CellPress
Partner Journal

Molecular Plant
Research Article

Genome sequences of five Sitopsis species of *Aegilops* and the origin of polyploid wheat B subgenome

Lin-Feng Li[1,2,5,*], Zhi-Bin Zhang[1,3,5], Zhen-Hui Wang[4], Ning Li[1], Yan Sha[1], Xin-Feng Wang[2], Ning Ding[2], Yang Li[1], Jing Zhao[1], Ying Wu[1], Lei Gong[1], Fabrizio Mafessoni[3], Avraham A. Levy[3,*] and Bao Liu[1,*]

[1]Key Laboratory of Molecular Epigenetics of the Ministry of Education (MOE), Northeast Normal University, Changchun 130024, China

[2]Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, School of Life Sciences, Fudan University, Shanghai 200438, China

[3]Department of Plant and Environmental Sciences, The Weizmann Institute of Science, 76100 Rehovot, Israel

[4]Faculty of Agronomy, Jilin Agricultural University, Changchun 130118, China

[5]These authors contributed equally to this article.

***Correspondence: Lin-Feng Li (lilinfeng@fudan.edu.cn), Avraham A. Levy (avi.levy@weizmann.ac.il), Bao Liu (baoliu@nenu.edu.cn)**

https://doi.org/10.1016/j.molp.2021.12.019

ABSTRACT

Common wheat (*Triticum aestivum*, BBAADD) is a major staple food crop worldwide. The diploid progenitors of the A and D subgenomes have been unequivocally identified; that of B, however, remains ambiguous and controversial but is suspected to be related to species of *Aegilops*, section Sitopsis. Here, we report the assembly of chromosome-level genome sequences of all five Sitopsis species, namely *Aegilops bicornis*, *Ae. longissima*, *Ae. searsii*, *Ae. sharonensis*, and *Ae. speltoides*, as well as the partial assembly of the *Amblyopyrum muticum* (synonym *Aegilops mutica*) genome for phylogenetic analysis. Our results reveal that the donor of the common wheat B subgenome is a distinct, and most probably extinct, diploid species that diverged from an ancestral progenitor of the B lineage to which the still extant *Ae. speltoides* and *Am. muticum* belong. In addition, we identified interspecific genetic introgressions throughout the evolution of the *Triticum*/*Aegilops* species complex. The five Sitopsis species have various assembled genome sizes (4.11–5.89 Gb) with high proportions of repetitive sequences (85.99%–89.81%); nonetheless, they retain high collinearity with other genomes or subgenomes of species in the *Triticum*/*Aegilops* complex. Differences in genome size were primarily due to independent post-speciation amplification of transposons. We also identified a set of Sitopsis genes pertinent to important agronomic traits that can be harnessed for wheat breeding. These newly assembled genome resources provide a new roadmap for evolutionary and genetic studies of the *Triticum*/*Aegilops* complex, as well as for wheat improvement.

Key words: *Aegilops*, Sitopsis, genetic introgression, genome evolution, polyploid wheat, *Triticum*

Li L.-F., Zhang Z.-B., Wang Z.-H., Li N., Sha Y., Wang X.-F., Ding N., Li Y., Zhao J., Wu Y., Gong L., Mafessoni F., Levy A.A., and Liu B. (2022). Genome sequences of five Sitopsis species of *Aegilops* and the origin of polyploid wheat B subgenome. Mol. Plant. **15**, 488–503.

INTRODUCTION

The genus *Triticum*, which is of paramount agricultural importance, contains several domesticated wheats at different ploidy levels. The hexaploid common or bread wheat (*Triticum aestivum*, 2*n* = 6*x* = 42, BBAADD) is the most widely grown and largest acreage crop in the world, providing about 20% of the global calories and protein in the human diet (Shewry and Hey, 2015). Common wheat contains three closely related subgenomes (A, B, and D) donated by distinct diploid species, which were reunited via a recent allohexaploid speciation event between cultivated forms of *Triticum turgidum* (BBAA) and *Aegilops tauschii* (DD) fewer than 10 000 years ago (Feldman et al., 1995; Nesbitt and Samuel, 1996). The diversification of the *T. turgidum* lineage was initiated after the domestication of cultivated emmer wheat (ssp. *dicoccon*) from wild emmer wheat (ssp. *dicoccoides*) (Nesbitt and Samuel, 1996; Feldman and Kislev, 2007; Matsuoka, 2011). Cultivated emmer wheat has undergone considerable varietal

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

diversification in the process of domestication, leading to the establishment of durum (*T. turgidum* ssp. *durum*) and several other minor tetraploid wheats (Feldman and Kislev, 2007; Luo et al., 2007; Matsuoka, 2011). Wild emmer wheat itself was formed via an earlier allotetraploidization event (<0.8 million years ago, mya) between two wild diploid species of the *Triticum*/*Aegilops* complex, which donated the A and B subgenomes, respectively (Gornicki et al., 2014; Marcussen et al., 2014). The second hexaploid wheat, *Triticum zhukovskyi* ($GGA^tA^tA^mA^m$), has evolved through the hybridization of cultivated tetraploid Timopheevi (*Triticum timopheevii*, GGA^tA^t) and diploid einkorn (*Triticum monococcum* ssp. *monococcum*, A^mA^m) wheat (Gill and Friebe, 2002). Cultivated Timopheevi wheat was domesticated from its wild relative *Triticum araraticum* (GGA^tA^t), which is believed to have arisen via allotetraploidization between *Aegilops speltoides* (SS) and *Triticum urartu* (AA) <0.4 mya (Tsunewaki et al., 1996; Gornicki et al., 2014). It has been established that the A and D subgenomes of common wheat are derived from the wild diploid wheat *T. urartu* (AA) and the goatgrass *Ae. tauschii* (DD), respectively (Kihara, 1944; Dvořák, 1976). Likewise, the three extant wild diploid *Triticum*/*Aegilops* species, *T. urartu*, *Ae. speltoides*, and *T. monococcum*, have been proposed as the respective donors of the three subgenomes (A^t, G, and A^m) of *T. zhukovskyi* (Wagenaar, 1966; Dvořák et al., 1993; Matsuoka, 2011). Although recent molecular marker-based phylogenetic analyses also suggested that the B subgenome of common wheat had originated from some variant accessions or cryptic subspecies within *Ae. speltoides* (Kilian et al., 2007), this conclusion was incongruent with other aspects of previous studies, rendering the origin of the common wheat B subgenome still unclear and controversial.

The hypothesis that the polyploid wheat B subgenome originated from a diploid *Aegilops* species of the section Sitopsis was proposed in the 1950s. This inference was mainly based on the close morphological similarity of the spikelet and karyotype structure between the polyploid wheats and the five Sitopsis species (Sarkar and Stebbins, 1956; Riley et al., 1958). Yet, comparisons of chromosome structure and meiotic pairing behavior revealed the almost complete absence of homologous synapsis between the polyploid wheat B subgenome and all Sitopsis species genomes (Kimber and Athwal, 1972; Gill and Kimber, 1974; Ruban and Badaeva, 2018). Molecular phylogeny and population genetic inferences showed either high genetic similarity (Miki et al., 2019) or the closest phylogenetic relationship (Huang et al., 2002; Petersen et al., 2006; Kilian et al., 2007; Gornicki et al., 2014) of *Ae. speltoides* (rather than the other Sitopsis species) to the wheat B subgenome. Molecular and cytological evidence has led to the monophyletic origin hypothesis purporting that the wheat B subgenome evolved from *Ae. speltoides* or a closely related species but was modified at the polyploid level (Sarkar and Stebbins, 1956; Salse et al., 2008). An alternative hypothesis holds that the origin of the modern wheat B subgenome or its diploid progenitor was polyphyletic, shaped by hybridization or introgression of diverse genomic sequences from different *Triticum*/*Aegilops* species (Zohary and Feldman, 1962; Natarajan and Sarma, 1974; El Baidouri et al., 2017). This scenario appeared to be congruent with inferences from transcriptome-based phylogeny and exome-based population genomics that revealed frequent interspecific hybridizations in the species complex (Glémin et al., 2019; He et al., 2019). However, genome-scale evidence supporting or refuting the monophyletic or polyphyletic origin of the polyploid wheat B subgenome is lacking.

The reference genomes of both hexaploid common/semi-wild wheats and tetraploid wild emmer/cultivated durum wheats, as well as their diploid progenitors *Ae. tauschii* and *T. urartu*, have been released in recent years (Appels et al., 2018; Avni et al., 2017; Guo et al., 2020; Ling et al., 2018; Luo et al., 2017; Maccaferri et al., 2019; Walkowiak et al., 2020), but a high-quality whole-genome sequence of the diploid species related to the B subgenome of polyploid wheat is not yet available. Here, we report chromosome-level genome assemblies of all five diploid species of *Aegilops*, section Sitopsis, i.e., *Ae. bicornis*, *Ae. longissima*, *Ae. searsii*, *Ae. sharonensis*, and *Ae. speltoides*, as well as a partial assembly of the *Amblyopyrum muticum* genome. The reference-quality genome assemblies of these diploid species, together with those available for polyploid wheat and the A and D subgenome diploid progenitor species, provide a comprehensive repository of genome resources for deeper evolutionary studies of the *Triticum*/*Aegilops* species complex. Our results also shed new light on the evolution of the B lineage, confirming that *Am. muticum* is the present-day species that diverged earliest from the most common recent ancestor of the B lineage (Glémin et al., 2019) and supporting the notion that the donor of the polyploid wheat B subgenome was a single, distinct, and most probably extinct diploid species most closely related to the still extant *Ae. speltoides*. Likewise, we also show that *Ae. speltoides* is not the direct progenitor of the G subgenome of the tetraploid wheat *T. timopheevii*. In addition, we show that the novel genomic resources of the members of the Sitopsis section can be mined to identify new genes and natural allele variants that can be utilized to cope with the ever-increasing global demand for wheat improvement.

RESULTS

Sequence assemblies and genome features

The identities of the five diploid Sitopsis species ($2n = 2x = 14$) of *Aegilops* were confirmed by spike morphology and fluorescence *in situ* hybridization (Supplemental Figure 1). The same multi-generation-selfed individual (bagged) of each species was used for genome sequencing and assembly. Sizes of the assembled genomes of the five species ranged from 4.11 to 5.89 Gb, broadly consistent with the values (4.60–6.22 Gb) estimated by flow cytometry (Table 1). Notably, among the five species, *Ae. speltoides* (4.11 Gb) has the smallest genome, which is close in size to those of the common wheat D subgenome (3.95 Gb) (Appels et al., 2018) and its donor *Ae. tauschii* (4.30 Gb) (Luo et al., 2017). By contrast, the remaining four Sitopsis species, *Ae. bicornis* (5.64 Gb), *Ae. longissima* (5.80 Gb), *Ae. searsii* (5.34 Gb), and *Ae. sharonensis* (5.89 Gb), all have much larger genomes similar to the polyploid wheat A (4.86–4.94 Gb) and B subgenomes (5.11–5.18 Gb) (Appels et al., 2018; Avni et al., 2017; Maccaferri et al., 2019) and to the genome of *T. urartu* (4.94 Gb) (Ling et al., 2018).

To determine the origin of the variable genome content, we annotated both protein-coding genes and repetitive sequences of the five Sitopsis species and compared them with genomes/subgenomes of the other relevant wheat species,

	Assembly parameter	*Ae. bicornis*	*Ae. longissima*	*Ae. searsii*	*Ae. sharonensis*	*Ae. speltoides*
Genome assembly	Genome size[a] (Gb)	5.73	6.22	5.55	6.07	4.60
	Total length of contigs (Gb)	5.64	5.80	5.34	5.89	4.11
	GC content (%)	46.42	46.40	46.01	46.24	46.34
	N50 length (contig) (Mb)	9.16	1.05	0.56	1.01	1.78
Repetitive sequence	Retrotransposons (Gb)	3.82	4.15	3.55	4.21	2.94
	DNA transposons (Gb)	0.94	0.89	1.03	0.91	0.52
	Total (Gb)	4.86	5.11	4.64	5.19	3.54
Protein-coding genes	Predicted protein-coding genes	61 354	63 326	62 804	61 849	61 084
	High-confidence genes	40 222	37 201	37 995	38 440	37 607
	Average transcript length (bp)	1484	1657	1554	1627	1622
	Average coding sequence length (bp)	1193	1293	1290	1289	1319
	Average exon length (bp)	325	358	329	351	348
	Average intron length (bp)	507	527	860	818	834
	Functionally annotated	37 082	36 539	37 489	37 798	37 301
	BUSCO integrity (%)	97.99	94.03	92.92	93.19	93.75

Table 1. Statistics of genome features of the five Sitopsis species of *Aegilops*.
[a]Genome size was estimated by flow cytometry.

including *Ae. tauschii* and *T. urartu*, wild emmer, domesticated durum, and common wheat (cv. Chinese Spring). Our gene annotation predicted from 37 201 to 40 222 high-confidence protein-coding genes in the five Sitopsis genome assemblies, with >92.2% being functionally annotated in the GO/KEGG/KOG/NR databases (Table 1). The average transcript lengths in the Sitopsis genomes are 1193–1319 bp, comparable to those of the three common wheat subgenomes (1310–1351 bp) and *Ae. tauschii* (1144 bp) but longer than that of *T. urartu* (998 bp)[23–25]. In addition, our results show that differences in genome size among the five Sitopsis species are mainly attributable to the variable total length of repetitive sequences (3.54–5.19 Gb, 86.13%–88.11% of the total), including 2.94–4.21 Gb (66.48%–71.47%) of retrotransposons and 0.52–1.03 Gb (12.54%–19.21%) of DNA transposons (Table 1).

Distribution patterns of GC content, protein-coding genes, and repetitive sequences were assessed for the five Sitopsis species and the common wheat B subgenome. Broadly consistent with previously published genomes of wheat species, all five Sitopsis species show higher gene density and lower GC content in distal compared with proximal chromosomal regions (Figures 1A–1C). A general genomic feature of the repetitive sequences of the five species and the common wheat B subgenome is that *copia*-like long terminal repeat (LTR) retrotransposons tend to cluster at telomeric regions of all seven chromosomes (Figure 1D), whereas a reverse distribution pattern is observed for *gypsy*-like LTR retrotransposons (Figure 1E). It is notable that *Ae. speltoides* shows a distinct distribution density of *copia*-like retrotransposons and *CACTA* DNA transposons compared with the common wheat B subgenome and the other four Sitopsis species across all seven chromosomes (Figures 1D and 1F). Nevertheless, estimates of the overall unique *k*-mer frequency reveal similar densities of repetitive sequence between the five Sitopsis species and the wheat B subgenome (Figure 1G), and these are far lower than those of the A and D subgenomes detailed in a previous study (Wicker et al., 2018) (Kruskal-Wallis test, $p < 0.001$). We also performed genome collinearity analyses to assess differences in genome structure. Although these *Triticum*/*Aegilops* species contain large proportions of repetitive sequences and differ substantially in genome size, they still retain highly collinear genomes (Supplemental Figure 2). Notably, two previously identified species-specific translocation events, 4A/5A/7B in tetraploid/hexaploid wheat (Dvorak et al., 2018) and $7S^{l}/4S^{l}$ translocation in *Ae. longissima* (Ankori and Zohary, 1962; Ruban and Badaeva, 2018; Chen et al., 2020), were confirmed in our genome collinearity analyses, corroborating the quality of our genome assemblies.

