA quantification was performed using the Qubit dsDNA HS Assay (Q32851, Thermo Fisher Scientific), purity was confirmed using the Nanodrop spectrophotometer by checking the 260/280 and 260/230 ratios, and the DNA size was validated by using the FemtoPulse system (Agilent). HiFi libraries were then prepared according to the manual 'Procedure & Checklist - Preparing HiFi SMRTbell Libraries using the SMRTbell Express Template Prep Kit 2.0' (PN 101-853-100, Pacific Biosciences) with 10 µg DNA sheared by using the Megaruptor 2 system (Diagenode) to obtain a 15–20 kb average size. Size-selected libraries were sequenced on the PacBio Sequel II system in CCS mode for 30 h. For each accession, we obtained 121 Gb PacBio HiFi reads, corresponding to a coverage of around ~21-fold (Supplementary Table 13).

Bionano optical map. Einkorn grains were germinated on filter paper in the dark for 4 days at 4 °C followed by 3 days at 25 °C. Ultra HMM DNA was isolated from fresh root meristem tissue using the Plant DNA Isolation Kit protocol (Bionano Genomics). Labelling was performed using direct labelling enzyme (DLE1) and staining of the HMM DNA according to the Bionano Prep Direct Label and Stain (DLS) protocol (30206-Bionano Genomics). Optical maps were generated using the Bionano Genomics Saphyr System (Saphyr Chip G1.2) according to the Saphyr System User Guide (3024-Bionano Genomics). Data processing was performed using the Bionano Solve v.3.6 software (https://bionanogenomics.com/support/software-downloads).

Omni-C library construction and sequencing. Omni-C libraries were prepared using the Dovetail Omni-C Kit for plant tissues according to the manufacturer's protocol. In brief, chromatin was fixed in place in the nucleus from 100 mg of dark-treated young leaves. An in situ digestion of the fixed chromatin was performed using an endonuclease enzyme. Chromatin was then released by lysing the cells. Chromatin ends were repaired and ligated to a biotinylated bridge adapter to capture contacts, before proximity ligation of the adapter-containing ends. After proximity ligation, cross-link-reversal and DNA purification from proteins was performed. The purified DNA was treated to remove biotin that was not internal to ligated fragments. Two sequencing libraries were generated using Illumina-compatible adapters for each accession. Biotin-containing fragments were isolated using Streptavidin beads before PCR enrichment of the library. The two libraries were sequenced on the Illumina HiSeq X platform. Around 400 million paired-end reads (2 × 150 bp) were generated.

Genome assembly. PacBio HiFi reads were assembled using hifiasm (v.15.1)[61] with the default parameters (https://github.com/chhylp123/hifiasm/) to generate primary contig assemblies. We generated hybrid scaffolds by combining contig assemblies and optical maps using the hybridScaffold pipeline (Bionano Solve v.3.6) with the default parameters. We then integrated Omni-C read data to produce pseudomolecule assemblies using Juicer (v.1.6; https://github.com/aidenlab/juicer)[62] and the 3D-DNA pipeline (https://github.com/aidenlab/3d-dna)[63]. First, to generate Hi-C contacts (with duplicates removed), Omni-C Illumina short reads were processed with juicer.sh (parameter: -s none). The output file 'merged_nodups.txt', the hybrid scaffolds and contigs that were not integrated into hybrid scaffolds (in one fasta file) were then used to produce an assembly with 3D-DNA[63] (using run-asm-pipeline.sh with -r 0 parameter). We used Juicebox (v.1.11.08)[64] to visualize the Hi-C contact matrix, and to manually curate the assembly (orient and order hybrid scaffolds and pseudomolecules). The final Hi-C contact maps and assemblies were saved using run-asm-pipeline-post-review.sh from the 3D-DNA pipeline.

Assembly validation and quality control. We validated the two einkorn assemblies by mapping PacBio HiFi reads (with minmap2 v.2.21), Illumina short reads (with bwa mem v.0.7.17) and the optical map (Bionano Solve v.3.6) to the final assemblies and we found no major discrepancies. Assemblies were further validated using a genetic map constructed from a recombinant inbred line (RIL) population (see below). Using the genetic map, we manually corrected three misorientations in the telomeric regions (chromosomes 2A and 4A in TA299 and chromosome 2A in TA10622), and introduced a 1.04 Mb segment into chromosome 4A of TA10622 that was placed in the unanchored chromosome (Supplementary Table 14). We revalidated the corrected assembly by (1) re-calling SNPs from the RILs, (2) mapping raw-reads and optical maps and (3) mapping the individual contigs to both assemblies to ensure correct orientation of the regions. Assembly completeness was evaluated using BUSCO (v.5.0.0)[65] with the plant dataset (poales_odb10). Moreover, we generated Illumina short reads (150 bp paired-end reads, ~40-fold coverage) from leaf tissues of TA299 and TA10622 to evaluate the assemblies. Merqury (v.1.3)[66] was used to estimate the assembly consensus quality (QV) and completeness on the basis of the comparison of *k*-mers.

Genetic map construction

RIL population. We constructed a genetic map using a recombinant inbred line (RIL) population consisting of 827 lines resulting from a cross between a wild (TA291, also identified as TA4342_L95) and a domesticated (TA10868, also identified as TA4342_L96) einkorn accession. The two parents of the RIL population were sequenced at high depth (9.1-fold) using a TruSeq library, whereas the RILs were sequenced using a low-coverage (0.03-fold) skim-sequencing (skim-seq) protocol that used a low-volume Illumina Nextera library[67]. In the skim-seq panel, we also included five replicates of each parent.

Marker discovery and genotyping of RILs. Both the TruSeq and Nextera libraries were sequenced on the Illumina HiSeq X10 system with 2 × 150 bp reads (Psomagen). We used custom Perl scripts for demultiplexing raw FASTQ files (https://github.com/sandeshsth/SkimSeq_Method) obtained from Nextera sequencing[67] and TruSeq (https://github.com/sandeshsth/Fastq). Adapters and primers were trimmed using fastp[68]. Trimmed high-quality reads from the two parents were aligned to the TA299 and TA10622 assemblies separately using SAMtools[69] (v.1.8) and variants were called using BCFtools (v.1.9)[70]. Variants were filtered for minimum and maximum filtered read depths of ≥6 and ≤100, respectively, and reference and alternative allele depths of ≥3. Missing and heterozygous genotypes called in either RIL parents were removed. The remaining homozygous SNPs were then called on the RIL population[67]. Owing to the low sequence coverage, we used a bin mapping approach in 1 Mb sliding windows to call consensus genotypes[71]. Genotypes called on RILs were coded according to the parental

SNPs replacing the genotypes as either wild (P1) or domesticated (P2). A consensus genotype was called within the 1 Mb sliding windows on the basis of the proportions of P1 and P2 within the window. If P1/P2 ≥ 0.7, then we coded the window as P1, if P2/P1 ≥ 0.7 we coded the window as P2, otherwise as heterozygous (H). A custom Python script (Data availability) was used to genotype the 1 Mb windows and to identify the recombination breakpoints. The genotyping file with filtered recombination bins for missing and heterozygous loci and individual RILs was used to construct the genetic linkage maps.

Genetic linkage map construction. The genetic linkage maps were constructed using JoinMap (v.5.0; https://www.kyazma.nl/index.php/JoinMap/) using the bins as markers. We filtered the markers for missing data (removed >30% missing) at the population level. In JoinMap, we removed identical markers (similarity = 1) and mapped only one marker of the identical pair. We grouped the markers using minimum LOD of 6 and the markers were mapped using a regression mapping approach and the Kosambi function. The linkage maps were visualized using Mapchart (v.2.32; https://www.wur.nl/en/show/mapchart.htm). Linkage maps were constructed using this approach with both wild and domesticated einkorn assemblies.

Comparing genetic maps to TA299 and TA10622 assemblies. We visually compared the marker order of the genetic map to the two einkorn assemblies and corrected discrepancies if all of the following three criteria were met: (1) The genetic distance between two mis-ordered loci was ≥0.5 cM to ensure it is not a mistake in the genetic map; (2) the correction does not require breaking of a hybrid scaffold; and (3) the reorientation was supported by the mapping of raw reads and the optical maps.

Genome annotation

Transcriptome sequencing. Around 100 mg of frozen and ground tissues from roots, whole aerial parts at the seedling stage, flag leaves, fully emerged spikes, glumes and grains were used for RNA isolation using the Maxwell RSC Plant RNA Kit (AS1500) and the Maxwell RSC 48 instrument according to the kit protocol (Promega). For RNA-seq, around 10 Gb of Illumina 150 bp paired-end reads were generated for each tissue. PacBio Iso-seq SMRTbell libraries were constructed according to the standard isoform sequencing protocol (Pacific Biosciences, 101-763-800). Full-length complementary DNA was synthesized from total RNA from the six tissues separately using NEBNext Single Cell/Low Input cDNA Synthesis & Amplification Module (New England Biolabs, E6421S). The ProNex Size-Selective Purification System (Promega, NG2001) was used for size selection. One SMRT Cell 8M (Pacific Biosciences, 101-389-001) was sequenced on the PacBio Sequel II system using the Sequencing Kit 2.0 (Pacific Biosciences, 101-820-200).

Gene model prediction. For both TA299 and TA10622, gene model prediction was performed according to a previously described method[72] with minor modifications, combining transcriptomics data, ab initio prediction and protein homology. First, TA299 and TA10622 RNA-seq data from the six tissues were mapped to their respective reference assemblies using STAR[73] (v.2.7.0f; parameters: --outFilterMismatchNoverReadLmax 0.02) and assembled into transcripts with StringTie[74] (v.2.1.4; parameters : --rf -m 150 -f 0.3 -t). Iso-seq data were mapped using minimap2[75] (v.2.21; parameters: -ax splice -uf –secondary=no -C5) and the redundant isoforms were further collapsed into transcript loci using cDNA_Cupcake (v.12.4.0; http://github.com/Magdoll/cDNA_Cupcake; parameters: --dun-merge-5-shorter). The RNA-seq and Iso-seq transcripts were merged using StringTie[74] (v.2.1.4; parameters: --merge -m 150) for each accession into a pool of candidate transcripts and Transdecoder (v.5.5.0; https://github.com/TransDecoder/TransDecoder) was used to find potential open reading frames and to predict protein sequences within the candidate transcript set. For the ab initio gene predictions, we used BRAKER2 (v.2.1.2)[76] and FgeneSH (v.8.0.0; http://www.softberry.com). In brief, BRAKER2 gene prediction was trained supported by RNA-seq and Iso-seq data (parameters: --softmasking --gff3 --cores=48 --nocleanup --bam='list of BAM files'). For the FgeneSH prediction, the TA299 and TA10622 pseudomolecules were repeat masked using a de novo repeat library constructed using the EDTA pipeline[77] and the TREP database[78] (v.19). FgeneSH annotation was performed using the monocot matrix for the gene prediction. For the protein homology evidences, we used the translated proteins from gene annotations of *T. urartu*[79], *Aegilops tauschii*[80], wild emmer (Zavitan)[81], hexaploid wheat (Kariega[72] and ArinaLrFor[23]), barley (Morex v.3)[82], the related grass species *Brachypodium distachyon*[36] and rice[83], and the Triticeae and Poaceae protein sequences downloaded from the UniProt database (2021_03). All protein sequences were mapped against the TA299 and TA10622 assemblies using GenomeThreader[84] (v.1.7.1; parameters: -startcodon -finalstopcodon -species rice -gcmincoverage 70 -prseedlength 7 -prhdist 4 -gff3out). We used EVidenceModeler[85] (v.1.1.1) to join all of the gene evidences from transcriptomics, ab initio predictions and protein alignments with weights adjusted according to the input source (FgeneSH = 2; BRAKER2 = 1; protein homology = 6; transcriptomics = 12). Finally, we performed two rounds of isoform and UTR prediction using the PASA pipeline (v.2.5.1)[86] with the default parameters. Gene models were classified into high and low confidence according to the classification criteria used by the International Wheat Genome Sequencing Consortium[24] and a previous study[82]. In brief, protein-encoding gene models were considered to be complete when start and stop codons were present. A comparison against PTREP[78], UniPoa (Poaceae database of annotated proteins from UniProt_2021_03) and UniViri (Viridiplantae database) was performed using DIAMOND[87] (v.2.0.9) and a BUSCO (v.5.2.2) analysis against the poales database (v.10; parameters: -m prot -c 20 -l poales). Gene candidates were further classified using the following criteria: a high-confidence (HC) gene model is complete with a hit in the UniViri database and/or in UniPoa and/or BUSCO poales database and not PTREP. A low-confidence (LC) gene model is incomplete and has a hit in the UniViri or UniPoa or BUSCO poales database but not in PTREP, or the protein sequence is complete with no hit in UniViri, UniPoa, BUSCO poales and PTREP. Putative functional annotations were assigned to HC and LC transcripts using a protein comparison against the UniProt database (2021_03) and PFAM domain signatures and Gene Ontology were assigned using InterproScan[88] (v.5.55–88.0).