Molecular phylogeny, divergence time, and genetic similarity

Phylogenetic relationships of the five Sitopsis and other *Triticum*/*Aegilops* species were reconstructed based on single-copy orthologous gene, reduced representative genomic region (RRGR), and whole-genome single-nucleotide polymorphism (SNP) datasets. In line with previously inferred phylogenies (Marcussen et al., 2014; Glémin et al., 2019), the diploid species and polyploid wheat subgenomes fall into three independent clades corresponding to the A, B, and D lineages. *Ae. speltoides* is clustered with the polyploid wheat B subgenome (B lineage), and the remaining four Sitopsis species are grouped with the common wheat D subgenome (D lineage) and its diploid donor, *Ae. tauschii* (D lineage) (Figure 2A and Supplemental Figure 3).

Next, we estimated the time at which the *Triticum*/*Aegilops* species diverged from each other based on the same datasets. A genome-wide average was calculated: overall, *Ae. speltoides* diverged from the wheat B-subgenome donor ca. 4.44 mya

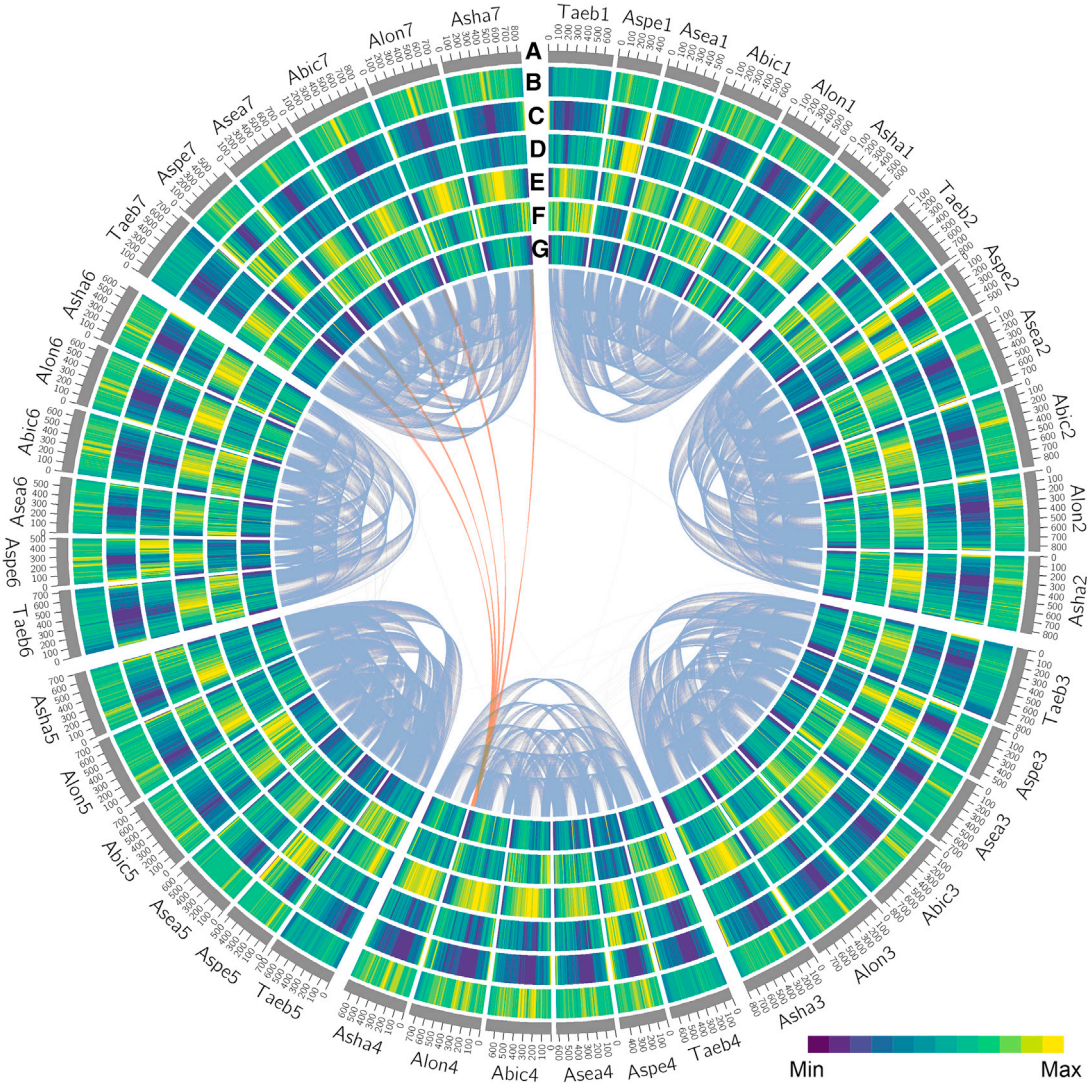


Figure 1. Structural, functional, and syntenic landscape of the five Sitopsis species and the common wheat B subgenome.
The lanes of the circular diagram from outside to inside are **(A)** species and chromosome names, with tick marks placed in 100-Mb intervals; (**B**) percentage of GC content (min 42.5% and max 48.4%); (**C**) density of high-confidence genes (count/Mb; min 1 and max 21); (**D**) *copia*-like retrotransposon density (min 8.3% and max 27.6%); (**E**) *gypsy*-like retrotransposon density (min 14.9% and max 63.7%); (**F**) *CACTA* DNA transposon density (min 5.0% and max 21.4%); and (**G**) distribution of unique 20-mer frequencies across physical chromosomes (count/Mb; min 4 and max 144). The color of the links is blue between homologous chromosomes and orange in cases of large translocations. Abic, *Ae. bicornis*; Alon, *Ae. longissima*; Asea, *Ae. searsii*; Asha, *Ae. sharonensis*; Aspe, *Ae. speltoides*; Taeb, B subgenome of common wheat (Chinese Spring cultivar, IWGSC reference sequence v1.0).

(95% highest posterior density [HPD] 4.26–4.62 mya) (Figure 2A). By contrast, the common wheat A and D subgenomes diverged from their respective diploid progenitors at much later times: A subgenome versus *T. urartu*, 1.15 mya, 95% HPD 1.10–1.19 mya; D subgenome versus *Ae. tauschii*, <0.68 mya. Given that wild emmer wheat formed no earlier, and probably much later, than 0.80 mya (Marcussen et al., 2014), our results rule out the possibility that *Ae. speltoides* is the direct donor of the polyploid wheat B subgenome. Of the five D-lineage species, our estimates suggest that *Ae. tauschii* evolved independently around 5.23 mya (95% HPD 5.02–5.44 mya), probably soon after the homoploid hybridization event(s) between the ancient A and B lineages (~5.50 mya) (Marcussen et al., 2014). The four Sitopsis species, *Ae. bicornis*, *Ae. longissima*, *Ae. searsii*, and *Ae. sharonensis*, diversified more recently from a common ancestor, <3.61 mya (95% HPD 3.47–3.76 mya). In parallel, we also calculated the divergence times between the seven diploid species and the wheat B subgenome along each of the seven chromosomes (Supplemental Figure 4). We found that *Ae. speltoides* showed later divergence time in centromeric regions than in telomeric regions from the B subgenome. Likewise, the five modern D-lineage species (*Ae. bicornis*, *Ae. longissima*, *Ae. searsii*, *Ae. sharonensis*, and *Ae*. *tauschii*) also showed variable divergence times along the chromosome lengths. In particular, the four D-lineage Sitopsis species (*Ae. bicornis*, *Ae. longissima*, *Ae. searsii*, and *Ae. sharonensis*) possessed apparently later divergence times from the wheat B subgenome than did *Ae. speltoides* in several subtelomeric regions (i.e., chromosomes 2 and 7). Nonetheless, all seven diploid species showed earlier divergence times from the wheat B subgenome (>1.00 mya)

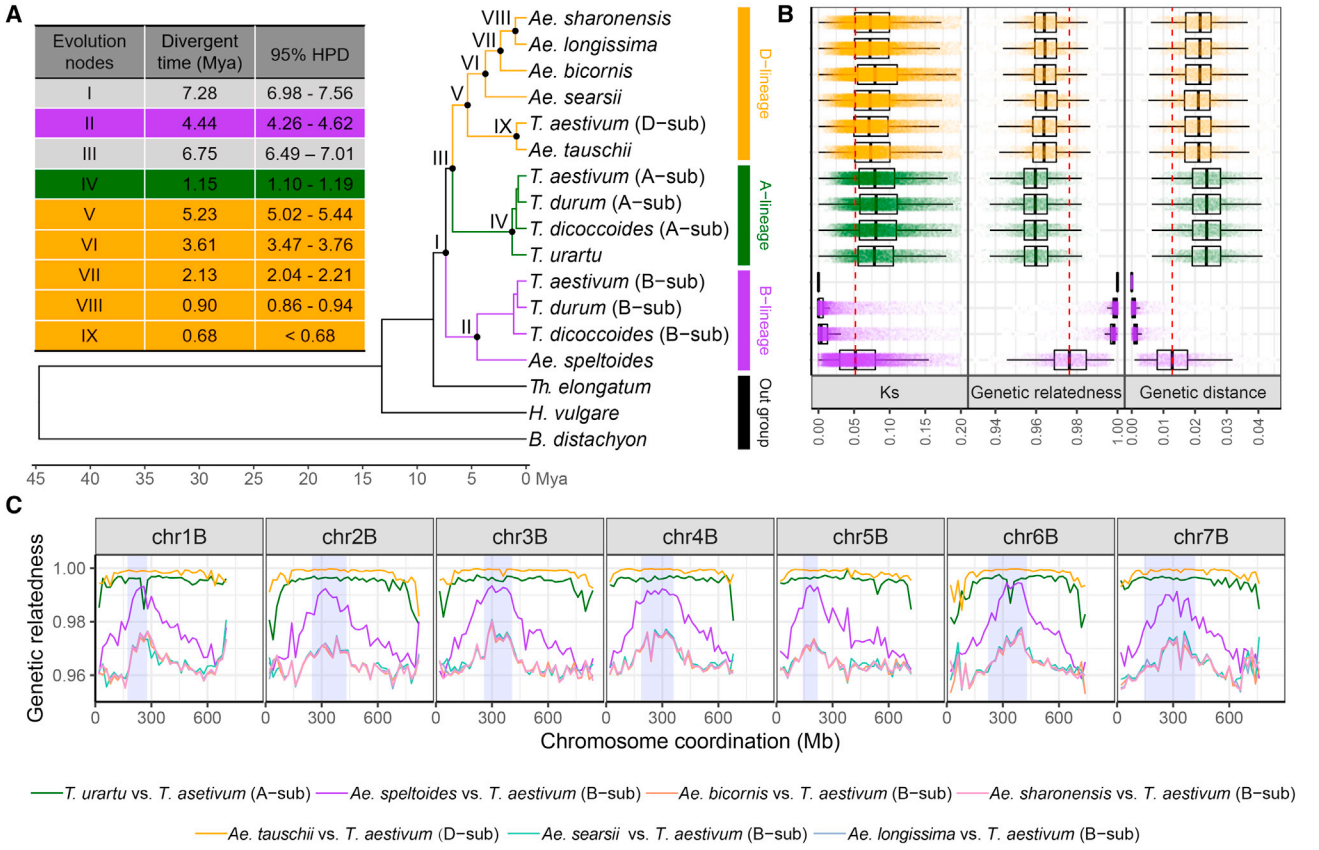


Figure 2. Phylogenetic relationships, divergence times, and genetic similarities of the *Triticum/Aegilops* species complex.

(A) Maximum likelihood tree and divergence times of the *Triticum/Aegilops* species complex based on single-copy orthologous genes. The green, purple, and orange branches represent the A, B, and D lineages, respectively. Black branches represent the common ancestor of *Triticum/Aegilops* species and outgroup species. The column "95% HPD" contains the 95% highest posterior density of divergence time for each of these representative nodes.

(B) Synonymous mutation rate (Ks), genetic relatedness, and genetic distance between the common wheat B subgenome and the other diploid species and polyploid wheat subgenomes based on collinear genes and reduced representative genomic regions. Green, purple, and orange indicate comparisons obtained from the A, B, and D lineages. Numbers on the x axis are the values of genetic similarity. Red dashed line indicates the median value between *Ae. speltoides* and the common wheat B subgenome.

(C) Distribution of genetic relatedness between the five Sitopsis species and the common wheat B subgenome, as well as between the common wheat A and D subgenomes and their diploid donors (*T. urartu* and *Ae. tauschii*) along the seven chromosomes of the common wheat B subgenome (IWGSC reference sequence v1.0) based on reduced representative genomic regions. The centromere of each chromosome is highlighted in purple. Numbers on the y axis are the values of genetic relatedness. Coordinates of each chromosome are shown on the x axis. Note that four Sitopsis species (*Ae. bicornis*, *Ae. searsii*, *Ae. longissima*, and *Ae. sharonensis*) are almost equally divergent from the B subgenome of common wheat and therefore show similar/overlapping patterns.

than the speciation time of wild emmer wheat (<0.80 mya) across all seven chromosomes, indicating that none of the five Sitopsis species is a direct progenitor of the B subgenome of polyploid wheats.

Taking advantage of a recently assembled common wheat cultivar (LongReach Lancer) with more than half of its chromosome 2B (ca. 450 Mb of a total length of ~800 Mb) substituted by the *T. timopheevii* G subgenome (Walkowiak et al., 2020), together with our assembled sequence contigs of *Am. muticum* (B lineage) (Supplemental Dataset 1), we constructed a separate phylogeny based on 285 RRGRs within this chromosomal segment from all pertinent diploid *Triticum/Aegilops* species and polyploid wheat subgenomes (Supplemental Figure 5A). Overall, we found that the segment-based inferences were highly consistent with the whole-genome molecular phylogenies and divergence times inferred above (Figure 2A and Supplemental Figure 3) and in previous studies (Marcussen et al., 2014; Glémin et al., 2019). For example, *Am. muticum* belongs to the B lineage and diverged from *Ae. speltoides* and the B subgenome at a more ancient time (6.35 mya, 95% HPD 5.89–6.80 mya), confirming that *Am. muticum* can definitely be considered the extant representative most directly related to the B-lineage ancestor (Glémin et al., 2019). Notably, *Ae. speltoides* diverged from the *T. timopheevii* G subgenome ca. 2.93 mya (95% HPD 2.57–3.30 mya), i.e., after its divergence from the B-subgenome progenitor (ca. 4.44 mya). This makes the donor of the G subgenome substantially older than the estimated allotetraploidization time (<0.4 mya) leading to the speciation of *T. araraticum*, the wild progenitor of *T. timopheevii* (Gornicki et al., 2014). From this analysis, it is clear that *Ae. speltoides* is also not the direct donor of the G subgenome of *T. timopheevii*, although it is more closely related to the G subgenome than to the B subgenome. This is consistent with earlier reports showing that *Ae. speltoides* shares nearly

identical cytoplasmic genomes with *T. timopheevii* (donated by the G-subgenome progenitor) but not with *T. turgidum* and *T. aestivum* (donated by the B-subgenome progenitor) (Ogihara and Tsunewaki, 1988). Similar relationships were also seen in gel blotting patterns revealed by repeated nuclear sequence probes (Dvorak and Zhang, 1990).

To gain further insight into the genome-wide genetic similarities of these *Triticum/Aegilops* species, we calculated genetic relatedness, genetic distance, and synonymous (d_S) and nonsynonymous (d_N) substitution rates based on collinear genes, single-copy orthologous genes, and RRGRs. In accordance with the divergence times detailed above, the wheat B subgenome is highly divergent from all the extant diploid *Triticum/Aegilops* species and is most closely related to *Ae. speltoides* (Figure 2B and Supplemental Figure 6). It is notable that the polyploid wheat A and D subgenomes display high genetic similarity to their respective diploid donors, *T. urartu* and *Ae. tauschii*, across almost the entire length of each of the seven chromosomes (Figure 2C). However, *Ae. speltoides* shows higher genetic similarity to the wheat B subgenome in centromeric regions than in telomeric regions (Figure 2C), mirroring the pattern of divergence time detailed above (see Supplemental Figure 4). This grossly bipartite divergence pattern was also observed in comparisons of the three recently split B-lineage species/subgenomes (*Ae. speltoides* and the B and G subgenomes) (Supplemental Figure 5B). By contrast, *Am. muticum* (B lineage) and other Sitopsis species (D lineage) show relatively lower genetic similarities to the wheat B subgenome (Figure 2C and Supplemental Figure 5B). Together, these genomic features suggest that (i) the ancestral B lineage should have contained at least four distinct diploid species, namely *Ae. speltoides*, *Am. muticum*, and the progenitors of the hexaploid common wheat B subgenome and the Timopheevii wheat G subgenome; (ii) the B subgenome is of monophyletic origin, i.e., from a single distinct diploid species, now extinct or yet to be discovered, that is phylogenetically most close to the extant *Ae. speltoides*; and (iii) it is possible that the diploid progenitor of the B subgenome may have experienced genetic introgression with other *Aegilops* species before its hybridization with *T. urartu* leading to the establishment of polyploid wheat, as detailed in later sections.

Heterogeneous variation pattern and genetic introgression

It has been proposed that interspecific hybridization occurred frequently in many of the *Triticum/Aegilops* species at various evolutionary stages (El Baidouri et al., 2017; Glémin et al., 2019; Huynh et al., 2019; Bernhardt et al., 2020). We therefore investigated whether hybridization/introgression also occurred and, if so, to what extent it had shaped the genomes of the five Sitopsis species. We found that the phylogenetic relationship between the B-lineage *Ae. speltoides* and the polyploid wheat B subgenome varied in 260 (11.3%) of the 2318 representative genomic regions, especially at the recombination-active distal chromosome regions (Supplemental Figure 7). Likewise, the five D-lineage species (including the remaining four Sitopsis species) also showed distinct phylogenetic topologies, but in a higher ratio than the B lineage, namely in >26.2% of the total genomic regions (Supplemental Figure 7). The observed heterogeneous patterns along all seven chromosomes suggest the possibility of either incomplete lineage sorting or genetic introgression between the five Sitopsis species and their relatives. Yet, the B-lineage donor of the ancestral homoploid hybridization speciation remains controversial (Jiang et al., 2020), and both *Ae. speltoides* and *Am. muticum* have been proposed as the parental donor of the D lineage (Marcussen et al., 2014; Glémin et al., 2019; Huynh et al., 2019; Bernhardt et al., 2020). Our estimates based on the *D* statistic, *fd*, hybrid index (γ), and χ^2 goodness of fit test confirmed the homoploid hybridization origin of the D lineage between the ancestral A and B lineages (Figures 3A and 3B and Supplemental Figure 8).

Previous cytogenetic studies showed that karyotypes of the four Sitopsis species (D lineage) are more similar to that of *Ae. speltoides* (B lineage) than to that of *Ae. tauschii* (D lineage) (Kihara, 1954). This observation was also supported by transcriptome-based phylogenetic inference that genetic introgression probably occurred from *Ae. speltoides* to the common ancestor of D-lineage Sitopsis species after its separation from *Ae. tauschii* (Glémin et al., 2019). We propose a different scenario in which the introgression event was more likely to have occurred from the last common ancestor of *Ae. speltoides* and the diploid donor of the wheat B subgenome (earlier than 4.44 mya) to the four D-lineage Sitopsis species (Figure 3B). This scenario of ancestral genetic introgression is also confirmed by the distribution pattern of introgressed sites (*i* sites) (Supplemental Figure 9A). For example, all four D-lineage Sitopsis species possess moderate *i* sites with both *Ae. speltoides* and the common wheat B subgenome (1.44% of the total SNPs) (putatively derived from their last common ancestor). However, fewer *i* sites were identified between the four D-lineage Sitopsis species and each of the two B-lineage species (0.90% and 1.04%) and the putative A-lineage donor *T. urartu* (1.31%). By contrast, the same D-lineage species *Ae. tauschii* harbors similar proportions of *i* sites with *Ae. speltoides* (0.26%), the B subgenome (0.26%), and the most recent common ancestor of the two species (0.30%), all of which are markedly lower than the proportion of *i sites with the* A-lineage *T. urartu* (0.89%). In line with these observations, allele frequency-based inference of migration also confirmed the ancestral genetic introgression from the B to the D lineage (Supplemental Figure 10). In particular, we identified several genomic regions that show high genetic similarity between the D-lineage Sitopsis species and either *Ae. speltoides* or the common wheat B subgenome (Supplemental Figures 11 and 12), suggesting the possibility of genetic introgressions between the B and the D lineage. These features together may explain why the four D-lineage Sitopsis species have higher genetic similarity to the wheat B subgenome in some genomic regions than to their otherwise phylogenetically closer relative, *Ae. tauschii*. It is notable that previous studies have proposed some additional post-ancestral homoploid hybridization genetic introgressions among the A-, B-, and D-lineage species (El Baidouri et al., 2017; Glémin et al., 2019; Huynh et al., 2019; Bernhardt et al., 2020). Broadly consistent with these studies, our integrated analyses also identified genetic introgressions from the D to the B lineage, i.e., moderate genetic introgression from *Ae. tauschii* (D lineage) to *Am. muticum* (B lineage) (Supplemental Figure 10).

Among the four D-lineage Sitopsis species, Waines and Johnson (Waines and Johnson, 1972) proposed that *Ae. sharonensis* was

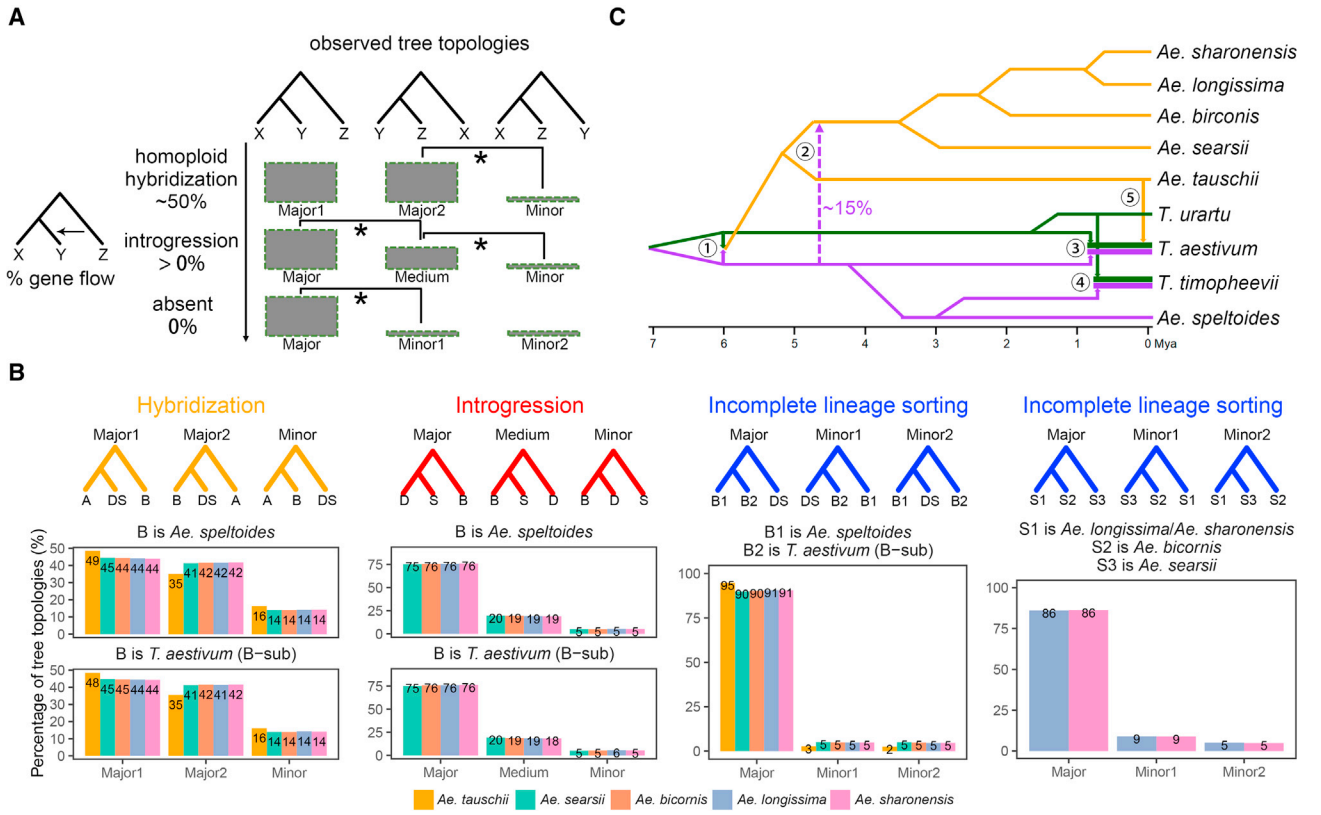


Figure 3. Estimates of genetic introgression in the *Triticum*/*Aegilops* species complex.

(A) Three typical incongruent types between the species tree and the gene tree based on reduced representative genomic regions. From top to bottom: homoploid hybridization, genetic introgression, and incomplete lineage sorting (ILS). The three incongruent types are distinguished by the numbers of each tree topology. Significance is determined by the χ^2 test with $^*p < 0.001$. In the homoploid hybridization type, numbers of both of the "major" topologies (major1 and major2, $2 > n_{major1}/n_{major1} > 1/2$) are significantly higher than that of the "minor" topology. In the genetic introgression type, the number of the major topology is significantly higher than that of the "medium" topology. Likewise, the medium topology is also significantly higher than the minor topology. In the ILS type, numbers of both of the minor topologies (minor1 and minor2) are less than that of the major topology, but this difference is not statistically significant.

(B) Examples of the three incongruent types identified in the *Triticum*/*Aegilops* species complex, including homoploid hybridization between ancestral A and B lineages (orange), genetic introgressions from the B lineage (ancestor of *Ae. speltoides* and the B subgenome) to the ancestor of the four D-lineage Sitopsis species (red), ILS between the B and D lineages (blue), and ILS among the four D-lineage Sitopsis species (blue). In the hybridization model, A, B, and DS represent the A-, B-, and D-lineage species (see Figure 2A), respectively. In the introgression model, S, D, and B indicate the D-lineage Sitopsis species, *Ae. tauschii*, and the B lineage (either polyploid wheat B subgenome or *Ae. speltoides*), respectively. In the ILS model, DS indicates the four D-lineage Sitopsis species and *Ae. tauschii*. Percentages of these trees are shown below the tree topology. Colors in the bar plot represent the five D-lineage species.

(C) Evolutionary scenario of the *Triticum*/*Aegilops* species complex based on the integrated estimates of genetic introgression. The numbers ① and ② indicate the ancestral homoploid hybridization between the A and B lineages and the B- to D-lineage ancestral genetic introgression event, respectively. The remaining three numbers (③, ④, and ⑤) represent the allopolyploidization events that formed tetraploid *T. turgidum* ssp. *dicoccoides*, tetraploid *T. araraticum*, and hexaploid *T. aestivum*. The A, B, and D lineages are marked as green, purple and orange, respectively.

likely to be a hybrid between *Ae. longissima* and *Ae. bicornis* based on morphology and cytogenetic analyses. Our estimates, however, did not find evidence of this possibility. The observed heterogeneous pattern was more likely due to the incomplete sorting of ancestral polymorphisms (Figure 3B). In line with this conclusion, we found that *Ae. sharonensis* not only possesses a high proportion of species-specific SNPs (8.80% of the total SNPs) but also shares a low proportion of species-shared SNPs with *Ae. bicornis* (1.27%) (Supplemental Figure 9B). It is notable that all the extant D-lineage species were probably established through a single ancestral homoploid hybridization event, as reported by Marcussen et al. (Marcussen et al., 2014) and modified by Glémin et al. (2019). We therefore asked how the above-identified genetic introgressions have shaped the genomes of the five extant D-lineage species (including the four Sitopsis species). By comparing the distribution patterns of A- and B-lineage-specific SNPs, we found that genomic regions containing more A-lineage-specific SNPs (A dominant) are clustered at the recombination-inert proximal regions across all seven chromosomes (Supplemental Figure 13). By contrast, species-specific SNPs identified in either *Ae. speltoides* (B lineage) or the common wheat B subgenome (B lineage) were distributed mainly at the recombination-active distal chromosomal regions. In particular, the Sitopsis species (excluding *Ae. speltoides*) harbor more B-lineage species-specific SNPs compared with *Ae. tauschii* (D lineage), confirming the above-identified B-lineage to D-lineage