Circos and synteny

Dot plot comparisons were performed using chromeister[89]. For circos, we performed a BLAST search using HC protein sequences with DIAMOND[87] (v.2.0.9; parameter -e 1e-10) and syntenic blocks were identified with MCScanX[90] between TA299 and TA10622. Only relationships between the same chromosomes were retained. The Circos plot was generated using the Circos software[91]. The density of genes and repeat elements as well as CENH3 read coverage shown in the ciros plot were calculated in non-overlapping 10 Mb windows.

Identification of 1 Mb tandem duplication

The 1 Mb tandem duplication was initially identified because of a conflict in the long arm of chromosome 4A between the contig-level assembly of TA10622 and the hybrid scaffolds after integrating the optical map. The contig-level assembly suggested an approximately 1 Mb tandem duplication that was collapsed in the optical map. Manual inspection revealed that multiple PacBio reads spanned the putative duplication breakpoints (Supplementary Fig. 18a–c). We designed a primer pair across the junction between the duplicated segments, which amplified in TA10622, but not in TA299 (Supplementary Fig. 18d), confirming that the tandem duplication in TA10622 is real. We manually corrected the disagreement based on the genetic map and read mapping, which supported the contig-level assembly. We defined the breakpoints to

the base-pair level and analysed the sequences located at the breakpoints using Gepard (v.1.40)[92] by generating dot plots. A PCR marker (forward primer, 5′-GGTCCCAGGCCATGATACCTC-3′; reverse primer, 5′-CTATGTCTCCCACGTGTCGAGGT-3′) was developed to validate the presence of the duplication in T10622. The amplicon (361 bp) was verified by Sanger sequencing. To inspect the sequence similarity between the two segments, we aligned one segment (558331155–559371848 bp) against the other (557272410–558331154 bp) using minimap2 (v.2.21)[75], and we counted the number of SNPs and the covered sequences in 5 Mb genomic windows. We assessed the einkorn diversity panel for the duplication using whole-genome alignments and analysed the read depths at the MADS-box gene from sequencing reads of all 218 einkorn accessions using SAMtools[69] depth command (v.1.8). The variant frequency between the duplicated segments was above the HiFi sequence read accuracy (mean read QV = 30), an important factor that probably enabled the differentiation and assembly of the two highly similar segments.

Centromere analysis

ChIP–seq. ChIP was performed according to a described previously method[93], standardized with anti-wheat-CENH3 antibodies[94]. An antigen with the peptide sequence 'RTKHPAVRKTKALPKK' corresponding to the N terminus of wheat CENH3 was used to produce the antibody using the custom-antibody production facility provided by Thermo Fisher Scientific. In total, 0.396 mg of customized antibody was purified and obtained as a pellet. The pellet was dissolved in 2 ml of PBS buffer, pH 7.4 resulting in 198 ng μl^{-1} of CENH3 antibody. Antibodies against H3K4me3 (04-475) were purchased from Sigma-Aldrich. The specificity of the anti-CENH3 antibodies was validated using immunofluorescence assays on mitotic and meiotic chromosomes of diploid (*T. monococcum*) and hexaploid (*T. aestivum*) wheat. Nuclei were isolated from 2-week-old seedlings and digested with micrococcal nuclease (Sigma-Aldrich) to liberate nucleosomes. The digested mixture was incubated overnight with 3 µg of antibody at 4 °C. Target antibodies were captured from the mixture using Dynabeads Protein G (Invitrogen) to obtain ChIP DNA. Mock DNA control was maintained with the input DNA using the same conditions as described above without antibodies. The ChIP experiments were performed in two biological replicates. Library construction was performed using the TruSeq ChIP Sample Prep Kit (Illumina) according to the manufacturer's instructions.

PacBio DNA methylation analysis. Methylation in the CpG context for TA299 and TA10622 was inferred with ccsmeth (v.0.3.2)[95], a deep-learning method to detect DNA 5mCpGs using kinetics features from PacBio CCS reads. The methylation prediction for CCS reads was called using the model 'model_ccsmeth_5mCpG_call_mods_attbigru2s_b21.v1.ckpt' and then aligned to their respective genome using BWA (v.0.7.17)[96] and reads were filtered for hard/soft clips and quality (MAPQ ≥ 60) using SAMtools (v.1.8)[69]. The methylation frequencies were calculated at the genome level from their respective modbam files and using the aggregate mode of ccsmeth with the model 'model_ccsmeth_5mCpG_aggregate_attbigru_b11.v2.ckpt'. To generate meta plots of *RLG_Cereba* elements, the bedmethyl file resulting from methylation frequency call was converted to bedgraph format. The bedgraph file was converted to bigwig using the script bedGraphToBigWig[97]. CpG methylation for *RLC_Cereba* copies was then calculated using the deeptools computeMatrix function and visualized using the deeptools plotProfile function[98].

ChIP–seq data analysis. Raw ChIP and input control sequencing reads were quality filtered and adapter sequences were removed with trimmomatic[99] using LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:50. Trimmed reads were then mapped to the respective genomes using bowtie2[100] with the default parameters. The SAM output file was converted to .bam format, the reads were sorted by position and duplicates were removed using SAMtools (v.1.8)[69]. Secondary alignments (that is, multi-mapping reads) were removed using the flag -F 0x0100. The resulting .bam file was filtered using SAMtools view to retain only reads that align over their full read length. The filtered .bam files were indexed using SAMtools index with the -c flag. The ratio of ChIP/input coverage was calculated using the deeptools function bamCompare using MAPQ ≥ 30 as a threshold[98]. To define centromere boundaries, the epic2 peak caller was used to identify peaks of CENH3 enrichment with a MAPQ ≥ 30 filtering[101]. Centromere boundaries were then defined by the density of epic2 peaks with a resolution of 100 kb (Supplementary Fig. 4). To assess the contiguous assembly of the centromeres, we obtained the breakpoints (start and end) of each contig in the pseudomolecule assemblies using MUMmer (v.4.0.0.2)[102]. We then compared contig breakpoints to the CENH3 read density plots to see whether the centromeres were present on a single contig.

Tandem repeat annotation. Tandem repeats were identified and annotated using tandem repeats finder[103] using the recommended standard settings but with 2,000 bp max periods size (match = 2, mismatch = 7, delta = 7, PM = 80, PI = 10, minscore = 50, maxperiod = 2000). To complement the analysis with tandem repeats finder, we also searched for instances of tandemly repeated *RLG_Cereba* elements (for example, *RLG_Cereba* elements with three long terminal repeats and two internal domains, which resulted from unequal recombination between LTRs) using BLASTN queries against the TA299 genome assembly. In this way, 61 instances of tandemly repeated *RLG_Cereba* elements were identified. As a control, we also searched for such recombinant *RLC_Angela* elements, which are around ten times more abundant than *RLG_Cereba* elements, but largely absent from functional centromeres. This revealed 620 tandemly repeated *RLC_Angela* elements. From this analysis, we conclude that, although there are tandemly repeated *RLG_Cereba* elements, the number is in the range of what could be expected from other non-centromeric TEs, revealing no higher-order structure of *RLG_Cereba* elements.

RNA-seq read mapping and feature counting. RNA-seq reads from the six tissues of TA299 and TA10622 were mapped to the respective genomes using STAR aligner (v.2.5.2a)[73] with the flags --outFilterMultimapNmax 20000, --outFilterMismatchNoverLmax 0.0, --alignIntronMax 1000. Features were counted using featureCounts (v.2.0.0)[104].

Genome-wide TE annotation. TE annotation was performed using EDTA with the settings --overwrite 1 --sensitive 1 --anno 1 --evaluate 1 and using the current version of TREP (v.19) as a curated input library. The identified TEs were subsequently extracted from the assemblies and BLAST searched against TREP to make family-level classifications (because, for full-length elements, EDTA is sometimes hesitant to assign a TE family and instead gives a unique identifier and superfamily tag). To determine the TE content of einkorn centromeres, centromeric DNA was annotated using the automated TE annotation software EDTA[77] with TREP as curated input library[78]. For the analysis of genes in centromeres, all annotated genes that showed homology to TEs were removed, as it is very probable that they are TEs that were annotated as genes by mistake.

***RLG_Cereba* dating.** Full-length copies of *RLG_Cereba* and *RLG_Quinta* retrotransposons were identified using our previously described pipeline[105]. This was done in addition to the automated TE annotation described above to extract high-quality datasets of full-length TEs. The insertion ages of full-length LTR retrotransposons were determined on the basis of the divergence of the two LTRs, which are identical at the time of insertion, and which accumulate mutations over time[106]. This produced information on insertion age and precise chromosomal location for each full-length *RLG_Cereba* retrotransposon.

Chromosome collinearity and similarity analysis. Collinearity of chromosomal segments was visualized using 1 kb sequence segments of one chromosome in BLASTN searches against the other chromosome. The positions of the top BLASTN hits were used for the dot plot alignments. A sliding step of 10 kb was used. The chromosome comparison was done using the original Perl script blast_compare_chromosome that is available at GitHub (https://github.com/Wicker-Lab/Monococcum_genome_scripts). To complement our homology and annotation-based repeat analysis, we used the ChIP-Seq mapper tool, which is part of the RepeatExplorer2 software collection. First, repeats were identified by clustering short sequencing reads with RepeatExplorer2, which does not depend on the reference genome and would therefore also identify repetitive elements missing from the reference. For this, around 20× coverage Illumina short reads were downsampled to the recommended coverage equivalent of 0.5× using seqkit (https://github.com/shenwei356/seqkit). CENH3 ChIP and control reads were then mapped onto the identified repeat clusters using the ChIP-Seq mapper tool. Two repeat clusters passed the threshold of greater than fivefold enrichment. The unique sequences contained in these two clusters were then queried using BLASTN searches against the nrTREP20 repeat database. In one of the clusters, all 38 sequences showed very high homology to either the *RLG_Cereba* or *RLG_Quinta* consensus sequence, whereas, in the other cluster, 37 out of 42 sequences were *RLG_Cereba* or *RLG_Quinta* sequences. Thus, ChIP-Seq mapper did not identify any additional repeat clusters enriched in CENH3 that were not found in the homology and annotation-based repeat analysis.

Whole-genome sequencing of the einkorn diversity panel

Illumina short-read sequencing. We extracted genomic DNA from one or two young leaves of 219 einkorn accessions using CTAB extraction[46]. DNA quantification was performed using the Qubit dsDNA HS Assay (Q32851, Thermo Fisher Scientific), the purity was assessed using the Nanodrop spectrophotometer by checking the 260/280 and 260/230 ratios and the integrity was confirmed by analysing 1 µl per sample on a 1% TAE agarose gel. Library preparation and sequencing (150 bp paired-end libraries) were performed by Novogene using the Illumina NovaSeq 6000 system.