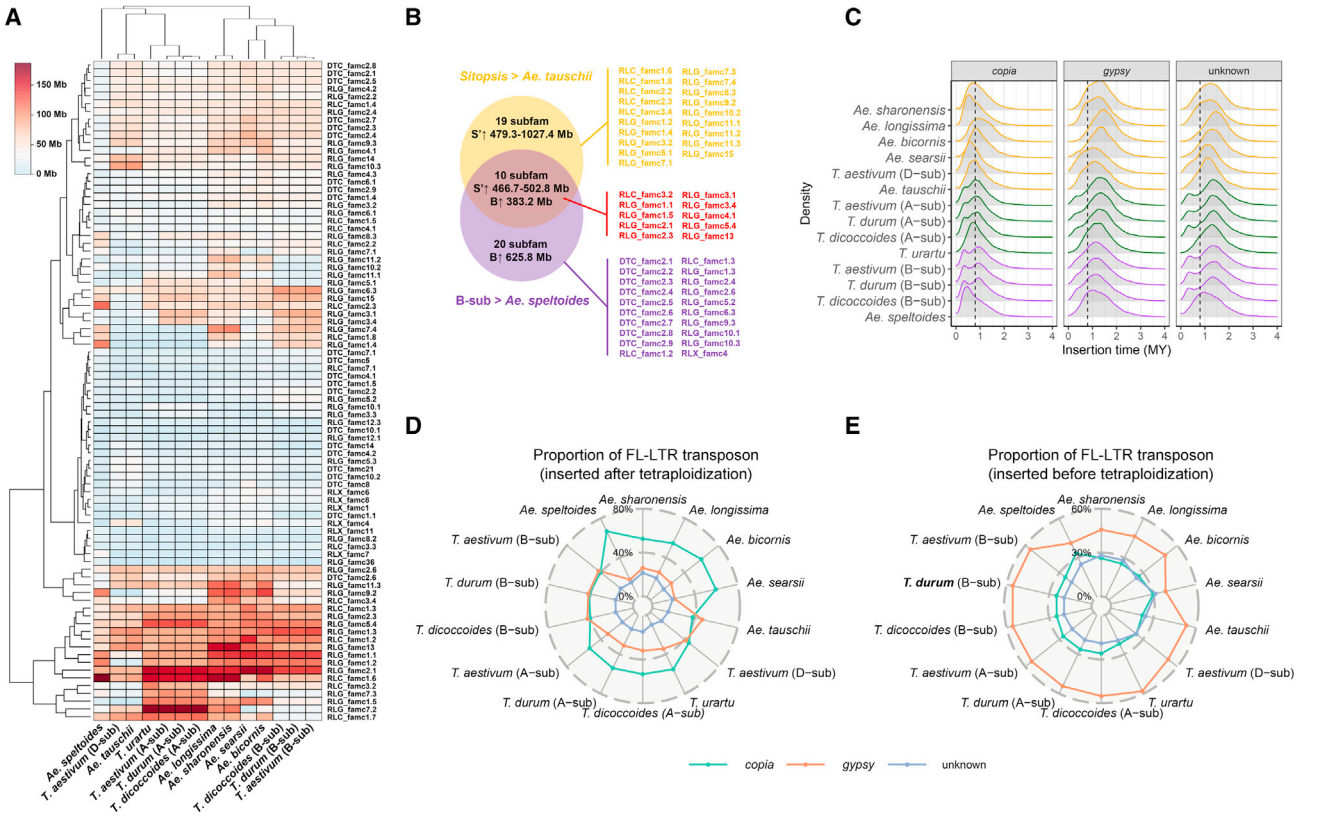


Figure 4. Evolutionary dynamics and insertion times of transposable elements in the *Triticum/Aegilops* species complex.

(A) Heatmap of the lengths of the 85 transposable element (TE) subfamilies in the diploid species and polyploid wheat subgenomes. Scale bar at the top left corner denotes the length of each TE subfamily.

(B) Intersection analysis of the TE subfamilies that are expanded specifically in the B (purple, 20 subfamilies) and D lineages (orange, 19 subfamilies), as well as those shared between the two lineages (red, 10 subfamilies). Total length of these expanded TE subfamilies is shown in each section. The characters S′ and B indicate the four D-lineage Sitopsis species and the common wheat B subgenome, respectively. Black arrows after S′ and B indicate the increased genome size in these genomes compared with their close relatives. The expanded TE subfamilies are listed on the right.

(C) Insertion times of full-length long terminal repeat (FL-LTR) retrotransposons in the diploid species and polyploid wheat subgenomes. Green, purple, and orange represent the A, B, and D lineages, respectively. The vertical line represents the inferred time of the tetraploidization speciation event (~0.8 mya).

(D and E) Proportions of the FL-LTR retrotransposons transposons in the diploid species and polyploid wheat subgenomes after **(D) and** before **(E)** the tetraploidization speciation event (~0.8 mya). Light green, orange, and blue indicate the *copia*-like, *gypsy*-like, and unknown retrotransposons, respectively.

introgression in the Sitopsis species. Together, our genome-scale estimates revealed frequent post-ancestral homoploid hybridization introgressions among the *Triticum/Aegilops* species (Figure 3C), which may have shaped the genomes of extant Sitopsis species.

Post-speciation amplification of transposable elements

The genomic features detailed above revealed differences in genome content between the five Sitopsis species and their close relatives (see Figure 1 and Table 1). We therefore assessed whether the differences in genome content are due to genetic introgressions (see Figure 3) or independent expansion/contraction of transposable elements (TEs). At the overall level, our analyses reveal that 2.73–3.95 Gb (77.6%–81.2%) of the genome components of these *Triticum/Aegilops* species are composed of the *gypsy*-like, *copia*-like, and *CACTA* TE families (Supplemental Figures 14A and 14B). About 0.97 Gb (90.4%) of the genome size difference between the polyploid wheat B subgenome and *Ae. speltoides* can be attributed to the high copy numbers of *gypsy*-like and *CACTA* TEs in the B subgenome (Supplemental Figure 14C). In the D lineage, compared with *Ae. tauschii*, all three types of TEs (*gypsy*-like, *copia*-like, and *CACTA*) show higher copy abundance in the two earlier established species, *Ae. bicornis* (1.12 Gb, accounting for 84.2% of the genome size difference) and *Ae. searsii* (1.00 Gb, 89.8% of the difference). By contrast, only *gypsy*-like and *copia*-like TEs exhibit high copy numbers in the two more recently established Sitopsis species, *Ae. longissima* (1.47 Gb, 93.6% of the difference) and *Ae. sharonensis* (1.54 Gb, 92.3% of the difference).

To examine whether specific repetitive sequences have contributed to the differences in genome content, we further characterized 85 retrotransposon and transposon subfamilies that are responsible for >90.0% of the genome size differences among the *Triticum/Aegilops* species (Figure 4A). In line with the above results, the differential abundance of two *gypsy*-like subfamilies (*RLG_famc3.1* and *RLG_famc3.4*) accounts for about 10.7% of the genome size differences between the polyploid wheat B subgenome and *Ae. speltoides* (Figure 4A). Likewise, different

copy numbers of 29 subfamilies are responsible for the different genome contents among the five D-lineage species. Intersection analysis showed that the 20 and 19 lineage-specific retrotransposon and transposon subfamilies contributed to 625.8 Mb (58.5% of B lineage) and 479.4–1027.4 Mb (43.0%–63.0% of D lineage) of the genome size differences in the B- and D-lineage species, respectively (Figure 4B and Supplemental Table 1). By contrast, seven subfamilies that were shared between the B and the D lineages account for 383.2 Mb (35.8% of B lineage) and 466.7–502.8 Mb (29.5%–42.0% of D lineage) of the differences in genome contents. This result suggests that genome size differences within the B and D lineages are primarily due to distinct proportions of specific TE subfamilies.

We next estimated the burst time of retrotransposons to reexamine whether they were expanded or contracted independently in the B and D lineages. If the retrotransposons expanded in the B and D lineages were directly derived from the above-identified interspecific genetic introgressions (detailed in Figure 3B), we would expect to identify a pre-introgression (>4.44 mya) burst of retrotransposons in the two lineages. However, our estimates identified relatively recent retrotransposon amplifications (<3.00 mya) in all the diploid species and polyploid wheat subgenomes (Figure 4C). It is notable that both the A and the B subgenome of the three polyploid wheats (emmer, durum, and common) have experienced a common recent retrotransposon expansion (~0.5 mya), most likely after the allotetraploidization event ca. 0.8 mya. Consistent with this inference, their diploid donors (*T. urartu* and *Ae. tauschii*) and the five Sitopsis species do not share this recent retrotransposon burst. We then compared the insertion times of these retrotransposon families relative to the allotetraploidization event (0.8 mya). In the B lineage, about 55.6%–55.7% of the earlier expanded retrotransposons (>0.8 mya) in the polyploid wheat B subgenome can be attributed to *gypsy*-like retrotransposons (Figures 4D and 4E and Supplemental Table 2). However, *Ae. speltoides* possesses more recently (<0.8 mya) amplified *copia*-like retrotransposons (67.0% of the total) compared with the polyploid wheat B subgenome (40.5%–41.2%). This may explain why *Ae. speltoides* shows a distinct distribution density of *copia*-like retrotransposons compared with the B subgenome and the other Sitopsis species (see Figure 1). In the D lineage, the *copia*-like families are responsible for 52.6%–59.8% of the recently amplified retrotransposons (<0.8 mya) in the four Sitopsis species (Figures 4D and 4E and Supplemental Table 2). By contrast, only 37.8%–45.1% of the recently expanded retrotransposons come from *copia*-like families in the polyploid wheat D subgenome and its donor *Ae. tauschii*, supporting the distinct evolutionary histories of the five modern D-lineage species observed above. In the A lineage, slightly lower proportions of recently expanded *copia*-like families were identified in the polyploid wheat A subgenome (31.2%–31.9%) compared with their diploid donor *T. urartu* (34.6%). We also estimated the insertion times for the expansion of the above-identified TE subfamilies in the seven diploid species and the wheat B subgenome. All these TE subfamilies possess insertion times <3.00 mya (Supplemental Figure 15), later than the ancestral B- to D-lineage introgression (4.49 mya) (detailed in Figure 2A). In particular, all four D-lineage Sitopsis species show distinct expansion patterns of these TE subfamilies compared with *Ae. speltoides* and the B subgenome, even for those that are expanded in both the B and the D lineages (Supplemental Figure 15). As these retrotransposon and transposon subfamilies are responsible for >90% of the genome size differences, this result suggests that the increased genome content in D-lineage Sitopsis species compared with *Ae. tauschii* is more likely to be due to post-speciation expansions/contractions of a few specific active TEs rather than to direct B- to D-lineage genetic introgression.

Pan-genomic analyses of the *Triticum*/*Aegilops* species

Pan-genomic analyses of the *Triticum*/*Aegilops* species were performed based on protein-coding genes and genome structural variations (SVs). In 49 384 gene families among the *Triticum*/*Aegilops* species, the five Sitopsis species contain 23 805–25 042 gene families, 12 197 (48.2%–51.2%) of which are shared among all species, probably representing the core gene set of the *Triticum*/*Aegilops* species complex (Figure 5A). In addition, a total of 20 086 (40.7%) dispensable and 17 101 (34.6%) species-specific gene families were also identified from these *Triticum*/*Aegilops* species. In these orthologous gene families, from 399 (1.00%) to 967 (2.40%) species-specific genes and from 1373 (4.50%) to 1788 (4.70%) specifically expanded genes were identified in the five Sitopsis species (Figure 5A). Functional analyses of the Sitopsis-specific and expanded genes revealed significant enrichment in basic cellular activities, including DNA recombination, DNA integration, and metabolic processes (Figure 5B and Supplemental Table 3). Based on the same protein-coding gene set, we characterized evolutionarily conserved genomic regions by identifying shared syntenic orthologous genes in the *Triticum*/*Aegilops* species. Our results revealed that centromeric regions of all seven chromosomes possess very few core putative protogenes (present in all diploid species and polyploid wheat subgenomes) (Supplemental Figure 16). This pattern can be explained by either the uneven distribution of protein-coding genes along the chromosomes or a low level of genetic conservation of centromeric regions in the *Triticum*/*Aegilops* species (Brinton et al., 2020; Hao et al., 2020). In addition, we identified several large genomic regions that are not evolutionarily conserved in the B and D lineages. For example, two large nonconserved genomic regions near the telomere of chromosome 2 are potentially correlated with the evolutionary divergence between the five Sitopsis species and their close relatives, the wheat B subgenome (B lineage), and *Ae. tauschii* (D lineage) (Supplemental Figures 16C and 16E).

Pan-genomic analyses were also performed with the genome-wide SVs characterized from the seven diploid *Triticum*/*Aegilops* species based on the corrected Nanopore long reads. We identified from 37 039 to 37 721 genomic SVs in the five Sitopsis species, 18 153 of which are shared by the two diploid species, *Ae. tauschii* (DD) and *T. urartu* (AA) (Supplemental Figure 17A). Of these shared SVs, 16 337 monomorphic SVs that are fixed in the seven diploid species compared with the polyploid wheat B subgenome were excluded, leaving 1816 polymorphic insertions/deletions in the genome-scale SV dataset. Further analyses of the 1816 polymorphic SVs showed that, although these SVs were scattered randomly along the seven chromosomes (Supplemental Figure 18), the seven diploid species possessed 5 to 120 species-specific SVs (Figure 5C and Supplemental Figures 17B and 17C). For example, although *Ae. speltoides* is

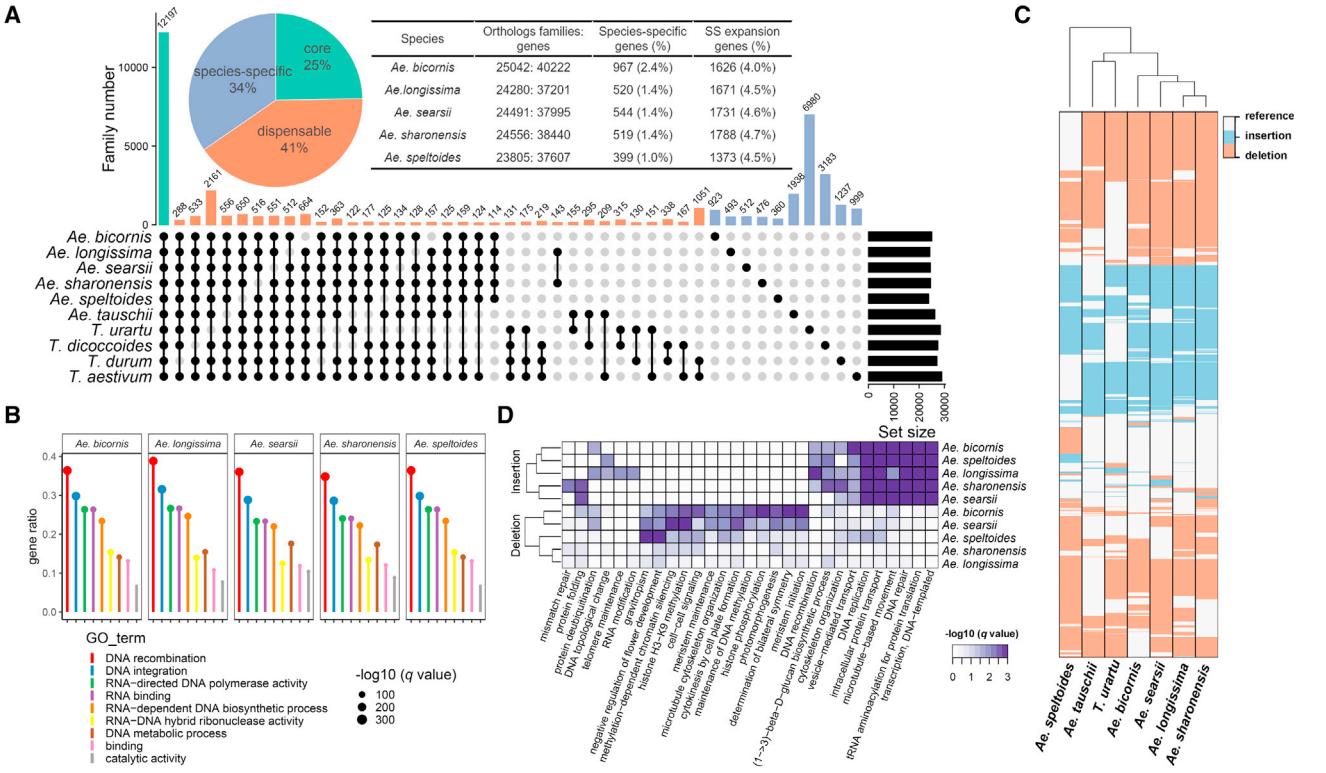


Figure 5. Pan-genomic analyses of the *Triticum/Aegilops* species complex based on protein-coding genes and genome structural variations.

(A) Intersection analysis of the protein-coding genes in the *Triticum/Aegilops* species. Light green, orange, and blue are the core, dispensable, and species-specific gene families in the *Triticum/Aegilops* species complex. SS expansion gene families (170, 0.3% of the total) are defined as expanded specifically in the five Sitopsis species compared with the rest of the *Triticum/Aegilops* species/subgenomes according to the phylANOVA test. Numbers in the top-corner table indicate the gene numbers of Sitopsis-specific and SS expansion family groups.

(B) Functional enrichment analysis of the SS expansion genes in the Sitopsis species.

(C) Heatmap of the distribution of genome structural variations in the seven diploid species compared with the common wheat B subgenome. Orange, blue, and white represent deletion, insertion, and identity relative to the common wheat B subgenome.

(D) Functional annotations of the structural variation-related genes in the five Sitopsis species. The full list of GO terms is shown in Supplemental Table 5.

phylogenetically close to the B subgenome, it still carries 120 (6.61% of the total polymorphic SVs) species-specific SVs compared with the B subgenome and all the other diploid species. Compared with the above Sitopsis-specific and expanded genes that are mainly involved in basic cellular activities, the SV-associated genes identified in the five Sitopsis species are correlated with several important phenotypes, such as photomorphogenesis and meristem and flower development, and genome functions, such as DNA methylation, chromatin silencing, and topology (Figure 5D and Supplemental Table 4).

The Sitopsis genomic resource

The *Aegilops* species represent secondary gene and germplasm resources for wheat genetic improvement (Feldman and Levy, 2015). We therefore examined whether the five Sitopsis species contain homoeologous genes related to agronomic and pathogen-resistant traits of durum and common wheat. Our genome-wide screening of the functional nucleotide-binding site and leucine-rich repeat (NBS-LRR) domains identified a total of 5859 genes in the 14 diploid species and polyploid wheat subgenomes. Further redundancy analysis revealed that the total NBS-LRR gene pool consists of 2573 (95% identity) to 4439 (100% identity) unique genes in the *Triticum/Aegilops* species, with 90% of the NBS-LRR genes contained in 12 (95% sequence identity) and 13 (100% sequence identity) wheat genomes, respectively (Supplemental Figure 19A). This result suggests that the 5859 NBS-LRR genes may represent the core resistance gene set of *Triticum/Aegilops* species. In particular, the 5859 NBS-LRR genes are mostly distributed at the distal chromosomal regions in all the *Triticum/Aegilops* species (Supplemental Figure 19B). This is broadly consistent with previous findings that gene families and quantitative trait loci associated with adaptation to biotic and abiotic stresses are mainly clustered near the subtelomeric chromosome regions in polyploid wheats (Appels et al., 2018; Maccaferri et al., 2019; Walkowiak et al., 2020).

We noted that the five Sitopsis species contain more NBS-LRR genes (388–490) than do the two other diploid species, *Ae. tauschii* (350) and *T. urartu* (318), and the two subgenomes of wild emmer wheat (239–286), but their NBS-LRR gene numbers are comparable to those of domesticated durum (326–435) and common wheat (401–542) (Supplemental Figure 20A). Intersection analysis of these NBS-LRR gene families allocated 112 (31.7%), 221 (62.6%), and 17 (5.7%) to core, dispensable, and Sitopsis-specific gene families (Supplemental Figure 20B). The

112 core and 221 dispensable NBS-LRR gene families are the major components of the innate immune system in the *Triticum*/*Aegilops* species. Thus, the 14 Sitopsis-specific NBS-LRR gene families may provide specific genetic resources for the improvement of disease resistance in domesticated durum and common wheat. In addition, we also characterized a set of homoeologous genes in the five Sitopsis species that are related to stripe rust (*Puccinia striiformis* f. sp. *tritici*, *Pst*) and powdery mildew (*Blumeria graminis* f. sp. *tritici*, *Bgt*) resistance (Supplemental Table 5). Because *Pst* and *Bgt* are two major fungal diseases that cause heavy yield losses in wheat production worldwide (Zhang et al., 2014), the novel resistance genes we identified in the Sitopsis species may be important genetic resources for future wheat breeding.

For agronomic traits, we checked the copy number and nucleotide variation pattern of two major domestication genes related to the nonshattering phenotype: free-threshing seed (*Q*/*q*) and nonfragile rachis (*Btr*/*btr*) (Supplemental Table 5). The *Q*/*q* gene encodes an *AP2*-like transcription factor that confers free threshing and also has pleiotropic effects on a number of other domestication traits, including rachis fragility, spike architecture, and flowering time (Faris and Gill, 2002; Simons et al., 2006; Zhang et al., 2011). The seven diploid *Triticum*/*Aegilops* species and their attendant natural and resynthesized polyploids with diverse genome combinations, including BBAA, $S^{sh}S^{sh}A^mA^m$, S^lS^lAA, S^bS^bDD, and AADD, all show substantial morphological differences in inflorescence structure (Zhang et al., 2013) and distinct *Q*/*q* allele expression patterns (Wang et al., 2016). Here we show that, although all five Sitopsis species harbor the wild-type *q* allele (L_{329}), it is different from that of the durum and common wheat A subgenome domesticated *Q* allele (I_{329}) and from the *q allele* ($V_{329)\ of}$ their diploid donor *T*. *urartu*; it also contains numerous unique synonymous and nonsynonymous mutations (Supplemental Figure 21 and Dataset 2). A similar phenomenon was observed in the two nonfragile rachis genes (*Btr1* and *Btr2*), many genetic variants of which are found in the five Sitopsis species (Supplemental Table 5 and Dataset 3). It has been documented that novel SNPs in the miRNA binding site at the *Q*/*q* gene are correlated with changes in transcriptional regulation and play pleiotropic roles in growth and reproductive development (Greenwood et al., 2017). Thus, the natural variations we identified in the Sitopsis species are potentially valuable for breeding new wheat cultivars. In addition, we also characterized candidate genes that are functionally associated with other important agronomic traits (i.e., tiller number and kernel size) and floral development (i.e., vernalization and photoperiod sensitivity) (Supplemental Table 5). With the availability of high-quality genomes, these genic and genetic resources can be efficiently tapped for future wheat breeding or *de novo* domestication of new types of wheat.

DISCUSSION

We have assembled chromosome-level reference genomes of all five *Aegilops* species of the Sitopsis section and performed comparative genomic analyses both among the five species and with the other available diploid species and polyploid wheat subgenomes. Our main motivation was to better understand the evolutionary histories and trajectories of the Sitopsis species and, especially, the origin of the polyploid wheat B subgenome, as it has long been and is still being debated. A long-standing hypothesis posited that the wheat B subgenome was derived monophyletically from *Ae. speltoides* (Riley et al., 1958). This hypothesis was formulated based on multiple lines of observational and empirical evidence, including botanical, cytological, phylogenetic, and biogeographical data (Feldman and Levy, 2015). This hypothesis was already being questioned nearly 50 years ago based on the near absence of homologous synapsis between the S- and B-subgenome chromosomes in artificial hybrids involving both higher- and lower-pairing types of *Ae. speltoides* (Kimber and Athwal, 1972; Gill and Kimber, 1974; Natarajan and Sarma, 1974). However, it was revitalized by an extensive molecular marker-based population study, which concluded that a yet-undiscovered variant accession or cryptic subspecies of *Ae. speltoides* donated the B subgenome (Kilian et al., 2007). Our genome-scale comparative analyses show that *Ae. speltoides* and the B subgenome diverged ~4.49 mya, i.e., much earlier than the speciation of tetraploid emmer wheat, which occurred approximately 0.8 mya (Marcussen et al., 2014). In other words, a major divergence between *Ae. speltoides* and the B-subgenome diploid donor should have occurred at the diploid level. Moreover, the estimates of genome-wide genetic similarity between the B subgenome and *Ae. speltoides* are far lower than those of the A and D subgenomes from their respective diploid donors, *T. urartu* and *Ae. tauschii*. Together, our results lead to the unequivocal conclusion that *Ae. speltoides* is not the direct progenitor of the B subgenome, which was donated by a distinct, most probably extinct, diploid species of the B lineage.

Another hypothesis for the origin of the wheat B subgenome suggested that it had formed through multiple hybridizations and introgressions of diverse genomic sequences from Sitopsis species at the tetraploid level. According to this polyphyletic scenario, the tetraploid wild emmer wheat (*T. turgidum* ssp. *dicoccoides*) was probably established through the intercrossing of two or more amphiploids with the same A-genome species (*T. urartu*) but different S-genome donors (Sitopsis species) (Sarkar and Stebbins, 1956; Zohary and Feldman, 1962). However, the observed heterogeneous genomic patterns between the five Sitopsis species and the wheat B subgenome are more likely to be due to incomplete sorting of ancestral alleles and B-to-D-lineage genetic introgressions. This refutes the polyphyletic origins of the wheat B subgenome from diverse Sitopsis species at the tetraploid level. Alternatively, the direct progenitor of the wheat B subgenome itself might be of polyphyletic origin through the hybridizations/introgressions between two or more distant ancestral B-lineage species at the diploid level (El Baidouri et al., 2017). However, our comparisons clearly show that the four extant B-lineage species/subgenomes (*Ae. speltoides*, *Am. muticum*, the B subgenome, and the G subgenome) are evolutionarily independent of one another. Taking these results together, we conclude that the direct donor of the B subgenome is a distinct diploid species that diverged from *Ae. speltoides* 4.49 mya but which experienced genetic introgressions with the D-lineage Sitopsis species before its hybridization with *T. urartu* leading to the formation of *T. turgidum*. Similarly, the ca. 450-Mb single-segment-based analyses suggest that a similar situation applies to *T. timopheevii* (GGA^tA^t) and therefore to *T. zhukovskyi* ($GGA^tA^tA^mA^m$), i.e., the G subgenome of these two species was also donated by a distinct diploid species of the B lineage rather than by *Ae. speltoides* (Kilian et al., 2007). Thus,

our results raise an intriguing question for future investigation: why did both of the diploid progenitor species of the B and G subgenomes go extinct while their two congeneric species, *Ae. speltoides* and *Am. muticum*, remain extant? It may be that both diploid donors were outcompeted by their tetraploid progeny, *T. turgidum* ssp. *dicoccoides* and *T. araraticum*, the wild progenitor of *T. timopheevii*, or that the B and G donors remain to be discovered. The latter is possible but very unlikely, considering that significant effort has been devoted to finding these species in the Levant and other relevant regions.

We also performed genomic comparisons to elucidate the evolutionary dynamics of the *Triticum*/*Aegilops* species complex. Our results reveal a highly collinear genome structure among the Sitopsis species, although they all contain high but markedly variable proportions of repetitive sequences. The differences in genome size among the *Triticum*/*Aegilops* species are primarily due to independent post-speciation amplification of a few specific TEs. In addition, we show how detailed comparisons between the reference-quality genome assemblies of the Sitopsis species and the wheat subgenomes may open new avenues for the utilization of these important genic and genetic resources (i.e., homoeologous genes related to agronomic and pathogen-resistant traits) for future wheat breeding. The high-quality genome assemblies of the Sitopsis species, together with those of other species in the *Aegilops*/*Triticum* complex, represent an unprecedented opportunity for further evolutionary, genetic, and breeding studies in the wheat group.

METHODS

Plant materials and DNA and RNA extraction

All plant materials used in this study were collected by Moshe Feldman and curated at the plant germplasm resource of the Weizmann Institute of Science, Israel. One accession was selected for each of the five Sitopsis species, *Ae. bicornis* (accession TB01), *Ae. longissima* (accession TL05), *Ae. searsii* (accession searsii), *Ae. sharonensis* (accession TH02), and *Ae. speltoides* (accession TS01) (Supplemental Table 6). These accessions have been preserved by strict selfing (bagging) since collection. All plants were grown in a greenhouse under temperatures of 25°C/16°C (day/night) and a 16/8 h (day/night) photoperiod. Species status validation was performed by checking the spike morphology and karyotype. Inflorescence morphology was captured with a digital single-lens reflex camera (EOS 6D MARK II) in a photo studio. Karyotypes of the five species were analyzed using fluorescence *in situ* hybridization according to Han et al. (2005) and Kato et al. (2004). In brief, the repetitive DNA sequences pSc119.2 (McIntyre et al., 1990) and pAs1 (Rayburn and Gill, 1986) were labeled with Alexa Fluor 488–5-dUTP (green) and Texas red–5-dCTP (red), respectively. Metaphase chromosome spreads were prepared according to the protocol described in Kato et al. (2004). Karyotypes of the five Sitopsis species were examined with an Olympus BX61 fluorescence microscope (Olympus, Japan). The haploid genome size was estimated by flow cytometry using an Attune focusing analyzer (ABI, CA, USA). Propidium iodide fluorescence was collected using a 620-nm fluorescence-2 (FL2) filter. Parameters for data acquisition were kept constant for all wheat samples, with the genome sizes of *T. urartu* (accession TL38) and *T. turgidum* ssp. *durum* (cultivar TTR19) as controls. Sample flow rate was set to 100 nuclei/s, with each of the wheat samples acuminate for >6000 nuclei. The average of the coefficient of variation value for the G1 peak was used to evaluate the relative genome size. Genomic DNA was isolated from fresh leaf tissue using the CTAB method (Kidwell and Osborn, 1992). Total RNA was extracted from root, leaf, and inflorescence tissues independently based on the standard TRIzol protocol specified by the supplier (Invitrogen, CA, USA).