Read mapping and SNP calling. Whole-genome sequencing data for the 219 einkorn accessions were first trimmed using trimmomatic[99] (v.0.38; parameters: LEADING:20 TRAILING:20 SLIDINGWINDOW:5:20 MINLEN:5). Reads were mapped to the TA299 reference assembly using BWA[96] mem (v.0.7.17). The mapped reads were then sorted according to genomic coordinates using the SAMtools[69] command sort (v.1.8). Duplicated reads were marked and read groups were assigned using the Picard tools (http://broadinstitute.github.io/picard/). HaplotypeCaller from GATK[107] (v.4.1.8.0) was used to identify variants and generate individual-specific .gvcf files followed by a joint calling of variants performed by GenotypeGVCFs. We extracted SNPs using the GATK SelectVariants command. SNPs were hard filtered using VariantFiltration removing putative variants according to the following criteria: 'QD < 2.0 || FS > 60.0 || MQ < 40.00 || MQRankSum < −12.5 || ReadPosRankSum < −8.0 || SOR > 3.0'. In total, 208,855,939 SNPs were called from 219 einkorn accessions. After quality control using VCFtools[108] (v.0.1.17), the raw SNPs were filtered using GATK[107] (v.4.1.8.0) and VCFtools[108] (v.0.1.17) as follows: SNP clusters, defined as three or more SNPs located within 10 bp; low and high average SNP depth ($4 \leq DP \geq 15$); and SNPs located in the unanchored chromosome were removed. Moreover, one misclassified accession (TA574; initially was classified as γ) was removed on the basis of PCA and divergence analysis. Finally, only biallelic SNPs were retained for further analyses, representing a final VCF file of 121,459,674 SNPs (Supplementary Table 15). These SNPs were annotated using snpEff[109] (v.5.0e) with TA299 HC gene models. The false-positive error rate of variant calling (percentage of polymorphic sites in a resequenced TA299 sample compared with the TA299 reference) was 0.008%, which is comparable to the error rates of other studies[43–46] (Supplementary Fig. 19a). Variants were evenly distributed across the seven chromosomes, except for the centromeres that showed a marked reduction in variant densities due to reduced read mapping (Supplementary Fig. 19b, Supplementary Fig. 20 and Supplementary Table 16). Approximately 2.2% of the total SNPs were gene-proximal (2 kb upstream and downstream of a coding sequence). An additional 0.8% of the SNPs were located in introns and 0.5% in exons. Of the exonic SNPs, 317,023 (53.4%) were non-synonymous affecting 26,505 genes, of which 9,145 SNPs resulted in a disruption of coding sequences (premature stop codon) in 5,726 genes. Furthermore, 45.7% of the total SNPs (55,558,212 SNPs) represented rare variants with a minor allele frequency below 1% (Supplementary Fig. 19c and Supplementary Table 17). Variant calling using the TA10622 assembly revealed very similar results on the basis of population divergence, PCA and nucleotide diversity (α, $\pi = 0.0012$; β, $\pi = 0.0017$; γ, $\pi = 0.0022$; domesticated, $\pi = 0.0012$; Supplementary Fig. 21a–c), confirming the high accuracy of variant calling and the independence of population structure analyses from which reference assembly is used. The SNP calling against the TA10622 reference assembly was used for the analyses presented in Extended Data Fig. 7a,b,e.

Mapped reads and SNP data statistics. Mapping statistics for each accession were calculated from the BAM files using SAMtools[69] (v.1.8; option 'flag-stat') to get the number of mapped reads, and the mapping rate was then calculated as follows: (the number of mapped reads/the total number of reads) × 100 . The false-positive error rate of SNP calling was calculated as the proportion of segregating sites in a resequenced TA299 sample compared with the TA299 reference assembly. The numbers of homozygous reference, heterozygous reference and alternative alleles were obtained using VCFtools[108] (v.0.1.17) and awk command-line. The SNP density was calculated in bin sizes of 1 Mb, and the nucleotide diversity was calculated in sliding windows of 10,000 bp per chromosome and then averaged across the entire genome to measure the degree of polymorphism within each einkorn population using VCFtools[108] (v.0.1.17).

Population diversity and structure. We assessed the genetic relationships between accessions with PCA using all SNPs (121,459,674) with PLINK[110] (v.1.90). The unrooted neighbour-joining phylogenetic tree was generated from filtered SNPs (missing data > 10% and 5% randomly sampled SNPs; total SNPs = 5,318,268). First, the genetic distances were computed using Euclidean distances with the 'dist' function in the stats R package. The distance matrix was converted to a phylo object using the R package ape and the tree was generated using the phyclus R package. For estimating individual ancestry coefficients, the R package LEA 'snmf' function was used with the entropy option and with 10 independent runs for each K (K is the number of putative ancestral populations) from $K = 1$ to $K = 10$ using the same SNP subset used to generate the phylogenetic tree. The cross-entropy value decreased with increasing K and reached a plateau starting from $K = 6$ (Supplementary Fig. 14).

F_{ST} calculation. We defined the two domesticated einkorn groups on the basis of an ancestry threshold of 80% at $K = 4$ (because the split of the two domesticated einkorn groups occurred at $K = 4$). We then calculated the mean fixation index (F_{ST}) between these two domesticated einkorn groups in 1 Mb non-overlapping genomic windows using VCFtools[108] (v.0.1.17).

Wild einkorn γ race introgression analyses. We evaluated the divergence of domesticated einkorn accessions from the TA10622 reference assembly by calculating the proportion of segregating sites using only the diverged blocks (chromosome 2A: 261–406 Mb; chromosome 5A: 92–409 Mb; and chromosome 7A: 301–448 Mb). Moreover, we

performed PCA using all einkorn accessions with only SNPs located in the diverged regions. To estimate the proportion of γ race introgression into all domesticated einkorn accessions, we calculated pairwise nucleotide diversity in 1 Mb non-overlapping windows with VCFtools[108] (v.0.1.17) between one γ accession (TA10600) and each of the domesticated einkorn accessions. Accession TA10600 was selected because it showed a low divergence from the introgressed segments. A region was considered as introgression if (1) there was a continuous nucleotide diversity reduction for ≥10 Mb; (2) nucleotide diversity reduction in the region was not observed between an α race accession and a domesticated accession; and (3) the reduction of nucleotide diversity was not due to the lack of mapped reads. To identify γ accessions with the closest genetic relatedness to the introgressed segments in domesticated einkorn, we first performed a PCA using only wild γ race and domesticated einkorn accessions with the introgressed genomic regions (chromosome 2A: 261–406 Mb; chromosome 5A: 92–409 Mb; and chromosome 7A: 301–448 Mb, separately). On the basis of the clustering of accessions in the PCA, the geographical projection of the first (for the regions on chromosome 2A and 7A) and the second PCA axes (for the region on chromosome 5A) was done and visualized using the Kriging function in the fields v.10.3 R package (https://cran.r-project.org/web/packages/fields).

TreeMix analysis. We included *T. urartu* whole-genome sequencing data[44] in the analysis, and SNP calling was performed as described above. We filtered all missing SNPs, and we used PLINK[110] (v.1.90) to remove SNPs in linkage disequilibrium (parameter: indep-pairwise 100 5 0.2). The total number of SNPs retained for the TreeMix analysis was 1,042,531. To infer splits and admixture events, we first obtained a list of einkorn groups considering 80% ancestry threshold at $K = 6$. We then used TreeMix (v.1.13)[49] using jackknife blocks of 1,000 SNPs and modelling 5 migration events.

k-mer based approaches to detect einkorn introgressions in bread wheat

We used two *k*-mer-based approaches to identify putative einkorn introgressions into bread wheat.

***k*-mer mapping approach.** For generating *k*-mer datasets, we used the whole-genome sequencing data from all domesticated einkorn accessions in the panel and *T. urartu* accessions[44]. *k*-mers ($k = 51$) were counted from the Illumina raw data per accession using jellyfish (v.2.2.10)[111]. We extracted the *k*-mer nucleotide sequences from each accession of *T. monococcum* and *T. urartu*. We concatenated all *k*-mer sequences from all *T. monococcum* accessions and *T. urartu* accessions into one separate file for each species and retained one representative per *k*-mer. We removed common *k*-mers between *T. monococcum* and *T. urartu* and obtained a list of specific einkorn and *T. urartu k*-mer sequences, respectively. The lists of specific *k*-mers were later converted into fasta files. Each fasta file (*T. monococcum* and *T. urartu*) was mapped against the bread wheat genomes[23] using BWA[96] mem (v.0.7.17), requiring mapping of only full-length *k*-mers with no mismatches. Mapped *k*-mers in each .bam file (*T. monococcum* and *T. urartu*) were analysed for the coverage in genomic windows of 1 Mb using mosdepth[112] and visualized in R (v.4.0.4) using ggplot2. Putative introgressions were identified as an increased coverage of mapped *k*-mers from *T. monococcum* (with an average coverage of ≥5), but depleted mapping of *T. urartu*-specific *k*-mers. Two or more regions were grouped into one if they were no more than 1 Mb apart.

***k*-mer variation approach.** We implemented Identity-by-State Python (IBSpy; https://github.com/Uauy-Lab/IBSpy), a *k*-mer-based pipeline that counts variations in 50 kb windows, and used it to detect *T. monococcum* introgressions into the ten bread wheat genomes[23] (Supplementary Note 2). We first used KMC3[113] to build a *k*-mer ($k = 31$) databases from the Illumina raw data of 218 *T. monococcum* accessions, the two *T. monococcum* chromosome-scale assemblies and ten genome assemblies of wheat. We next compared the *k*-mers of the bread wheat reference sequence to the *k*-mers of each database and counted the number of variations within each 50 kb window. We used variations ≤ 30 as a cut-off (Supplementary Fig. 22 and Supplementary Note 2). We considered the six STRUCTURE groups to identify the putative donors for each introgressed segment (Fig. 3b, Supplementary Fig. 13 and Supplementary Note 2).

Positional cloning of *tin3*

Phenotyping and DNA isolation of tin3 mutants. A cross between the *tin3* mutant and the parental accession TA4342-L96 was made. A total of 375 F_2 plants were grown and phenotyped for tiller number (Zadoks growth scale ~Z29). Leaf tissues were harvested and DNA was extracted using the BioSprint 96 DNA Plant Kit (Qiagen, 576) on a KingFisher Flex robot (Thermo Fisher Scientific, 5400610) according to the manufacturer's instructions. Equimolar concentrations of 30 F_2 plants showing the *tin3* mutant phenotype were pooled to create a mutant bulk. SEM samples were fixed in 3% glutaraldehyde, post-fixed in 1% osmium tetroxide and dehydrated in a graded acetone series, and these samples were processed and SEM images were captured as described previously[114,115].

Whole-genome sequencing, alignment and SNP calling. Sequencing libraries were prepared, and sequencing was performed by Novogene using the Illumina NovaSeq 6000 platform. Samples were sequenced to >13-fold coverage. We used the Illumina reads of TA4342-L96 (Sequence Read Archive: SRR21543761) as the parental control. We followed the MutMap protocol with minor modifications[57]. High-quality filtered reads were aligned to the *T. monococcum* accession TA10622 using BWA[96]. SAM files were converted into .bam files using SAMtools[69]. SAMtools (markdup option) was used to mark and remove PCR duplicates. Improperly mapped read pairs were removed from the .bam files retaining only concordantly aligned reads with MAPQ ≥ 30. The BCFtools mpileup tool was used for SNP calling[70]. SNPs were filtered on the basis of the following criteria: minQ ≥ 30, Fisher Strand (FS) > 40, mapping quality (MQ < 40), minDP > 3 and genotype quality (GQ < 20). SNPs within 10 bp proximity of indels were removed and only the biallelic SNPs were retained. SNP positions with an identical allele in both TA4342-L96 and the *tin3* mutant bulk were treated as varietal SNPs and were removed from the analysis. SnpSift[109] was used to select EMS-type (G/C to A/T) transitions from the VCF file. We considered the positions with a SNP index of ≥0.9 to be homozygous, whereas SNPs with an SNP index of <0.3 were removed, and the rest were considered to be heterozygous. We used the mutplot tool (https://github.com/VivianBailey/Mutplot) to calculate the average SNP index using a window size of 100 kb[116]. The average SNP index was plotted along the chromosomes using ggplot2[117]. SnpEff 5.0c (build 2020-11-25 14:23) was used to calculate the effect of the variants on genes.

KASP assay. A KASP marker was designed on the basis of the SNP in *Tm.TA10622.r1.3AG0164370* using Primer3 software (primer F1, 5′-**GAAGGTGACCAAGTTCATGC**TCGCCGTCCTCACCCAGa-3′; primer F2, 5′-**GAAGGTCGGAGTCAACGGAT**TCGCCGTCCTCACCCAGg-3′-; primer R, 5′-ATCCCAACATAACCGACCCC-3′; the HEX and FAM tails are marked in bold, and the lower case base indicates the allele-specific SNP). KASP assays were performed using KASP-TF V4.0 2× Master Mix (LGC Bioscience Technologies, KBS-1050-102) with a 5 µl reaction in the Bio-Rad C1000 thermal cycler with a CFX96 module according to the manufacturer's instructions. Allelic discrimination was called using CFX Maestro Software (v.1.1).