Genome assembly, gene annotation, and quality assessment

Chromosome-level reference genomes of the five Sitopsis species were assembled by a combination of Oxford Nanopore Technologies (ONT) single-molecule real-time technology and Hi-C-based scaffolding strategy, followed by Illumina short read-based polishing. All the genome assemblies were performed by the Biomarker Company (Beijing, China). In brief, about 740–799 Gb (~114×–178× genome coverage) high-quality ONT long reads were generated using the PromethION platform, and 262–302 Gb (~39.8×–51.4×) short reads were obtained from the Illumina NovaSeq platform (Supplemental Table 6). The long ONT reads were corrected using Canu (Koren et al., 2017) and assembled to long contigs by wtdbg2 (Ruan and Li, 2020). The draft assemblies were then polished using Racon (Vaser et al., 2017) and Pilon (Walker et al., 2014) separately. The Hi-C data (~183×–256×) were used to link the polished contigs into seven pseudochromosomes using LACHESIS (Burton et al., 2013).

Protein coding genes and noncoding RNAs were predicted using *de novo* and homology gene blast strategies (supplemental information). All predicted protein-coding genes were annotated based on the KOG, KEGG, and GO databases. Repetitive elements were identified by LTR_FINDER (Xu and Wang, 2007) and RepeatScout (Price et al., 2005) and then annotated using RepeatMasker (http://www.repeatmasker.org). Quality validation of the genome assemblies was performed by estimating the completeness of the gene repertoire using CEGMA (Parra et al., 2007) and BUSCO (Simão et al., 2015).

Phylogeny, divergence time, and genetic introgression

Phylogenetic topologies and divergence times of the *Triticum*/*Aegilops* species and outgroups were estimated based on whole-genome resequencing data and single-copy orthologous genes. Genome sequences and resequencing data for the five Sitopsis species were generated in this study. Data for the other *Triticum*/*Aegilops* and outgroup species were obtained from previously released reference genomes (supplemental information). Phylogenetic trees were constructed using the maximum-likelihood method implemented in RAxML (v.8.2.12) (Stamatakis, 2014) with the GTR-GAMMA substitution model and 1000 bootstrap replicates. Divergence times of the *Triticum*/*Aegilops* species were estimated using BEAST (v.2.6.0) (Suchard et al., 2018). Tree topologies were summarized using TreeAnnotator (v.2.6.0) (Suchard et al., 2018) and visualized with ggtree (Yu et al., 2017).

Genome-wide interspecific genetic divergence was evaluated for the *Triticum*/*Aegilops* species by calculating the synonymous (d_S) and nonsynonymous (d_N) mutation rates based on PAML (Yang, 2007). In addition, we estimated pairwise branch length and genetic similarity using the *DendroPy* module (Sukumaran and Holder, 2010) in Python. Genetic introgression among the A, B, and D lineages was estimated using a combination of the *D* statistic, *fd*, hybrid index (γ), and χ^2 goodness of fit test (Martin et al., 2015; Blischak et al., 2018; Suvorov et al., 2021). Introgressed variants among the four D-lineage Sitopsis species were identified based on the shared-specific SNPs for each species pair compared with the other species.

Repetitive element classification and expansion

Full-length LTR retrotransposons in the five Sitopsis and the other *Triticum*/*Aegilops* species were characterized using LTRharvest (Ellinghaus et al., 2008) and LTR_FINDER (Xu and Wang, 2007). Then, the program RepeatScout (Price et al., 2005) was used to build consensus LTR retrotransposon sequences. DNA transposons were identified by a homology search against the REdat_9.7_Triticeae subset of the PGSB transposon library (Spannagl et al., 2016) using vmatch (http://www.vmatch.de). The above-identified DNA transposons and LTR

retrotransposons were classified into different families using RepeatMasker (http://www.repeatmasker.org) and CLARI-TE procedures (Daron et al., 2014). Insertion time of each LTR retrotransposon family was estimated using the formula age = $K/2r$, where K is the Kimura two-parameter distance and r is the mutation rate of 1.3×10^{-8} (Wicker et al., 2018). K-mer analyses of the wheat genomes were performed by KAT (Mapleson et al., 2017) for each genome/subgenome.

Genome collinearity and pan-genomic analyses

Genome collinearity between the five Sitopsis and the other *Triticum/Aegilops* species was performed using MCScanX (Wang et al., 2012). Interspecific homologous genes were characterized using BLASTP with the default parameters. The resulting syntenic genomic regions were used to identify orthologous genes using ColinearScan (Wang et al., 2006). Putative protogenes were characterized for the A, B, and D lineages according to previously published protocols (Murat et al., 2017; Pont et al., 2019). Pan-genomic analyses were performed based on both the protein-coding genes and the genome SVs. For the protein-coding genes, all three polyploid wheat subgenomes (A, B, and D) and their diploid donors (*Ae. tauschii* and *T. urartu*) were mixed as a new *in silico* species called "ABD." Then, the five Sitopsis species and "ABD" were used to identify gene families using OrthoFinder (Emms and Kelly, 2019). Gene family expansion and contraction were inferred using a previously established phylogenomic approach (Appels et al., 2018). Log-transformed gene family sizes among 15 genomes and/or subgenomes were compared using the phylANOVA function of the phytools package (https://github.com/liamrevell/phytools) in R according to the guidance of the phylogenetic species tree. Genome SVs of *Ae. tauschii*, *T. urartu*, and the five Sitopsis species were characterized based on the ONT data. The corrected ONT long reads were mapped onto the common wheat (Chinese Spring) B subgenome IWGSC_v1.0 using minimap2 (Li, 2018) and predicted using Sniffles (Sedlazeck et al., 2018). Only the uniquely mapped reads with mapping quality >30, depth >10, and alternative allele ratio >0.2 were kept for subsequent analyses. GO enrichment of the candidate protein-coding and SV-related genes was performed using clusterProfiler (Yu et al., 2012).

Identification of resistance and agronomic trait-related genes

The nucleotide-binding and leucine-rich repeat immune receptor (NLR) genes were identified using the NLR-Annotator pipeline (Zhang, 2020). In brief, candidate NLR genes were predicted by searching annotated CDS against the Pfam database (https://pfam.xfam.org). A custom Python script was used to classify NLR genes according to the position of the intron located inside or outside of the NB_ARC domain. The other agronomic trait-related genes in the five Sitopsis species were identified by searching the coding sequences of candidate genes against the genome assemblies using BLASTN with E value <10 e−5 and hit length >300 bp. All candidate hits were manually checked and compared with corresponding protein annotation databases.

ACCESSION NUMBERS

All data supporting the findings of this study are available in the paper and the supplemental information files. Raw sequence data and genome assemblies have been deposited at the National Center for Biotechnology Information under BioProject accession no. PRJNA700474 and the National Genomics Data Center under project no. PRJCA007359.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at *Molecular Plant Online*.

FUNDING

This study was supported by the Natural Science Foundation of China (31991211 to B.L. and 31970235 to L.F.L.), the Shanghai Pujiang Program (19PJ1401500 to L.F.L.), Israel Science Foundation (ISF)–China National Natural Science Foundation (NSFC) collaborative grants to B.L. (32061143001) and A.A.L. (3394/20), and a China Postdoctoral Science Foundation grant (2021M690683).

AUTHOR CONTRIBUTIONS

L.F.L., A.A.L., and B.L. conceived this project. L.G., L.F.L., A.A.L., and B.L. designed and supervised the project. Z.B.Z., Z.H.W., N.L., Y.S., X.F.W., N.D., Y.L., J.Z., Y.W., L.G., and F.M conducted the experiments and analyzed the data. L.F.L., A.A.L., L.G., F.M., and B.L. wrote the manuscript. All authors discussed the results and approved the manuscript.

ACKNOWLEDGMENTS

We thank Moshe Feldman for critical reading of the manuscript and constructive comments. No conflict of interest is declared.

Received: November 5, 2021
Revised: December 11, 2021
Accepted: December 28, 2021
Published: December 31, 2021

REFERENCES

Ankori, H., and Zohary, D. (1962). Natural hybridization between *Aegilops sharonensis* and *Ae. longissima*: a morphological and cytological study. Cytologia **27**:314–324.

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., Pozniak, C.J., Choulet, F., Distelfeld, A., and Poland, J. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science **361**:eaar7191.

Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S., Gundlach, H., Hale, I., Mascher, M., Spannagl, M., and Wiebe, K. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science **357**:93–97.

Bernhardt, N., Brassac, J., Dong, X., Willing, E.M., Poskar, C.H., Kilian, B., and Blattner, F.R. (2020). Genome-wide sequence information reveals recurrent hybridization among diploid wheat wild relatives. Plant J. **102**:493–506.

Blischak, P.D., Chifman, J., Wolfe, A.D., and Kubatko, L. (2018). HyDe: a python package for genome-scale hybridization detection. Syst. Biol. **67**:821–829.

Brinton, J., Ramirez-Gonzalez, R.H., Simmonds, J., Wingen, L., Orford, S., Griffiths, S., Haberer, G., Spannagl, M., Walkowiak, S., Pozniak, C., and Uauy, C. (2020). A haplotype-led approach to increase the precision of wheat breeding. Commun. Biol. **3**:712.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat. Biotechnol. **31**:1119–1125.

Chen, Y.M., Song, W.J., Xie, X.M., Wang, Z.H., Guan, P.F., Peng, H.R., Jiao, Y.N., Ni, Z.F., Sun, Q.X., and Guo, W.L. (2020). A collinearity-incorporating homology inference strategy for connecting emerging assemblies in Triticeae tribe as a pilot practice in the plant pangenomic era. Mol. Plant **3**:1694–1708.

Daron, J., Glover, N., Pingault, L., Theil, S., Jamilloux, V., Paux, E., Barbe, V., Mangenot, S., Alberti, A., and Wincker, P. (2014). Organization and evolution of transposable elements along the bread wheat chromosome 3B. Genome Biol. **15**:546.

Dvorak, J., and Zhang, H.B. (1990). Variation in repeated nucleotide sequences sheds light on the phylogeny of the wheat B and G genomes. Proc. Natl. Acad. Sci. **87**:9640–9644.

Dvorak, J., Wang, L., Zhu, T., Jorgensen, C.M., Luo, M.C., Deal, K.R., Gu, Y.Q., Gill, B.S., Distelfeld, A., and Devos, K.M. (2018). Reassessment of the evolution of wheat chromosomes 4A, 5A, and 7B. Theor. Appl. Genet. **131**:2451–2462.

Dvořák, J. (1976). The relationship between the genome of *Triticum urartu* and the A and B genomes of *Triticum aestivum*. Can. J. Genet. Cytol. **18**:371–377.

Dvořák, J., Terlizzi, P.d., Zhang, H.-B., and Resta, P. (1993). The evolution of polyploid wheats: identification of the A genome donor species. Genome **36**:21–31.

El Baidouri, M., Murat, F., Veyssiere, M., Molinier, M., Flores, R., Burlot, L., Alaux, M., Quesneville, H., Pont, C., and Salse, J. (2017). Reconciling the evolutionary origin of bread wheat (*Triticum aestivum*). New Phytol. **213**:1477–1486.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinf. **9**:18. https://doi.org/10.1186/1471-2105-9-18.

Emms, D., and Kelly, S.L. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. **20**:238.

Faris, J.D., and Gill, B.S. (2002). Genomic targeting and high-resolution mapping of the domestication gene Q in wheat. Genome **45**:706–718.

Feldman, M., and Kislev, M.E. (2007). Domestication of emmer wheat and evolution of free-threshing tetraploid wheat. Isr. J. Plant Sci. **55**:207–221.

Feldman, M., and Levy, A.A. (2015). Origin and evolution of wheat and related Triticeae species. In Alien Introgression in Wheat (Switzerland: Springer), pp. 21–76.

Feldman, M., Lupton, F., and Miller, T. (1995). Wheats. In Evolution of Crop Plants, 2nd, J. Smartt and N.W. Simmonds, eds. (London: Longman Scientific), pp. 184–192.

Gill, B., and Friebe, B. (2002). Cytogenetics, phylogeny and evolution of cultivated wheats. Bread Wheat: Improvement Production. Food Agric. Organ. United Nations, Rome, 71–88.

Gill, B.S., and Kimber, G. (1974). Giemsa C-banding and the evolution of wheat. Proc. Natl. Acad. Sci. **71**:4086–4090.

Glémin, S., Scornavacca, C., Dainat, J., Burgarella, C., Viader, V., Ardisson, M., Sarah, G., Santoni, S., David, J., and Ranwez, V. (2019). Pervasive hybridizations in the history of wheat relatives. Sci. Adv. **5**:eaav9188.

Gornicki, P., Zhu, H., Wang, J., Challa, G.S., Zhang, Z., Gill, B.S., and Li, W. (2014). The chloroplast view of the evolution of polyploid wheat. New Phytol. **204**:704–714.

Greenwood, J.R., Finnegan, E.J., Watanabe, N., Trevaskis, B., and Swain, S.M. (2017). New alleles of the wheat domestication gene *Q* reveal multiple roles in growth and reproductive development. Development **144**:1959–1965.

Guo, W., Xin, M., Wang, Z., Yao, Y., Hu, Z., Song, W., Yu, K., Chen, Y., Wang, X., Guan, P., et al. (2020). Origin and adaptation to high altitude of Tibetan semi-wild wheat. Nat. Commun. **11**:5085.

Han, F., Fedak, G., Guo, W., and Liu, B. (2005). Rapid and repeatable elimination of a parental genome-specific DNA repeat (pGc1R-1a) in newly synthesized wheat allopolyploids. Genetics **170**:1239–1245.

Hao, C., Jiao, C.Z., Hou, J., Li, T., Liu, H.X., Wang, Y., Zheng, J., Liu, H., Bi, Z., Xu, F., and Zhao, J. (2020). Resequencing of 145 landmark cultivars reveals asymmetric sub-genome selection and strong founder genotype effects on wheat breeding in China. Mol. Plant **13**:1733–1751.

He, F., Pasam, R., Shi, F., Kant, S., Keeble-Gagnere, G., Kay, P., Forrest, K., Fritz, A., Hucl, P., Wiebe, K., et al. (2019). Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nat. Genet. **51**:896–904.

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

Huynh, S., Marcussen, T., Felber, F., and Parisod, C. (2019). Hybridization preceded radiation in diploid wheats. Mol. Phylogenet. Evol. **139**:106554.

Jiang, Y., Yuan, Z., Hu, H., Ye, X., Zheng, Z., Wei, Y., Zheng, Y.L., Wang, Y.G., and Liu, C. (2020). Differentiating homoploid hybridization from ancestral subdivision in evaluating the origin of the D lineage in wheat. New Phytol. **228**:409–414.

Kato, A., Lamb, J.C., and Birchler, J.A. (2004). Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. Proc. Natl. Acad. Sci. **101**:13554–13559.

Kidwell, K.K., and Osborn, T.C. (1992). Simple plant DNA isolation procedures. In Plant genomes: methods for genetic and physical mapping (Springer), pp. 1–13.

Kihara, H. (1944). Discovery of the DD-analyser, one of the ancestors of *Triticum vulgare*. Agric. Hort. **19**:13–14.

Kihara, H. (1954). Considerations on the evolution and distribution of *Aegilops* species based on the analyser-method. Cytologia **19**:336–357.

Kilian, B., Özkan, H., Deusch, O., Effgen, S., Brandolini, A., Kohl, J., Martin, W., and Salamini, F. (2007). Independent wheat B and G genome origins in outcrossing *Aegilops* progenitor haplotypes. Mol. Biol. Evol. **24**:217–227.

Kimber, G., and Athwal, R. (1972). A reassessment of the course of evolution of wheat. Proc. Natl. Acad. Sci. U S A **69**:912–915.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. **27**:722–736.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34**:3094–3100.

Ling, H., Ma, B., Shi, X., Liu, H., Dong, L., Sun, H., Cao, Y., Gao, Q., Zheng, S., and Li, Y. (2018). Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. Nature **557**:424–428.

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

Luo, M., Gu, Y.Q., Puiu, D., Wang, H., Twardziok, S., Deal, K.R., Huo, N., Zhu, T., Wang, L., and Wang, Y. (2017). Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. Nature **551**:498–502.

Maccaferri, M., Harris, N.S., Twardziok, S., Pasam, R.K., Gundlach, H., Spannagl, M., Ormanbekova, D., Lux, T., Prade, V.M., and Milner, S.G. (2019). Durum wheat genome highlights past domestication signatures and future improvement targets. Nat. Genet. **51**:885–895.

Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., and Clavijo, B.J. (2017). KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics **33**:574–576.

Marcussen, T., Sandve, S.R., Heier, L., Spannagl, M., Pfeifer, M., Jakobsen, K.S., Wulff, B.B.H., Steuernagel, B., Mayer, K.F.X., and Olsen, O. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. Science **345**:1250092.

Martin, S.H., Davey, J.W., and Jiggins, C.D. (2015). Evaluating the use of ABBA–BABA Statistics to locate introgressed loci. Mol. Biol. Evol. **32**:244–257.

Matsuoka, Y. (2011). Evolution of polyploid *Triticum* wheats under cultivation: the role of domestication, natural hybridization and

allopolyploid speciation in their diversification. Plant Cell Physiol. **52**:750–764.

McIntyre, C., Pereira, S., Moran, L., and Appels, R. (1990). New Secale cereale (rye) DNA derivatives for the detection of rye chromosome segments in wheat. Genome **33**:635–640.

Miki, Y., Yoshida, K., Mizuno, N., Nasuda, S., Sato, K., and Takumi, S. (2019). Origin of wheat B-genome chromosomes inferred from RNA sequencing analysis of leaf transcripts from section Sitopsis species of *Aegilops*. DNA Res. **26**:171–182.

Murat, F., Armero, A., Pont, C., Klopp, C., and Salse, J. (2017). Reconstructing the genome of the most recent common ancestor of flowering plants. Nat. Genet. **49**:490–496.

Natarajan, A., and Sarma, N. (1974). Chromosome banding patterns and the origin of the B genome in wheat. Genet. Res. **24**:103–108.

Nesbitt, M., and Samuel, D. (1996). From staple crop to extinction? The archaeology and history of the hulled wheat. In Hulled Wheats, Promoting the Conservation and Used of Underutilized and Neglected Crops, S. Padulosi, K. Hammer, and J. Heller, eds. (Rome: IPGRI), pp. 40–99.

Ogihara, Y., and Tsunewaki, K. (1988). Diversity and evolution of chloroplast DNA in *Triticum* and *Aegilops* as revealed by restriction fragment analysis. Theor. Appl. Genet. **76**:321–332.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23**:1061–1067.

Petersen, G., Seberg, O., Yde, M., and Berthelsen, K. (2006). Phylogenetic relationships of *Triticum* and *Aegilops* and evidence for the origin of the A, B, and D genomes of common wheat (*Triticum aestivum*). Mol. Phylogenet. Evol. **39**:70–82.

Pont, C., Wagner, S., Kremer, A., Orlando, L., Plomion, C., and Salse, J. (2019). Paleogenomics: reconstruction of plant evolutionary trajectories from modern and ancient DNA. Genome Biol. **20**:29.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). *De novo* identification of repeat families in large genomes. Bioinformatics **21**:i351–i358.

Rayburn, A.L., and Gill, B.S. (1986). Isolation of a D-genome specific repeated DNA sequence from Aegilops squarrosa. Plant Mol. Biol. Rep. **4**:102–109.

Riley, R., Unrau, J., and Chapman, V. (1958). Evidence on the origin of the B genome of wheat. J. Hered. **49**:91–98.

Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nat. Methods **17**:155–158.

Ruban, A.S., and Badaeva, E.D. (2018). Evolution of the S-genomes in *Triticum-Aegilops* alliance: evidences from chromosome analysis. Front. Plant Sci. **9**:1756.

Salse, J., Chagué, V., Bolot, S., Magdelenat, G., Huneau, C., Pont, C., Belcram, H., Couloux, A., Gardais, S., and Evrard, A. (2008). New insights into the origin of the B genome of hexaploid wheat: evolutionary relationships at the SPA genomic region with the S genome of the diploid relative *Aegilops speltoides*. BMC Genom. **9**:555.

Sarkar, P., and Stebbins, G. (1956). Morphological evidence concerning the origin of the B genome in wheat. Am. J. Bot. **43**:297–304.

Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., and Schatz, M.C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. Nat. Methods **15**:461–468.

Shewry, P.R., and Hey, S.J. (2015). The contribution of wheat to human diet and health. Food Energy Secur. **4**:178–202.

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

Simons, K.J., Fellers, J.P., Trick, H.N., Zhang, Z., Tai, Y.-S., Gill, B.S., and Faris, J.D. (2006). Molecular characterization of the major wheat domestication gene *Q*. Genetics **172**:547–555.

Spannagl, M., Nussbaumer, T., Bader, K.C., Martis, M.M., Seidel, M., Kugler, K.G., Gundlach, H., and Mayer, K.F. (2016). PGSB PlantsDB: updates to the database framework for comparative plant genome research. Nucleic Acids Res. **44**:D1141–D1147. https://doi.org/10.1093/nar/gkv1130.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**:1312–1313.

Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. **4**:vey016.

Sukumaran, J., and Holder, M.T. (2010). DendroPy: a Python library for phylogenetic computing. Bioinformatics **26**:1569–1571.

Suvorov, A., Scornavacca, C., Fujimoto, M.S., Bodily, P., Clement, M., Crandall, K.A., Whiting, M.F., Schrider, D.R., and Bybee, S.M. (2021). Deep ancestral introgression shapes evolutionary history of dragonflies and damselflies. Syst Biol. syab063. https://doi.org/10.1093/sysbio/syab063.

Tsunewaki, K., Wang, G.-Z., and Matsuoka, Y. (1996). Plasmon analysis of *Triticum* (wheat) and *Aegilops*. 1. Production of alloplasmic common wheats and their fertilities. Genes Genet. Syst. **71**:293–311.

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. **27**:737–746.

Wagenaar, E. (1966). Studies on the genome constitution of *Triticum timopheevi* Zhuk. II. The *T. timopheevi* complex and its origin. Evolution **20**:150–164.

Waines, J.G., and Johnson, B.L. (1972). Genetic differences between *Aegilops longissima, A. sharonensis*, and *A. bicornis*. Can. J. Genet. Cytol. **14**:411–415.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., and Young, S.K. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One **9**:e112963.

Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M.T., Brinton, J., Ramirez-Gonzalez, R.H., Kolodziej, M.C., Delorean, E., and Thambugala, D. (2020). Multiple wheat genomes reveal global variation in modern breeding. Nature **588**:277–283.

Wang, X., Shi, X., Li, Z., Zhu, Q., Kong, L., Tang, W., Ge, S., and Luo, J. (2006). Statistical inference of chromosomal homology based on gene colinearity and applications to Arabidopsis and rice. BMC Bioinf. **7**:447.

Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., and Guo, H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. **40**:e49.

Wang, X., Zhang, H., Li, Y., Zhang, Z., Li, L., and Liu, B. (2016). Transcriptome asymmetry in synthetic and natural allotetraploid wheats, revealed by RNA-sequencing. New Phytol. **209**:1264–1277.

Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramirezgonzalez, R.H., De Oliveira, R., Mayer, K.F.X., Paux, E., and Choulet, F. (2018). Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biol. **19**:1–18.

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. **35**:W265–W268. https://doi.org/10.1093/nar/gkm286.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24**:1586–1591.

Yu, G., Wang, L.G., Han, Y., and He, Q. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. Omics A J. Integr. Biol. **16**:284–287.

Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.Y. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. **8**:28–36.

Zhang, W. (2020). NLR-Annotator: a tool for *de novo* annotation of intracellular immune receptor repertoire. Plant Physiol. **183**:418–420.

Zhang, Z., Belcram, H., Gornicki, P., Charles, M., Just, J., Huneau, C., Magdelenat, G., Couloux, A., Samain, S., and Gill, B.S. (2011). Duplication and partitioning in evolution and function of homoeologous *Q* loci governing domestication characters in polyploid wheat. Proc. Natl. Acad. Sci. U S A **108**:18737–18742.

Zhang, H., Bian, Y., Gou, X., Dong, Y., Rustgi, S., Zhang, B., Xu, C., Li, N., Qi, B., and Han, F. (2013). Intrinsic karyotype stability and gene copy number variations may have laid the foundation for tetraploid wheat formation. Proc. Natl. Acad. Sci. U S A **110**:19466–19471.

Zhang, H., Yang, Y., Wang, C., Liu, M., Li, H., Fu, Y., Wang, Y., Nie, Y., Liu, X., and Ji, W. (2014). Large-scale transcriptome comparison reveals distinct gene activations in wheat responding to stripe rust and powdery mildew. BMC Genom. **15**:898.

Zohary, D., and Feldman, M. (1962). Hybridization between amphidiploids and the evolution of polyploids in the wheat (*Aegilops-Triticum*) group. Evolution, 44–61.

Molecular Plant, Volume *15*

Supplemental information

Genome sequences of five Sitopsis species of *Aegilops* and the origin of polyploid wheat B subgenome

Lin-Feng Li, Zhi-Bin Zhang, Zhen-Hui Wang, Ning Li, Yan Sha, Xin-Feng Wang, Ning Ding, Yang Li, Jing Zhao, Ying Wu, Lei Gong, Fabrizio Mafessoni, Avraham A. Levy, and Bao Liu

Running title: Genome resources of the *Aegilops* Sitopsis species

Genome sequences of the five Sitopsis species of *Aegilops* and the origin of polyploid wheat B subgenome

Lin-Feng Li[1,2,#,*], Zhi-Bin Zhang[1,3,#], Zhen-Hui Wang[4], Ning Li[1], Yan Sha[1], Xin-Feng Wang[2], Ning Ding[2], Yang Li[1], Jing Zhao[1], Ying Wu[1], Lei Gong[1], Fabrizio Mafessoni[3], Avraham A. Levy[3,*], and Bao Liu[1,*]

***Correspondence**:
Lin-Feng Li
Email: lilinfeng@fudan.edu.cn
Avraham A. Levy
avi.levy@weizmann.ac.il
Bao Liu
Email: baoliu@nenu.edu.cn

[#] These authors contributed equally to this work.

Address:
[1] Key Laboratory of Molecular Epigenetics of the Ministry of Education (MOE), Northeast Normal University, Changchun 130024, China;
[2] Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, School of Life Sciences, Fudan University, Shanghai 200438, China;
[3] Department of Plant and Environmental Sciences, The Weizmann Institute of Science, 76100 Rehovot, Israel;
[4] Faculty of Agronomy, Jilin Agricultural University, Changchun 130118, China.

Supplementary information

This file includes:

Supplementary Notes

Supplementary Figures 1-23

Genome survey, assembly and Hi-C scaffolding

Genome features of the five Sitopsis species were surveyed by GenomeScope (Vurture *et al*., 2017) based on Illumina short reads. The Illumina DNA paired-end libraries were constructed with an insert length of 350 bp, and 262.71-302.73 Gb (44.53-66.42× coverage) sequencing data were produced on the Illumina Novaseq platform (Illumina, San Diego, CA, USA) according to the manufacturer's instructions (**Supplementary Fig. 22**). Genome size and heterozygosity of the five species varied from 4.49-6.60 Gb and 0.01-0.37, respectively. Reference genomes of the five species were assembled by a combination of short-read Illumina and long-range Nanopore sequences. DNA libraries for real-time single-molecule sequencing were prepared according to the manufacturer's instructions, and sequenced on Nanopore PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK). About 739.62-798.86 Gb long reads were obtained from the five species, with contig N50 length of 28,616-34,991 bp and mean length of 23,221-27,320 bp (**Supplementary Table 6**). The long sequence consensus was corrected using Canu (Koren *et al*., 2017) and assembled using Wtdbg2 (Ruan and Li, 2020) and NextDenovo (https://github.com/Nextomics/NextDenovo). Genome assemblies of the five species were polished by Racon (Vaser *et al*., 2017) and Pilon (Walker *et al*., 2014) based on long-read and short-read data, respectively. The draft assembled genomes consist of 1,154-35,663 contigs, with contig N50 length of 563,410-8,720,942 bp (**Supplementary Table 7**). The genome completeness was assessed using a terrestrial plant dataset from BUSCO (Simão *et al*., 2015) and a eukaryote dataset from CEGMA (Parra *et al*., 2007). All the five Sitopsis species show high level of genome completeness (BUSCOs: 92.92-95.10%; CEGs: 95.85-99.13%) (**Supplementary Table 8**). Chromosomal level reference genomes were generated by link the contigs into seven pseudochromosomes based on the Hi-C data using LACHESIS (Burton *et al*., 2013). About 1,051.32-1,234.93 Gb (179.15-280.09x genome coverage) Hi-C short reads were generated for the five species on Illumina Novaseq platform (Illumina, San Diego, CA, USA) (**Supplementary Table 9**). The construction of pseudomolecules was validated by counting the valid interaction read pairs within each 500-kb bin. The heatmap revealed high interactions between the adjacent genomic regions of all the seven pseudomolecules (**Supplementary Fig. 23**). A total of 49.06-57.68% (95.21-97.07% of the assembled genome) of the contigs were clustered, 47.55-53.62% (90.83-93.44% of the genome) of which were ordered on the on the seven pseudomolecules (**Supplementary Table 10**).