TILLING in hexaploid wheat cv. 'Jagger'. The TILLING population used in this study was first reported previously[59]. Markers were designed for

the A, B and D subgenome homeologous copies of *tin3* (*tin3A*, *tin3B*, *tin3D*) using GSP[118] and tested for specificity using nullisomic tetrasomic wheat lines. PCR products were amplified for each target gene on 4× pools. These products were heteroduplexed to form base pair mismatches and digested using homemade Cel-1 endonuclease. Individual target genes were then amplified from positive pools. Heterozygous mutants, such as in *tin3D*, were grown for an additional generation to retrieve homozygous mutants. Zygosity was confirmed by isolating DNA from the segregating plants and Sanger sequencing. Crosses between the different mutants were made and plants identified as homozygous for single mutations (that is, *tin3A*, *tin3B*, *tin3D*), double mutations (*tin3AB*, *tin3AD*, *tin3BD*) and triple mutations (*tin3ABD*) were selected. Plants were phenotyped for tiller numbers 8 weeks after vernalization.

Germplasm availability

The *T. monococcum* accessions used in this study are listed in Supplementary Table 8 and are available on request from the WGRC, Kansas State University (www.k-state.edu/wgrc).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw sequencing data used for de novo genome assemblies, the RNA-seq and Iso-seq data for the annotation, the ChIP–seq reads and the whole-genome sequencing reads of 218 einkorn accessions are available at the EBI-ENA under study number PRJEB61155. The two reference assemblies, annotations, VCF files and CpG methylation frequencies are available at DRYAD (https://doi.org/10.5061/dryad.v41ns1rxj). Raw fastq files and demultiplexed fastq files of the RIL population have been deposited at the National Center for Biotechnology Information (NCBI) SRA database under BioProject accession PRJNA879879. The barcode indices key file with required information for demultiplexing can be obtained at DRYAD (https://doi.org/10.5061/dryad.v41ns1rxj). CENH3 BED files (CENH3 peaks) and mapped files (.bam) of CENH3 and H3K4me3 for all replicates are available through the Dryad database (https://doi.org/10.5061/dryad.0p2ngf24b). Whole-genome sequencing data of the *tin3*-mutant bulk have been deposited at GenBank under BioProject PRJNA938447. IBSpy variations tables of the 218 einkorn accessions, and MUMer alignments of the two einkorn assemblies against the ten bread wheat assemblies are available online (https://opendata.earlham.ac.uk/wheat/under_license/toronto/Uauy_2022-09-24_IBSpy_Triticum_monococcum_introgressions/). An interactive webpage has been developed for this study to visualize various genome characteristics of TA299 and TA10622. The webpage features a JBrowse 2 explorer enabling visualization of the whole genome, gene models, transposable elements, variants positions and synteny. For the identification of homologous sequences in other wheat varieties, a BLAST server has been set up. This BLAST server enables searches against individual wheat subgenomes and chromosomes independently. Synteny between TA10622 and other wheat genomes can be visualized in the synteny tab located on the webpage. The database can be accessed online (https://wheat.pw.usda.gov/GG3/pangenome). Source data are provided with this paper.

Code availability

Custom scripts for variant calling are available at GitHub (https://github.com/IBEXCluster/Wheat-SNPCaller). The necessary codes and steps to genotype the RIL population were published previously[67] and are available in a separate file on the DRYAD database (https://datadryad.org/stash/share/v20dkVsSTj3toGn-CHG92eUSgre17uMT5AH_6LE2GDM). Codes and custom scripts for analysing einkorn introgressions into bread wheat are available at GitHub (https://github.com/Uauy-Lab/monococcum_introgressions).

60. Mayjonade, B. et al. Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. *Biotechniques* **61**, 203–205 (2016).
61. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
62. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
63. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
64. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
65. 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).
66. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
67. Adhikari, L. et al. A high-throughput skim-sequencing approach for genotyping, dosage estimation and identifying translocations. *Sci. Rep.* **12**, 17583 (2022).
68. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
69. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
70. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
71. Agarwal, G. et al. A recombination bin-map identified a major QTL for resistance to Tomato Spotted Wilt Virus in peanut (*Arachis hypogaea*). *Sci. Rep.* **9**, 18246 (2019).
72. Athiyannan, N. et al. Long-read genome sequencing of bread wheat facilitates disease resistance gene cloning. *Nat. Genet.* **54**, 227–231 (2022).
73. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
74. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
75. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
76. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. Bioinformatics* **3**, lqaa108 (2021).
77. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
78. Wicker, T., Matthews, D. E. & Keller, B. TREP: a database for Triticeae repetitive elements. *Trends Plant Sci.* **7**, 561–562 (2002).
79. Ling, H.-Q. et al. Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. *Nature* **557**, 424–428 (2018).
80. Luo, M.-C. et al. Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature* **551**, 498–502 (2017).
81. Avni, R. et al. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* **357**, 93–97 (2017).
82. Mascher, M. et al. Long-read sequence assembly: a technical evaluation in barley. *Plant Cell* **33**, 1888–1906 (2021).
83. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
84. 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).
85. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
86. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
87. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
88. Quevillon, E. et al. InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116–W120 (2005).
89. Pérez-Wohlfeil, E., Diaz-del-Pino, S. & Trelles, O. Ultra-fast genome comparison for large-scale genomic experiments. *Sci. Rep.* **9**, 10274 (2019).
90. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
91. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
92. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).
93. Nagaki, K. et al. Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of *Arabidopsis thaliana* centromeres. *Genetics* **163**, 1221–1225 (2003).
94. Koo, D.-H., Sehgal, S. K., Friebe, B. & Gill, B. S. Structure and stability of telocentric chromosomes in wheat. *PLoS ONE* **10**, e0137747 (2015).
95. Ni, P. et al. DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. *Nat. Commun.* **14**, 4054 (2023).
96. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
97. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).
98. Ramirez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).

99. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
100. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
101. Stovner, E. B. & Saetrom, P. epic2 efficiently finds diffuse domains in ChIP-seq data. *Bioinformatics* **35**, 4392–4393 (2019).
102. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
103. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
104. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2013).
105. Wicker, T. et al. Transposable element populations shed light on the evolutionary history of wheat and the complex co-evolution of autonomous and non-autonomous retrotransposons. *Adv. Genet.* **3**, 2100022 (2021).
106. SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y. & Bennetzen, J. L. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43–45 (1998).
107. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
108. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
109. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).
110. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
111. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
112. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
113. Kokot, M., Długosz, M. & Deorowicz, S. KMC 3: counting and manipulating *k*-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).
114. Venglat, P. et al. Gene expression analysis of flax seed development. *BMC Plant Biol.* **11**, 74 (2011).
115. Venglat, S. P. et al. The homeobox gene BREVIPEDICELLUS is a key regulator of inflorescence architecture in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **99**, 4730–4735 (2002).
116. Sugihara, Y. et al. High-performance pipeline for MutMap and QTL-seq. *PeerJ* **10**, e13170 (2022).
117. Wickham, H. *ggplot2—Elegant Graphics for Data Analysis* (Springer, 2016).
118. Wang, Y. et al. GSP: a web-based platform for designing genome-specific primers in polyploids. *Bioinformatics* **32**, 2382–2383 (2016).

Acknowledgements We thank the members of the KAUST Bioscience Core Laboratory for sequencing support; E. Cavalet-Giorsa for providing information on einkorn domestication and migration; L. Aouini for assistance with RNA extraction; L. Zou for greenhouse support; B. Gill for providing seeds of the *tin3* mutant; and T. Quilichini for SEM images. We acknowledge support by the plant growth facility and the GENTYANE platform of the Clermont-Ferrand INRAE Centre for assisting in NGS sequencing; the Genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul, http://bioinfo.genotoul.fr) and the KAUST supercomputing facilities (https://www.hpc.kaust.edu.sa) for providing computing resources; GrainGenes resources for hosting the online database; and the University of Maryland supercomputing resources (http://hpcc.umd.edu) for developing the einkorn database and *tin3* analysis. This publication is based on work supported by the King Abdullah University of Science and Technology, the UK Biotechnology and Biological Sciences Research Council (BBSRC; BB/P016855/1), the Mexican Consejo Nacional de Ciencia y Tecnología (CONACYT; 2018-000009-01EXTF-00306), the Global Institute for Food Security (to R.D.), the European Research Council (ERC-2019-COG-866328) and the United States Department of Agriculture National Institute of Food and Agriculture (USDA-NIFA; award 2020-67013-31460). D.-H.K. was supported by WGRC/IUCRC and NSF (grant 1822162).

Author contributions H.I.A., V.K.T., M.A., J.P. and S.G.K. designed the research. J.R. and V.K.T. maintained and provided plant materials. S.C., N. Rodde, C. Cravero and C. Callot performed sequencing and produced optical maps. H.I.A., M.A., S.C. and C. Cravero generated assemblies. L.A. selected two accessions for assemblies and the einkorn diversity panel and constructed genetic maps. D.-H.K. produced ChIP–seq data. H.I.A., N.A. and G.S. performed molecular experiments. M.A. performed genome annotations. N.K. prepared scripts for variant calling. H.I.A. analysed and validated genome assemblies, analysed whole-genome sequencing data and performed population genomics analyses. H.I.A., J.Q.-C., R.H.R.-G. and C.U. analysed einkorn introgressions into bread wheat. M.H., T.W. and J.P. performed centromere analyses. P.K.S., G.R.L. and V.K.T. developed the einkorn database. V.K.T. conceived the *tin3* experiments. A.S., I.M., I.S.Y., L.S. and R.D. performed *tin3* experiments and generated data. N. Rawat contributed *tin3* TILLING results in hexaploid wheat. H.I.A., M.H., T.W. and S.G.K. wrote the initial manuscript with input from J.Q.-C., L.A., C.U., M.A. and J.P. All of the authors have read and approved the final manuscript.

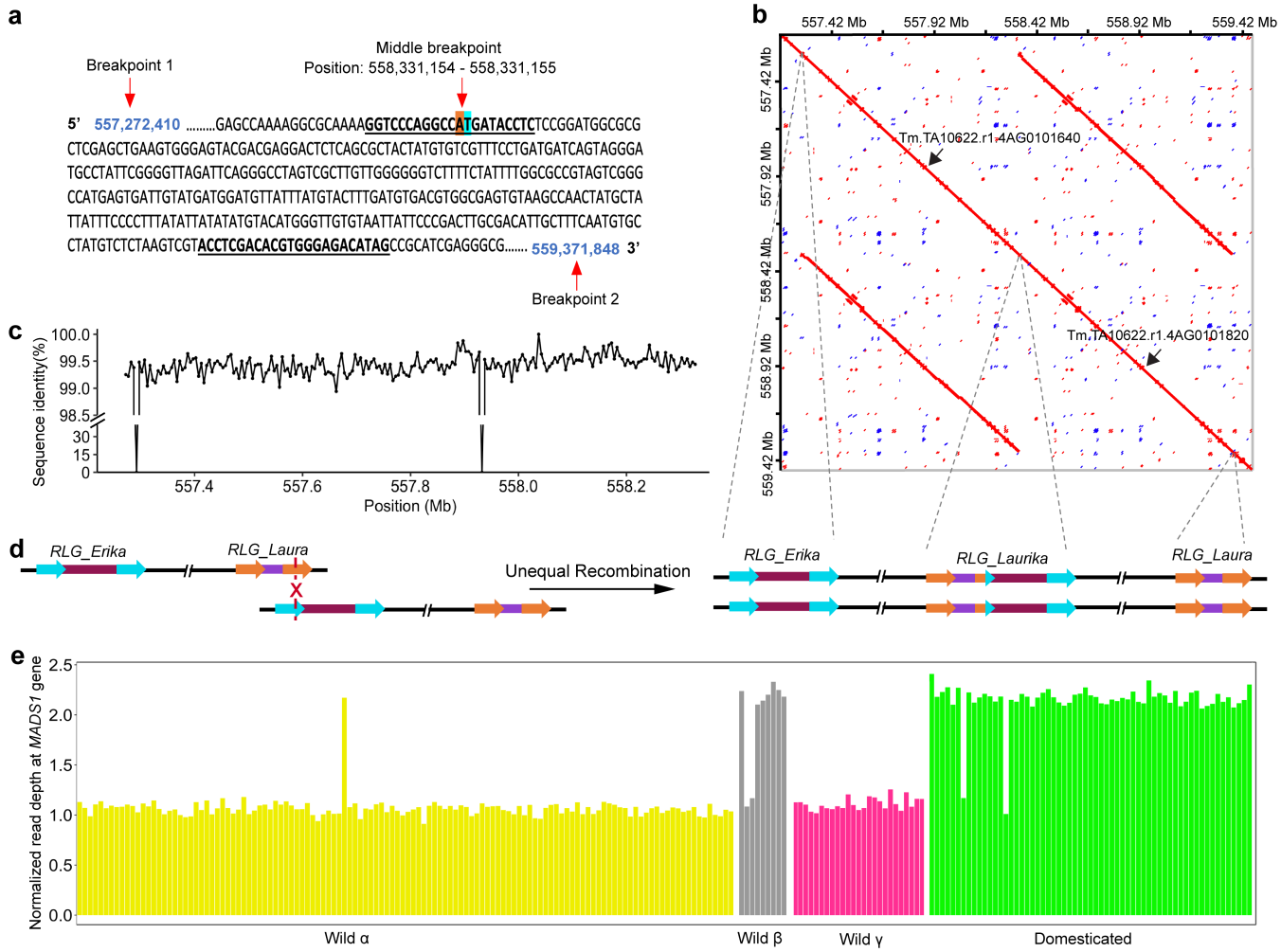
Competing interests The authors declare no competing interests.

Additional information

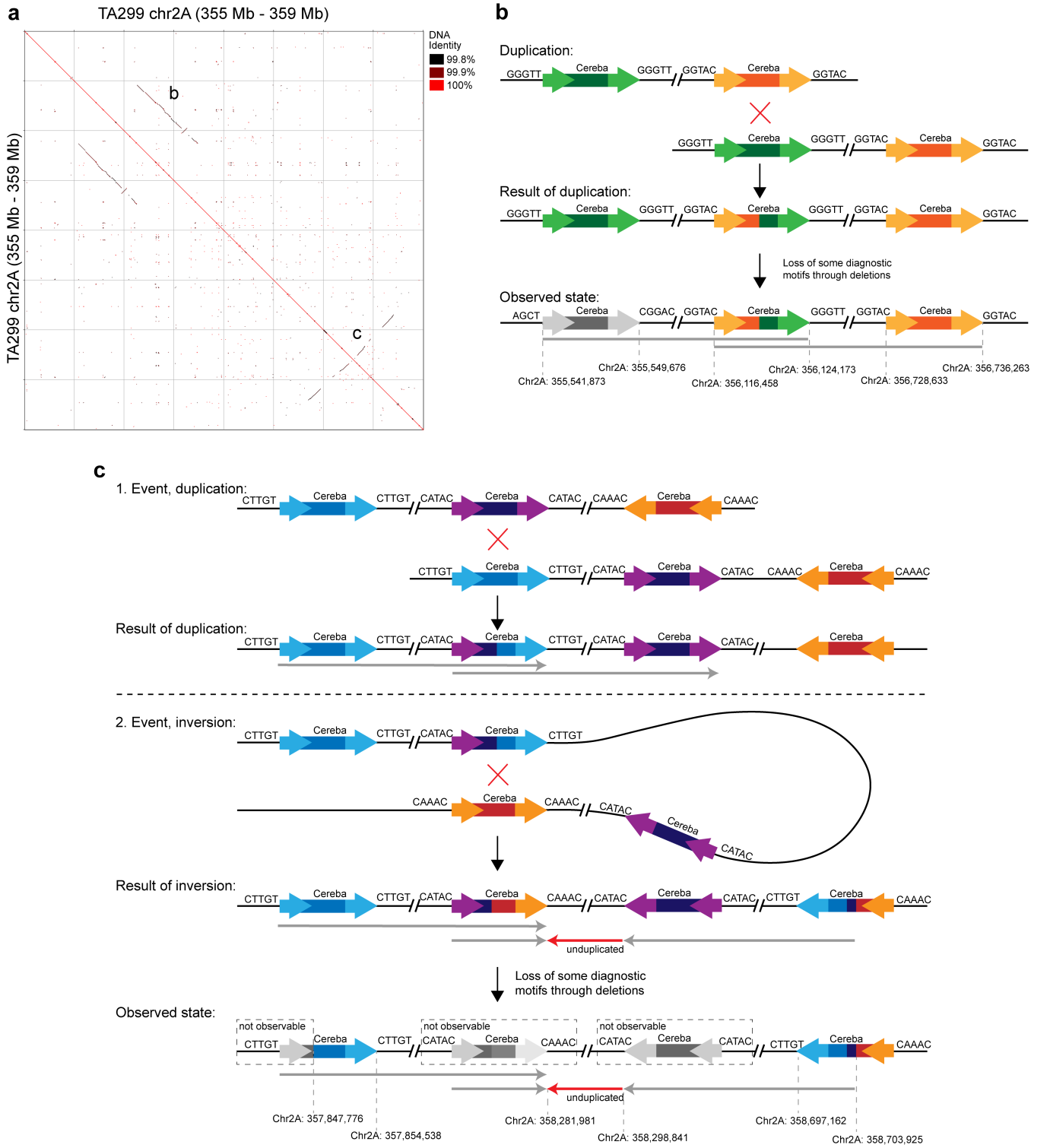
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-06389-7.

Correspondence and requests for materials should be addressed to Vijay K. Tiwari, Michael Abrouk, Jesse Poland or Simon G. Krattinger.

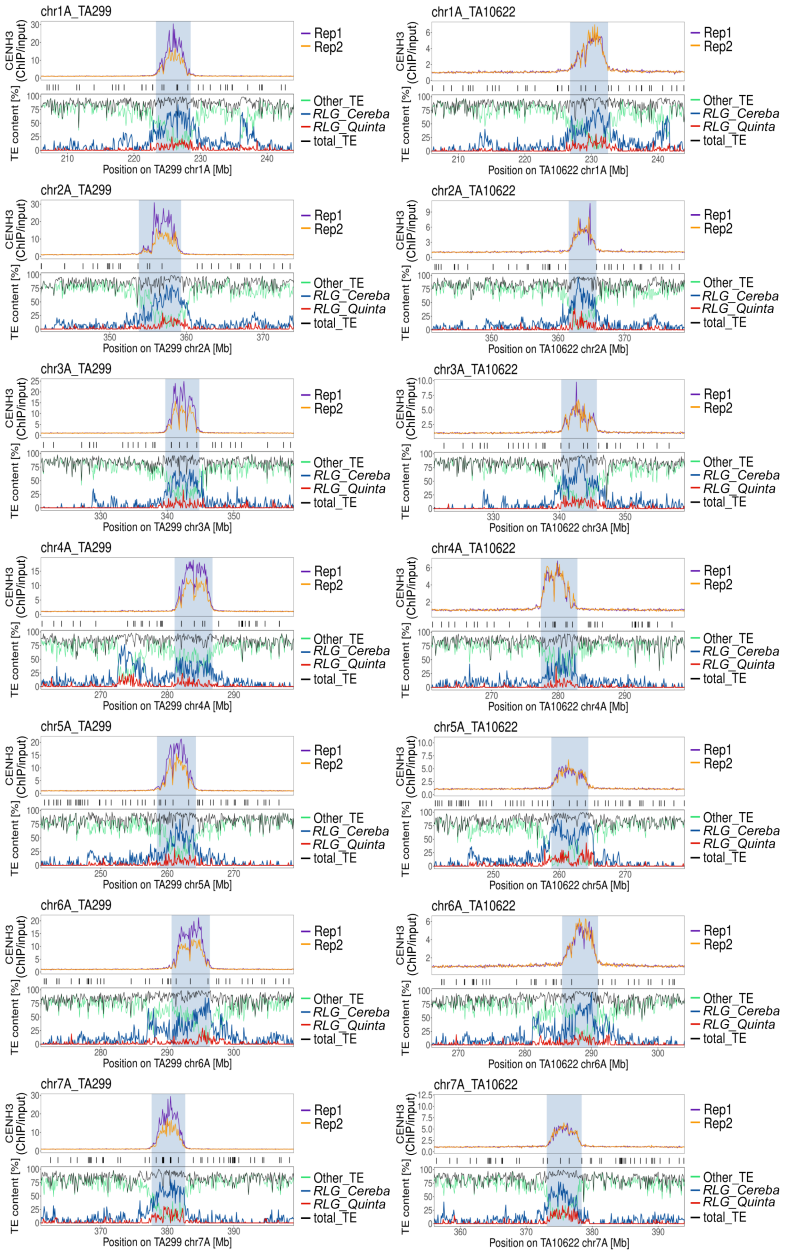
Peer review information *Nature* thanks André Marques and the other, anonymous, reviewer(s) for their contribution to this work.



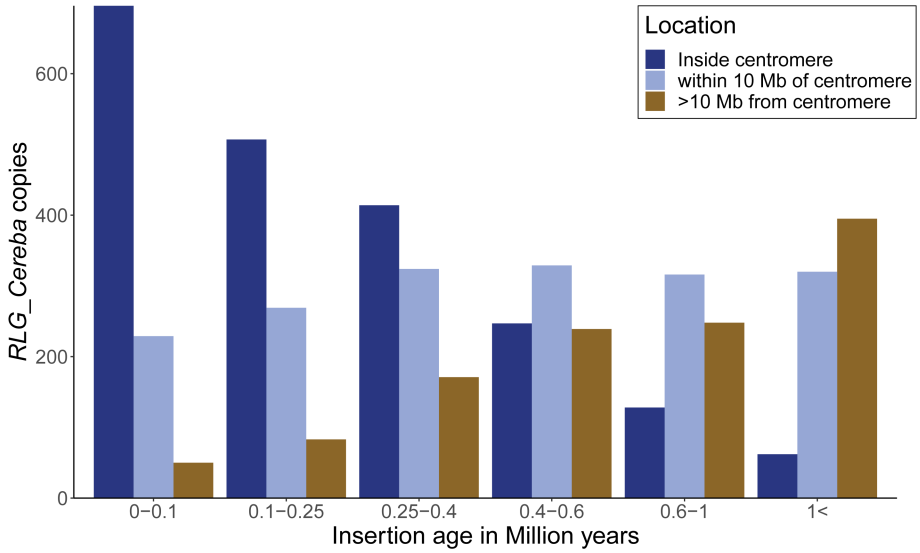
Extended Data Fig. 1 | Characterization of a large tandem duplication in einkorn. a, Sequence around the middle breakpoint of the two tandemly duplicated segments on chromosome 4A of TA10622. The 1 Mb duplication was confirmed by designing a PCR marker across the breakpoint. Primer sequences are underlined and indicated in bold. The nucleotides at the breakpoint are highlighted in orange (located in the *RLG_Laura* element) and in blue (located in the *RLG_Erika* element). **b**, Dot plot showing a comparison of a 2.3 Mb region of chromosome 4A of TA10622 against itself. The two red lines indicate the megabase-sized tandem duplication. The positions of the MADS-box transcription factor genes are indicated by black arrows. A schematic representation of the retroelements is shown at the bottom. Arrows indicate long terminal repeats (LTRs). **c**, Sequence identity across the two duplicated segments, calculated in 5 kb non-overlapping windows. **d**, Schematic diagram showing the proposed unequal recombination between two retrotransposons that led to the tandem duplication. **e**, The presence - absence of the tandem duplication on chromosome 4A was estimated by normalized read coverage across the *MADS1* gene (including 2 kb of flanking sequence). Outliers: α: TA316, β: TA10573 and TA10910, domesticated einkorn: TA10548 and TA10577.



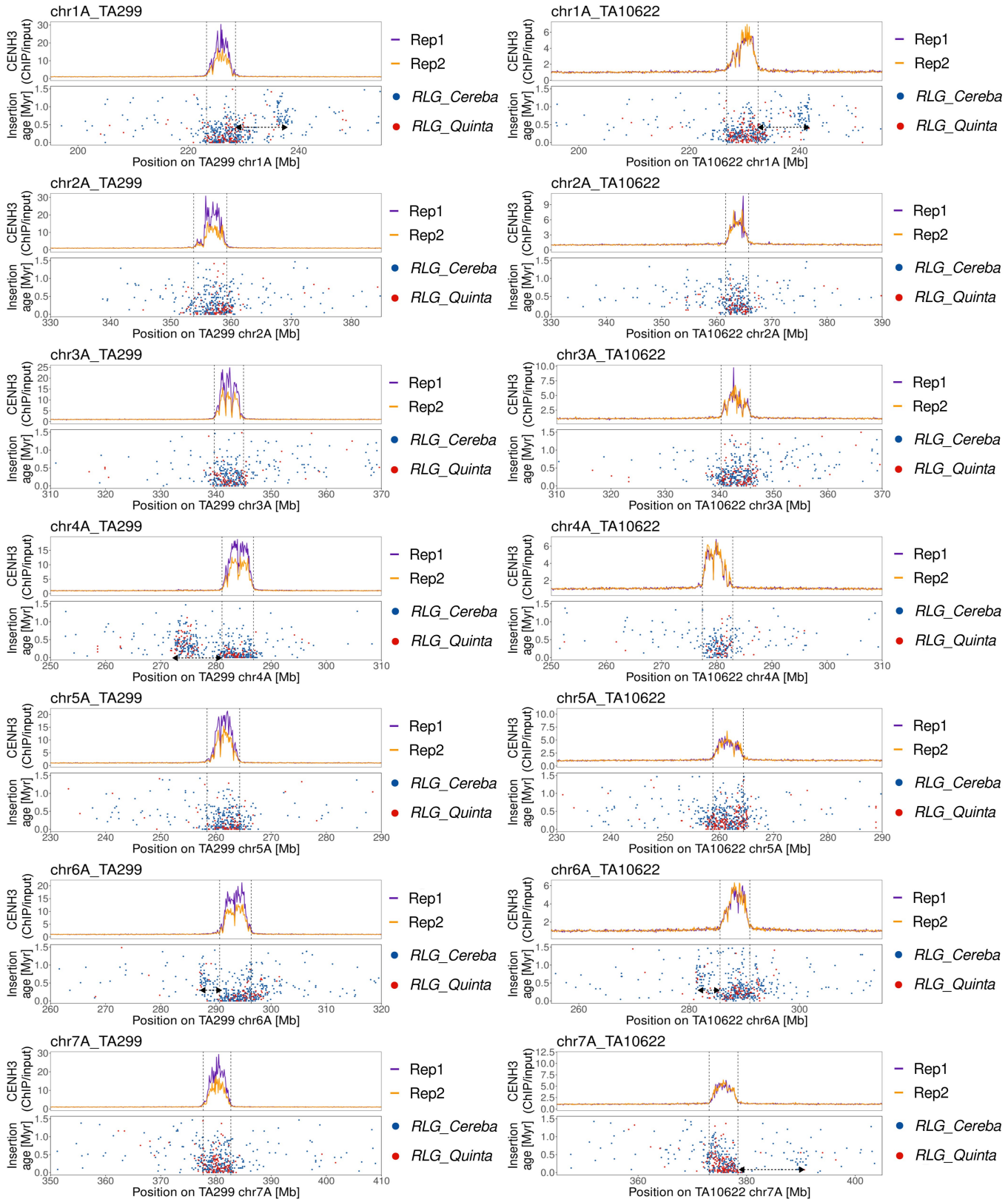
Extended Data Fig. 2 | Evolutionary origin of large-scale duplications and inversions in the centromere of TA299 chromosome 2A. a, Dot plot alignment of a segment inside the centromere of chromosome 2A. A large duplication is labelled with "b" and a duplication/inversion with "c". **b**, Evolutionary model for the duplication event b. We propose that the large duplication originated from unequal recombination between two *RLG_Cereba* retrotransposons that were ~700 kb apart, resulting in the duplication of the entire sequence between them. The *RLG_Cereba* elements that served as templates for the unequal recombination are shown in red and green. The 5 bp target site duplications (TSD) produced by their insertions were used as diagnostic sequences to identify the recombinant element in the centre of the two duplicated units. Parts of the duplication that were deleted in later events are shown in grey (this includes one of the *RLG_Cereba* copies that served as a template for the initial event). **c**, The second duplication/inversion "c" occurred following the same mechanisms. The duplication was followed by a second independent event resulting in an inversion that affected nearly the same region. The inversion was caused by recombination between different *RLG_Cereba* elements near the original duplication breakpoints but in the opposite orientation. Subsequently, deletions near the borders of the inverted segment removed some of the diagnostic motifs (indicated in grey). We emphasize that the presented model is based exclusively on homology-based recombination between different *RLG_Cereba* retrotransposons and that it is possible that events resulting in a duplication plus an inversion may also involve alternative mechanisms.



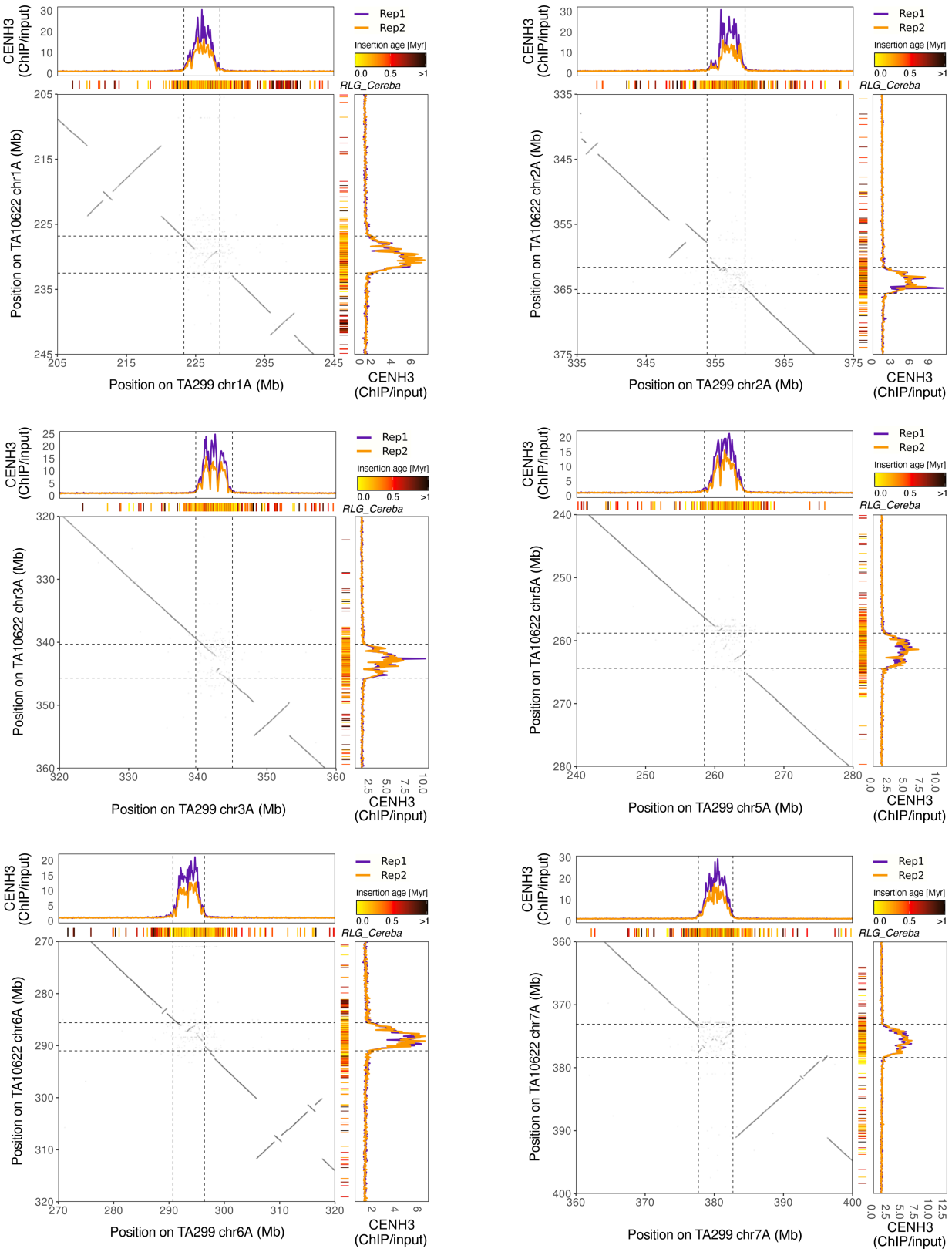
Extended Data Fig. 3 | Definition of functional centromeres (shaded in blue) in *T. monococcum* accessions TA299 (left column) and TA10622 (right column) from CENH3 ChIP-Seq data. Centromeric regions plus ~15 Mb of flanking regions are shown. The top graph of each panel shows the ratio of CENH3 ChIP-Seq reads divided by input control reads. ChIP-Seq and control reactions were performed in duplicates, which are shown in purple (Rep1) and orange (Rep2). The region of functional centromeres was determined based on epic2 ChIP-Seq peak call density and is indicated by a shaded area. The middle track below shows the positions of genes as vertical lines. The bottom graph shows the average transposable element (TE) content in 100 kb windows. Note that *RLG_Cereba* and *RLG_Quinta* retrotransposons are highly enriched in functional centromeres, while other TE families dominate outside of centromeres.



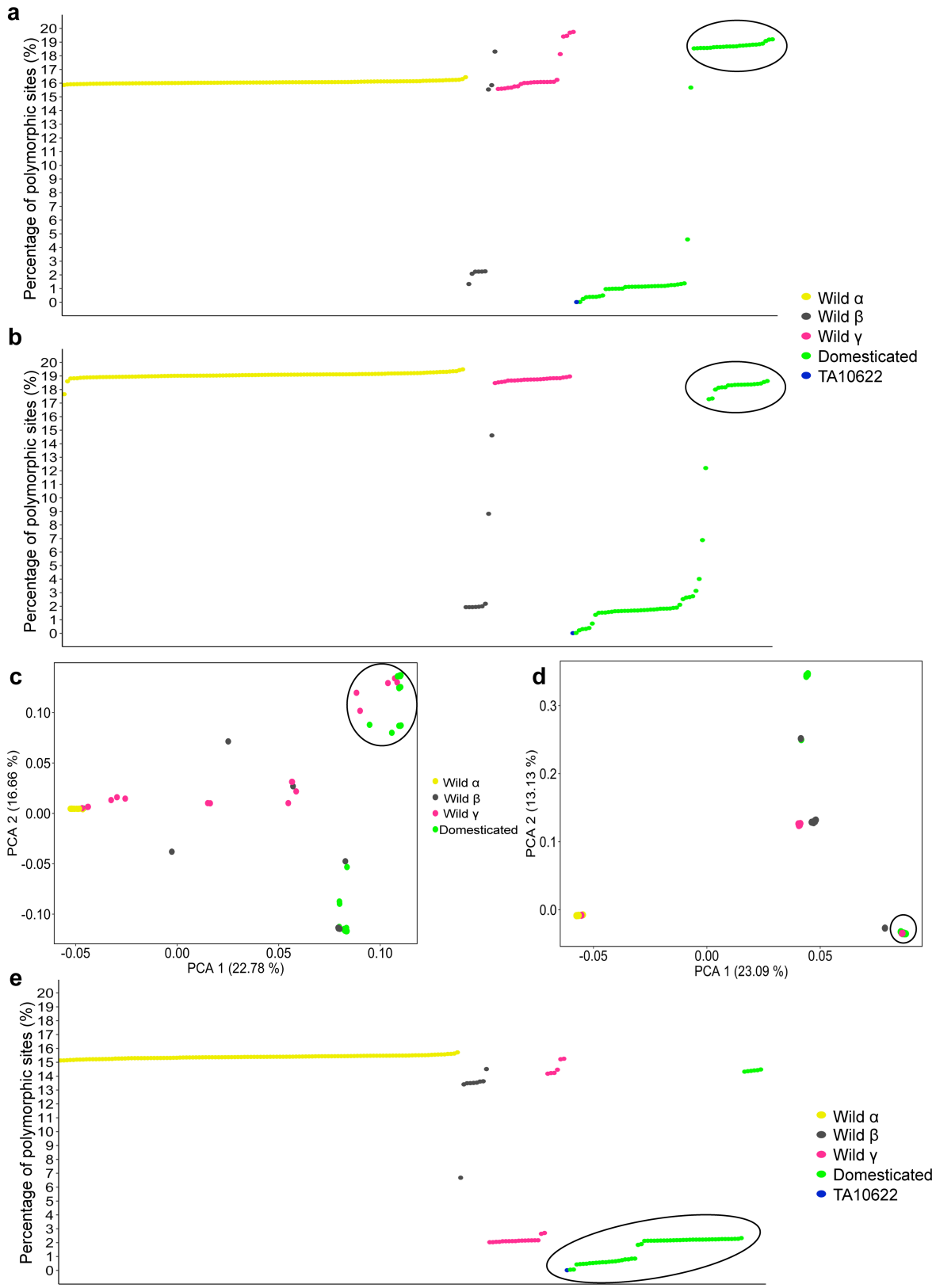
Extended Data Fig. 4 | Comparison of *RLG_Cereba* insertion ages with physical distance from centromeres. Note that ~95% of *RLG_Cereba* elements younger than 1 million years are found inside functional centromeres.



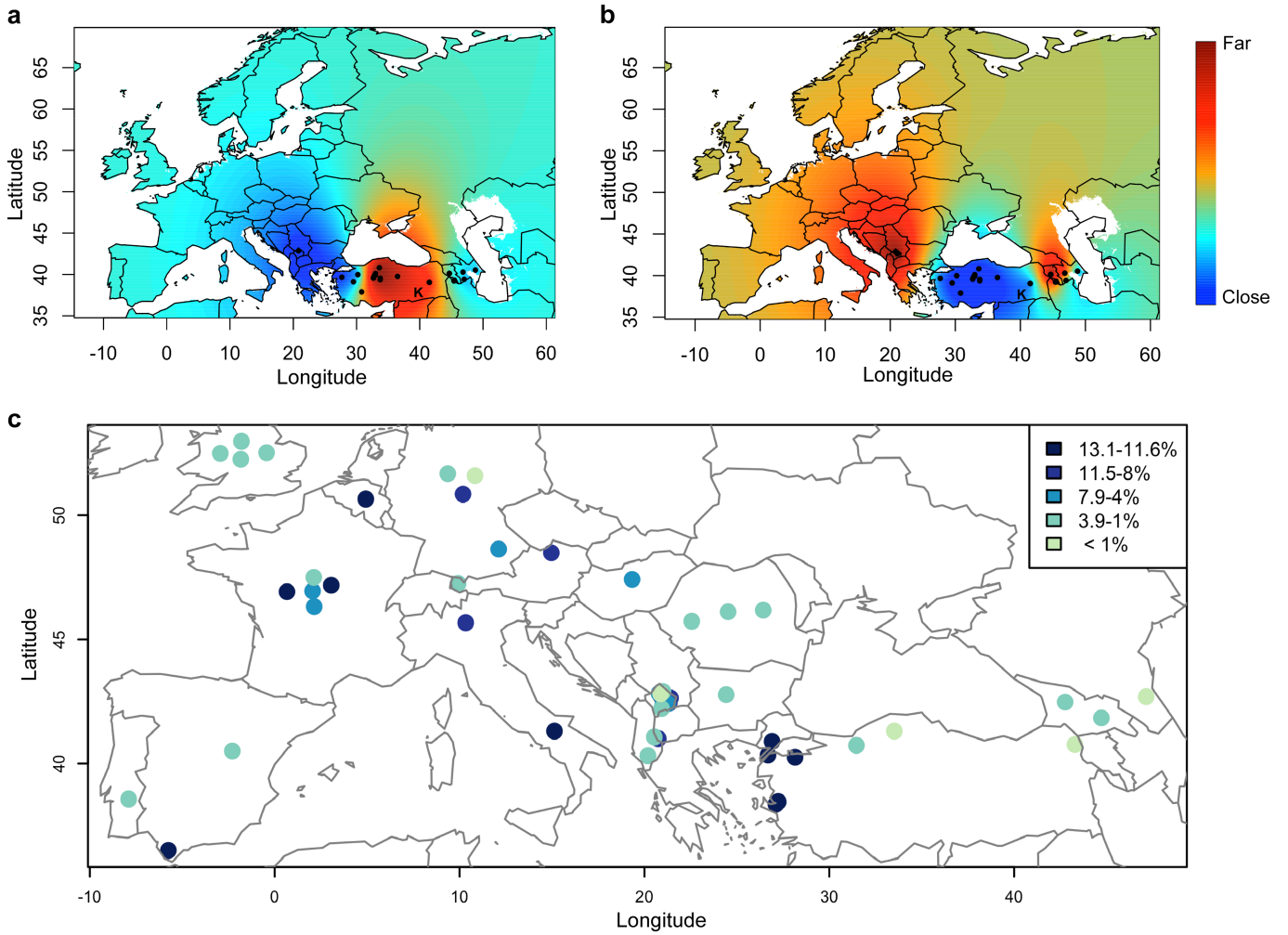
Extended Data Fig. 5 | CENH3 ChIP-Seq read coverage in relation to insertion ages of centromere-specific retrotransposons in *T. monococcum* accessions TA299 (left column) and TA10622 (right column). The top panels show the ratio of CENH3 ChIP-Seq reads divided by input control reads. ChIP-Seq and control reactions were performed in duplicates, which are shown in purple (Rep1) and orange (Rep2). The bottom panels show the chromosomal positions of *RLG_Cereba* (blue) and *RLG_Quinta* (red) retrotransposons (x-axis) and their insertion age (y-axis). The youngest retrotransposon insertions are generally found in the functional centromeres. Retrotransposon insertions on chromosomes 1A, 4A, 6A and 7A indicate that parts of the functional centromeres were moved at different evolutionary time points. The observed patterns can be explained by large-scale inversions that moved parts of centromeres (indicated by two-headed arrows, with y positions indicating the approximate time of the inversion event).



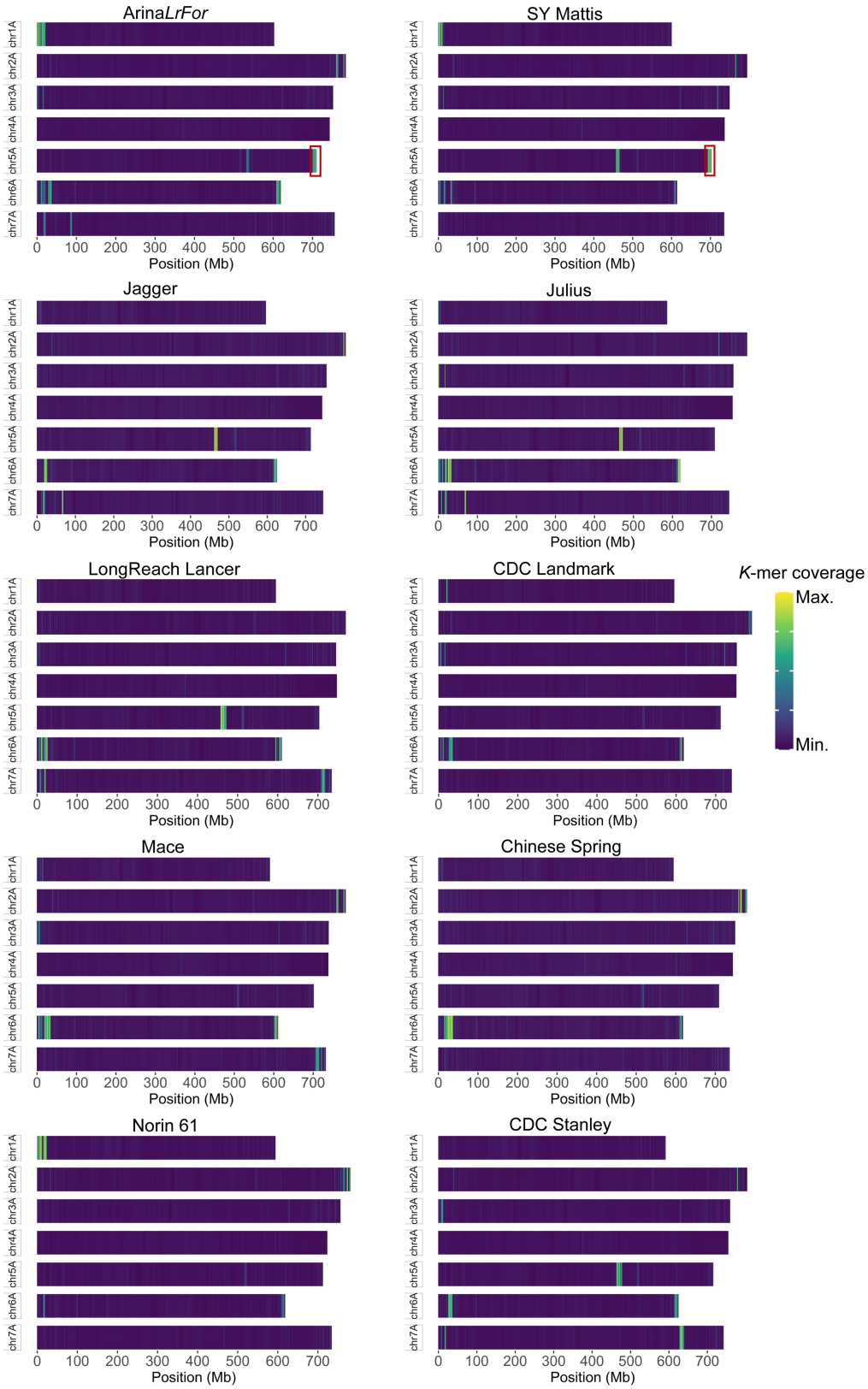
Extended Data Fig. 6 | Dot plot comparisons of centromeric and peri-centromeric regions of *T. monococcum* accessions TA299 (horizontal) and TA10622 (vertical). Aligned with the dot plots are plots of the average coverage of CENH3 ChIP-Seq reads and positions of *RLG_Cereba* retrotransposon insertions colour-coded according to their insertion ages. The plot for chromosome 4A is shown in the main Fig. 2.



Extended Data Fig. 7 | Population genomic analyses on diverged genomic regions. (**a**-**b**), The percentage of polymorphic sites of each einkorn accession compared to the TA10622 assembly considering SNPs present only in the two highly diverged genomic segments on chromosomes 2A (**a**) and 5A (**b**). The circles highlight domesticated einkorn accession that diverged from the TA10622 reference assembly. (**c**–**d**), Principal component analyses (PCA) based on SNPs found only in the two large introgressed segments on chromosomes 2A (**c**) and chromosome 7A (**d**) revealed that γ accessions cluster with some domesticated einkorn accessions. The circles highlight domesticated einkorn accessions that cluster with wild γ accessions instead of β. **e**, The percentage of polymorphic sites compared to the TA10622 assembly of the introgressed region on chromosome 7A showed that the majority of the domesticated einkorn accessions (highlighted in the circle) are not diverged from TA10622, which also carries the introgression on chromosome 7A.



Extended Data Fig. 8 | Wild einkorn γ race introgression into domesticated einkorn. a, Geographical projection of the second PCA axis based on variants found only in the large introgressed segment on chromosomes 5A. **b**, Geographical projection of the first PCA axis based on variants found only in the large introgressed segment on chromosomes 7A. Blue colour indicates γ accessions with close genetic relatedness to the introgressed segments found in domesticated einkorn accessions. Black dots represent the coordinates of γ accessions. The analysis was done excluding α and β accessions. Maps in both a and b were created using the Kriging function in the fields v10.3 R package (https://cran.r-project.org/web/packages/fields). **c**, The proportion of γ race introgression in domesticated einkorn. Each dot represents the coordinates of a domesticated einkorn accessions. Dark blue and light green represent the highest and the lowest proportions of γ introgressions in domesticated einkorn, respectively (legend at top right). The map was created using the graphic plot function in R.



Extended Data Fig. 9 | Einkorn introgression into 10 chromosome-scale bread wheat assemblies based on *k*-mer mapping approach. Putative introgressions are identified as regions with increased coverage of mapped *k*-mers from *T. monococcum* and visualized in the blue–yellow heat map (legend at the right). Red squares around chromosome 5AL in both Arina*LrFor* and SY Mattis indicate the *Yr34*-carrying region that was used as control.

Extended Data Table 1 | Summary of einkorn assemblies and annotations

	TA299	TA10622
Length of HiFi assembly (bp)	5,173,270,011	5,149,719,320
Number of contigs	1,240	1,142
Contig N50 (bp)	55,904,388	54,363,391
Contig N90 (bp)	12,937,174	13,108,674
Length of hybrid assembly (bp)[1]	5,174,460,332	5,154,588,727
Length of hybrid scaffolds (bp)	5,118,643,993	5,103,232,583
Number of hybrid scaffolds	22	24
Hybrid scaffold N50 (bp)	640,551,262	522,916,811
Length of pseudomolecule assembly (bp)[2]	5,174,572,132	5,154,690,404
Total length of anchored pseudomolecules (bp)	5,116,640,343	5,105,790,774
Number of anchored hybrid scaffolds or contigs	22	35
Number of gaps in anchored pseudomolecules	188	215
Total length of unanchored chromosome (bp)	57,782,004	48,795,835
Number of unanchored hybrid scaffolds or contigs	1,104	989
BUSCO scores		
Complete	97.6%	97.7%
Single	93.5%	93.5%
Duplicated	4.1%	4.2%
Fragmented	0.6%	0.6%
Missing	1.8%	1.7%
k-mer based completeness	98.9%	98.7%
Quality value (QV) score	49.28	49.63
Number of high-confidence genes on pseudomolecules	32,230	32,090

[1]Hybrid scaffolds were created by integrating optical maps into the contig-level assembly. The hybrid assembly includes all hybrid scaffolds and contigs that were not integrated into hybrid scaffolds.
[2]Includes contig-level assembly, optical map, and Omni-C data.

nature portfolio

Corresponding author(s): Vijay K. Tiwari, Michael Abrouk, Jesse Poland, Simon G. Krattinger

Last updated by author(s): 18 June 2023

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies 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

Whole genome sequencing of 218 einkorn accessions was done using Illumina NovaSeq platforms. Sequencing reads for genome assemblies were generated using PacBio circular consensus sequencing. Optical maps were generated using the Bionano Genomics Saphyr System. Omni-C reads were sequenced with Illumina NovaSeq platforms. RNA-Seq and Iso-Seq for six different tissues for two einkorn accessions were generated with Illumina NovaSeq platforms and PacBio circular consensus sequencing. CENH3 and H3K4me3 data were generated in this study.
The TREP database (v19) was used to obtain repeat information.
Translated proteins of Triticum urartu, Aegilops tauschii, wild emmer (Zavitan), hexaploid wheat (Kariega and ArinaLrFor), barley (Morex version 3), Brachypodium distachyon, rice, and the Triticeae and Poaceae protein sequences downloaded from the UniProt database (2021_03) were used for gene model prediction.

Data analysis

The software and tools used in this study are as follows:
Bionano Solve (v.3.6), hifiasm (v15.1), Juicer (v1.6), 3D-DNA (v180922), Juiecbox (v1.11.08), BUSCO (v5.0.0), Merqury (v1.3), SAMtools (v1.8), BCFtools (v1.9), JoinMap (v5.0), Mapchart (v2.32), STAR (v2.7.0f and v2.5.2a), Stringtie (v2.1.4), minimap2 (v2.21), cDNA_Cupcake (v12.4.0), Transdecoder (v5.5.0), BRAKER2 (v2.1.2), FgeneSH (v8.0.0), EDTA, GenomeThreader (v1.7.1), EVidenceModeler (v1.1.1), PASA pipeline (v2.5.1), DIAMOND (v2.0.9), MCScanX, Circos software (v0.69-9), Gepard, bowtie2, MUMmer (v4.0.0.2), Tandem Repeats Finder (v4.09.1), BLASTn (2.11.0+), trimmomatic (v0.38), BWA mem (v0.7.17), Picard tools, GATK (v4.1.8.0), VCFtools (v0.1.17), PLINK (v1.90), TreeMix (v1.13), jellyfish (v2.2.10), featureCounts (v2.0.0), ccsmeth (v0.3.2), InterproScan (vv5.55-88.0), deepTools, epic2 peak caller

R packages used in this study are as follows:
stats v4.1.3, LEA v2.0, fields v10.3, ggplot2 v3.3.6

Custom pipelines or scripts generated and used in this study:
Centromere comparison: (https://github.com/Wicker-Lab/Monococcum_genome_scripts)
IBSpy pipeline(Identity-by-State python; https://github.com/Uauy-Lab/IBSpy).
Wheat SNP calling scripts (https://github.com/IBEXCluster/Wheat-SNPCaller).
Codes and custom scripts for analyzing einkorn introgressions into bread wheat (https://github.com/Uauy-Lab/monococcum_introgressions).

Custom codes used to process data for genetic linkage maps are as follows:
Perl scripts for demultiplexing of raw FASTQ files (https://github.com/sandeshsth/SkimSeq_Method and https://github.com/sandeshsth/Fastq).
Codes and steps to generate RIL population: (https://datadryad.org/stash/share/v20dkVsSTj3toGn-CHG92eUSgre17uMT5AH_6LE2GDM).

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The raw sequence data of T. urartu was downloaded from the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) accession number PRJNA663409.
The ten bread wheat assemblies were downloaded from https://wheat.ipk-gatersleben.de/

Data availability
The raw sequencing data used for de novo genome assemblies, the RNA-Seq and Iso-Seq data for the annotation, the ChIP-Seq reads, and the whole-genome sequencing reads of 218 einkorn accessions are available on EBI_ENA under study number PRJEB61155. The two reference assemblies, annotations, VCF files, and CpG methylation frequencies are available on DRYAD [https://doi.org/10.5061/dryad.v41ns1rxj]. Raw fastq files and demultiplexed fastq files of the RIL population have been deposited at the National Center for Biotechnology Information (NCBI) SRA database with the BioProject accession PRJNA879879. The barcode indices key file with required information for demultiplexing can be obtained at DRYAD [https://doi.org/10.5061/dryad.v41ns1rxj]. CENH3 BED files (CENH3 peaks), and mapped files (bam) of CENH3 and H3K4me3 for all replicates are available through the Dryad database: [https://doi.org/10.5061/dryad.0p2ngf24b]. Whole genome sequencing data of the tin3 mutant bulk has been deposited into GenBank under BioProject PRJNA938447. The source data underlying Supplementary Figures 15, and 20, as well as Extended Data Fig. 9 are provided as a Source Data file. IBSpy variations tables of the 218 einkorn accessions, and MUMer alignments of the two einkorn assemblies against the 10 bread wheat assemblies are available through the following link: [https://opendata.earlham.ac.uk/wheat/under_license/toronto/Uauy_2022-09-24_IBSpy_Triticum_monococcum_introgressions/].
An interactive webpage has been developed for this study to visualize various genome characteristics of TA299 and TA10622. The webpage features a JBrowse 2 explorer allowing visualization of the whole genome, gene models, transposable elements, variants positions and synteny. For the identification of homologous sequences in other wheat varieties, a BLAST server has been set up. This BLAST server allows searches against individual wheat subgenomes and chromosomes independently. Synteny between TA10622 and other wheat genomes can be visualized in the synteny tab located on the webpage. The database can be accessed through the following link: https://avena.pw.usda.gov/genomes/mono

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	no human research participants
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

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	A recombinant inbred line (RIL) population consisting of 827 lines were used to construct the genetic maps. The larger the sample size is, the better in constructing the genetic map. The first set originally developed by Singh et al. (2007) [https://pubmed.ncbi.nlm.nih.gov/17565482/] consisted of 93 samples, including three blanks, and was sequenced at 0.2x coverage. The second set comprised 733 samples, sequenced at 0.03x coverage, with 35 blanks included as controls. In the skim-seq panel, we also included five replicates of each of the two RIL parents along with the RILs. The number of plants (n) for the tin3 analysis were as follows: Jagger (n = 20), tin3A (n=12), tin3B (n=6), tin3D (n=14), tin3AB (n=7), tin3AD (n=8), tin3BD (n=12), and tin3ABD (n=8). These samples were used for Tukey's HSD statistics. No statistical methods were used to establish sample sizes for genome assemblies and whole genome sequencing data for the einkorn diversity panel. Two einkorn accessions were selected for genome assemblies. A total of 218 of wild and domesticated einkorn accessions were used for the population genomics analyses. The chosen accessions cover a wide range of geographic and genetic diversity distribution of einkorn, which allowed to understand crop diversity and population structure in details. The einkorn accessions were selected from a larger diversity panel comprising 733 accessions and they were chosen based on genotyping-by-sequencing data (Reference: Adhikari et al., Genetic characterization and curation of diploid A-genome wheat species, Plant Physiology(2022)). A total of 6 different plant tissues per accession genome were used to extract RNA for RNA-Seq and Iso-Seq.
Data exclusions	218 out of 219 einkorn accessions were used for population genomics analysis. One accession was removed from the analysis due to mis-classification (i.e, this accession was not einkorn).
Replication	For validating the tandem duplication in the reference genome assembly TA10622, three different technical replicates were used for each primer combination. For the positional cloning of tin3 experiment: TA4342-L96 and tin3 mutants have been phenotyped four independent times with 15-20 plants per genotype each time. The SEM experiment was repeated three times. For ChIP-Seq experiment, two replicates have been used, all attempts were successful
Randomization	Randomization were not needed for this study as the study focuses on establishing and analyzing genomic resources and perform population genomics analyses. Our study does not include different treatment groups. Randomization is important to ensure that the allocation of participants or samples to different treatment groups is unbiased and free from systematic biases which in this case is not applicable to our study.
Blinding	Blinding does not apply to this study, as the main focus of our study is on the observational design, were we collect genetic data from individuals or populations without intervening or manipulating any variables. Our study does not invlove treatment or intervention being administered that would require blinding.

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
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	ChIP-Seq: Nuclei were isolated from 2-week-old seedlings and digested with micrococcal nuclease (Sigma) to liberate nucleosomes. The digested mixture was incubated overnight with 3 ug of antibody at 4oC. The target antibodies were captured from the mixture using Dynabeads Protein G (Invitrogen, Carlsbad, CA) to obtain ChIP DNA. Mock DNA control was maintained with the input DNA following the same conditions above without antibodies. The ChIP experiments were performed with two biological replications. Library construction was performed using the TruSeq ChIP Sample Prep Kit (Illumina, San Diego, CA) according to the manufacturer's instructions. Antigen with the peptide sequence 'RTKHPAVRKTKALPKK' corresponding to the N-terminus of wheat CENH3 was used to produce antibody utilizing the custom-antibody production facility provided by the Thermo Fisher Scientific, Illinois, USA (abs@thermofisher.com). A 0.396 mg of customized antibody was purified and obtained as pellet. The pellet was dissolved in 2 ml of PBS buffer, pH 7.4 resulting in 198 ng/ul of CENH3 antibody. Fifteen microliters of anti-CENH3 antibody was used for chromatin

	immunoprecipitation (ChIP). Anti-trimethyl-Histone H3 (Lys4) (H3K4me3) antibody (Cat.# 07-473) was purchased from Sigma (St. Louis, MO). A 3 ul of anti-H3K4me3 antibody was used for ChIP.
Validation	The specificity of anti-CENH3 and anti-H3K4me3 antibodies were validated using immunofluorescence assay on mitotic and meiotic chromosomes of diploid (Triticum monococcum) and hexaploid (Triticum aestivum) wheat.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	The raw sequencing data has been deposited to the EBI_ENA database under study number PRJEB61155 BED files (CENH3 peaks), and mapped files (bam) for both CENH3 and H3K4me3 are in Dryad database: https://doi.org/10.5061/dryad.0p2ngf24b.
Files in database submission	Samples names in EBI_ENA (project ID: PRJEB61155) are as follow: TA299-C1-CenH3_ChiP_replicat1, TA299-C2-CenH3_ChiP_replicat2, TA299-M1-CenH3_DNA_input_control_replicate1, TA299-M2-CenH3_DNA_input_control_replicate2, TA10622-C1-CenH3_ChiP_replicat1, TA10622-C2-CenH3_ChiP_replicat2, TA10622-M1-CenH3_DNA_input_control_replicate1, TA10622-M2-CenH3_DNA_input_control_replicate2, TA299-C1-H3K4me3_ChiP_replicat1, TA299-C2-H3K4me3_ChiP_replicat2, TA299-M1-H3K4me3_DNA_input_control_replicate1, TA299-M2-H3K4me3_DNA_input_control_replicate2. The Dryad link contains BED files, and mapped files (.bam)
Genome browser session (e.g. UCSC)	No longer applicable

Methodology

Replicates	The ChIP experiments were performed with two independent biological replicates. The ChIP-Seq profile map of the replicates were identical.
Sequencing depth	The CENH3 ChIP-Seq reads represent ~ 4x coverage per genome.
Antibodies	Wheat CENH3 antibody as described here: Koo DH, Sehgal SK, Friebe B, Gill BS (2015) Structure and stability of telocentric chromosomes in wheat. PLoS One 10: e0137747.
Peak calling parameters	The epic2 peak caller was used to identify peaks of CENH3 enrichment with a MAPQ≥30 filtering and with a resolution of 100 kb
Data quality	Quality filtering and adapter sequence removal were done with trimmomatic. Duplicates were removed using SAMtools. Secondary alignments (i.e. multi-mapping reads) were removed using the flag -F 0x0100 with SAMtools. Ratio of ChIP/input coverage was calculated using the deeptools function bamCompare using MAPQ≥30 as a threshold
Software	Trimmomatic, bowtie2, SAMtools, deeptools (function bamCompare), epic2 peak caller