Annotation of gene and repetitive sequence

Protein-coding genes were annotated by the combination of *de novo*, homeolog and unigene prediction strategies. The *de novo* annotation of the five species was performed using the programs

Genscan (Burge and Karlin, 1997), Augustus (Stanke and Waack, 2003), GlimmerHMM (Majoros et al., 2004), GeneID (Blanco *et al*., 2007) and SNAP (Korf, 2004). Gene homeolog was predicted using GeMoMa (Keilwagen *et al*., 2016). Unigenes were identified based on the transcriptome data generated from the leaf and root tissues of each species. The transcriptome data were mapped on the assembled genomes using Hisat2 (Kim *et al*., 2019) and Stringtie (Pertea *et al*., 2015). Then, the programs TransDecoder (https://github.com/TransDecoder/TransDecoder) and GeneMarkS-T (Tang *et al*., 2015) were employed to predict the protein-coding genes. In addition, we also used the program PASA (Campbell *et al*., 2006) to predict protein-coding genes. All the protein-coding genes annotated by the above three strategies were combined by EVM (Haas *et al*., 2008). About 61,354-63,326 protein-coding genes (37,201-40,222 of which are high confidence genes) were predicted in the five species (**Supplementary Table 11**). The total length of the protein-coding genes varied from 213,009,021-319,640,000 bp, with each gene being on average 3,364-5,168 bp long in the five species. Functions of these protein-coding gene were annotated by searching against the databases GO (http://geneontology.org/), TrEMBL (http://www.bioinfo.pte.hu/more/TrEMBL.htm), NR (https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/) KOG (Koonin *et al*., 2004), and KEGG (https://www.genome.jp/kegg/pathway.html) using BLAST (https://blast.ncbi.nlm.nih.gov/ Blast.cgi). We annotated 56,727-60,923 (92.22-96.21% of the total) protein-coding genes (**Supplementary Table 12**).

Repetitive sequences in the five species were identified using the programs LTR_FINDER (Xu and Wang, 2007) and RepeatScout (Price *et al*., 2005). The identified repetitive sequences were then classified using PASTEClassifier (Hoede *et al*., 2014) and annotated using RepeatMasker (http://repeatmasker.org/). A total of 2,936,990,453- 4,145,555,209 bp Class I (67.56-71.47% of the total genome) and 515,361,078- 1,025,079,256 bp Class II (12.54-19.21% of the total genome) repetitive sequences were characterized in the five species (**Supplementary Table 13**). In the Class I, the *Copia* and *Gypsy* families consist of 16.64-22.80% and 43.72-49.58% of the assembled genome, respectively. Likewise, the Class II CACTA represent 11.06-19.21% of the assembled genome in the five species.

Molecular phylogeny and divergence time

Phylogenetic inferences were performed based on single copy orthologous genes (SCOGs), reduced representative genomic regions (RRGRs) and whole genome single nucleotide polymorphisms (SNPs). The SCOGs were characterized from seven diploid *Triticum*/*Aegilops* species, seven polyploid wheat subgenomes and three outgroup species, including *Triticum urartu* (abbreviated as Tura) (Ling *et al.,*

2018), *Aegilops tauschii* (Atau) (Luo *et al*., 2017), *Ae. speltoides* (Aspe), *Ae. bicornis* (Abic), *Ae. longissima* (Alon), *Ae. searsii* (Asea), *Ae. sharonensis* (Asha), AA (Emaa) and BB (Emab) subgenomes of *T. turgidum*, ssp. *dicoccoides* (Avni *et al.*, 2017), AA (Sveb) and BB (Emab) subgenomes of *T. turgidum*, ssp. *durum* (Maccaferri *et al.*, 2019), AA (Taea), BB (Taeb) and DD (Taed) subgenomes of *T. aestivum* (Appels *et al*., 2018), *Brachypodium distachyon* (Bdis) (Vogel *et al.*, 2010), *Hordeum vulgare* (Hvul) (Mascher *et al.*, 2017), *Elymus elongatum* (Ele) (Wang *et al.*, 2020). Homologous proteins of these diploid species and polyploid wheat subgenomes were identified using BLASTP (https://blast.ncbi.nlm.nih.gov/Blast.cgi) with E-value < 10^{e-5}. The program OrthoFinder (Emms and Kelly, 2019) was performed to cluster the homologous as gene families. In total, 62,543 gene families were identified from these species, of which, 37,751 contain no more than one (none or one) gene copy. We then selected 3,588 gene families (singleton) that include one copy in all these species to performed phylogenetic inference. Coding domain sequences of the singleton genes were aligned based on their amino acid sequences using Guidance (Sela *et al*., 2015). A total of 2,099 gene families that generated alignments with MEAN_RES_PAIR_SCORE > 0.9 and MEAN_COL_Score > 0.9 were retained. Then, the program Gblocks (Castresana, 2000) was employed to further filter the low quality alignments and concatenated the high quality alignments as a single matrix. Phylogenetic tree was reconstructed using a maximum likelihood method implemented in RAxML (v.8.2.12) (Stamatakis, 2014) of a GTR-GAMMA substitution model with 1,000 bootstrap replicates.

Phylogenetic relationships of the *Triticum*/*Aegilops* species were also inferred based on RRGRs using the following procedures: (1) genic region together with 20-kb up-stream and 20-kb down-stream genomic regions were retrieved from the diploid species and polyploid wheat subgenomes; (2) all selected RRGRs (no overlaps) were aligned to bread wheat B-subgenome using the nucmer of MUMmer v3.9 (Kurtz *et al.*, 2004) with the parameters --mum -c 90 -l 40; (3) the program *delta-filter* was employed to extract one-to-one query-reference alignments with parameters -1; (4) the *show-coords* was used to obtain chromosome coordinates of each alignment; (5) the Intersect module of Bedtools (https://bedtools.readthedocs.io/en/latest/) was performed to search the intersections of alignment blocks of all query genome/subgenome regions to construct the originally conserved representative genomic sequence for bread wheat B-subgenome; (6) the reduced reference genome sequences were aligned against their own reference genomes to acquire the corresponding conserved representative genomic regions using Minimap2 (v2.17) (Li, 2018) with parameter "–x asm20". In total, we obtained ~30 Mb conserved representative genomic regions including 24,239 alignments among all genomes/subgenomes (excepting the two outgroups Bdis and Hvul). Then, these alignments were assigned to 2,318 genomic blocks (2-Mb in length) according to the coordinates of bread wheat B-subgenome. Based on this sequence matrix, we further retrieved

RRGRs from the chromosome 2B of bread wheat cultivar "LongReach Lancer", which supposed to be introgressed from *T. timopheevii*. In addition, we also reconstructed phylogenies of the *Triticum*/*Aegilops* species based on whole genome resequencing data. The clean short reads of seven diploid *Triticum*/*Aegilops* species were mapped onto bread wheat B-subgenome using BWA (http://bio-bwa.sourceforge.net/). Genome-wide SNPs were reported using SAMtools (http://www.htslib.org/). Phylogenetic tree was reconstructed using a maximum likelihood method implemented in RAxML (v.8.2.12) (Stamatakis, 2014) of a GTR-GAMMA substitution model with 1,000 bootstrap replicates.

Divergence times among these *Triticum*/*Aegilops* species were estimated based on SCOGs, RRGRs and genome-wide SNPs using the program BEAST (v.2.6.0) (Suchard *et al*., 2018) with the following parameters, strict molecular clock, HKY + gamma nucleotide substitution model, Yule priors and running for 100 million MCMC generations with parameters sampled every 10000 generations. The previously estimated divergence time between *B. distachyon* and *Triticum*/*Aegilops* (mean 44.4 ± stdev 3.53 Ma) (Marcussen *et al.*, 2014) was used as date calibration. TreeAnnotator (v.2.6.0) was used to summarize the output. The resulting phylogenies were visualized with ggtree (Yu *et al*., 2017).

Genome-wide genetic similarity

Genome-wide similarity was estimated based on the above generated RRGRs. Pair-wise nucleotide diversity between any two of these diploid species and polyploid wheat subgenomes were calculated using the DendroPy module (Sukumaran and Holder, 2010) in python. Maximum likelihood tree was reconstructed using RAxML (Stamatakis, 2014) with GTR-GAMMA model. The R package Ape v5.1 (Paradis *et al.*, 2004) was used to manipulate phylogenetic trees with the following parameters: chronos function was used to make ultrametric trees from input trees; conphenetic.phylo function computed the pairwise branch lengths between any pairs of phylogenetic tips; keep.tip function extracted needed tips and generate local tree topologies; tree topology was visualized with plot.phylo function. The Treedist function of R package phangorn was employed to calculate the difference between phylogeny trees to classify tree topologies of each reduced representative genome region. Tree topologies along each chromosome were visualized with RIdeogram (https://github.com/TickingClock1992/RIdeogram).

Estimation of genetic introgression

Reduced representative genomic regions of the nine diploid species and bread wheat B-subgenome were employed to estimate introgression events. Given a quartet of three species and a constant outgroup (*Elymus elongatum*) with the relationship (((P1, P2), P3), O) corresponding to *Triticum*/*Aegilops* phylogeny topology, we performed different statistics, such as *D*, *fd*, hybrid index (*γ*) and χ^2 goodness of fit test, to systematically infer the introgression/hybridization events among the selected *Triticum*/*Aegilops* species. The statistics *D* and *fd* were calculated with genomics_general tools (Martin *et al.*, 2015) (https://github.com/simonhmartin/genomics_general). We used a significance threshold of 0.001 for the *Z*-test. The program HyDe (Blischak *et al.*, 2018) was used to calculate the hybrid index (*γ*), which quantifies the proportion of P3 contributed to P2. In addition, *χ2* goodness of fit test is also applied to identify genetic introgressions based on the counts of three different sliding-window trees, with the (((P1, P2), P3) representing the topology of species tree and ((P2, P3), P1) and ((P1, P3), P2) representing discordant trees (Suvorov *et al*., 2020). Ratios between two discordant tree topologies that are significantly deviated from 1 (*p*-value < 0.001) indicate the presence of introgression event.

To further identify the precise introgression among the *Triticum/Aegilops* species, pair-wise alignment at the chromosome-level was performed for the available diploid species genomes and polyploid wheat subgenomes using the nucmer of MUMmer (v3.9) (Kurtz *et al.*, 2004) with the parameters "*--mum -c 90 -l 40*". These alignments including at least large 10 blocks (>20,000 bp) and spanning sequence larger than 1Mb were treated as candidate introgression regions. Then, we employed Minimap2 (v2.17) (Li, 2018) to assign Illumina short reads of query species to the reference genome of the putative introgressed species with parameters "*-ax sr -I 10G*". Introgressed genomic regions were then manually checked based on sequencing depth around the candidate introgressed genomic regions.

Evolutionary history of the gene families

To obtain the core protein-coding gene set of the *Triticum*/*Aegilops* species, we clustered the homologous into gene families using OrthoFinder (Emms and Kelly, 2019). All the three polyploid wheats (wild emmer, durum and bread wheat) and their A- and D-subgenome donors, *T. urartu* and *Ae. tauschii,* and five Sitopsis species were used to construct presence/absence matrix according to the member count in each family (presence marked as "1" whereas absence marked as "0"). In total, 49,583 gene families corresponding to three categories were reconstructed, including: (1) core gene family, present in all *Triticum*/*Aegilops* species; (2) dispensable gene family, present in either at least

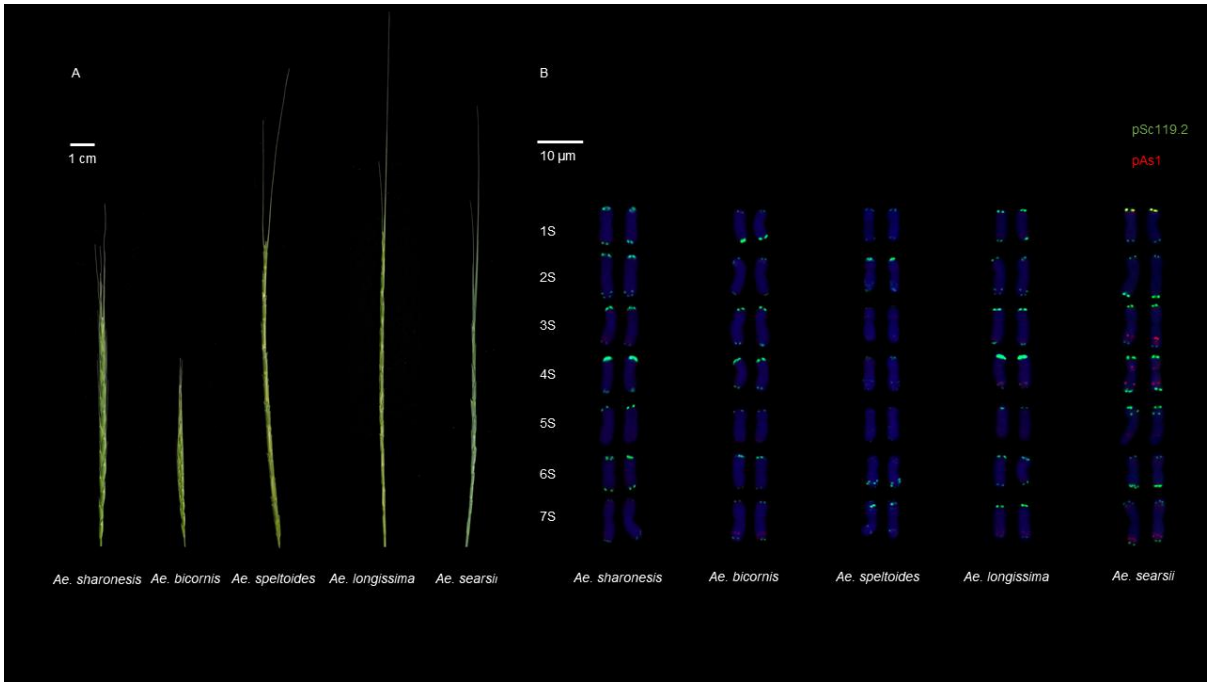
two species; (3) *Species*-specific gene family of species. Intersect analyses of the three gene categories were visualized by R package Upset (https://github.com/hms-dbmi/UpSetR/).

We also evaluated the gene family expansion and contraction histories of the *Triticum*/*Aegilops* species using a previously published phylogenomic approach (Appels *et al.*, 2018) with the same protein-coding gene set. Log-transformed gene family size of these *Triticum*/*Aegilops* species were compared using the phylANOVA function (with parameters "nsim = 1000, p.adj = 'fdr'") of phytools package (v0.7-90; https://github.com/liamrevell/phytools) in R. The *p*-value of ANOVA statistics was corrected using FDR method. Only these gene families that possess FDR < 0.1 were applied to infer the expansion and contraction history.

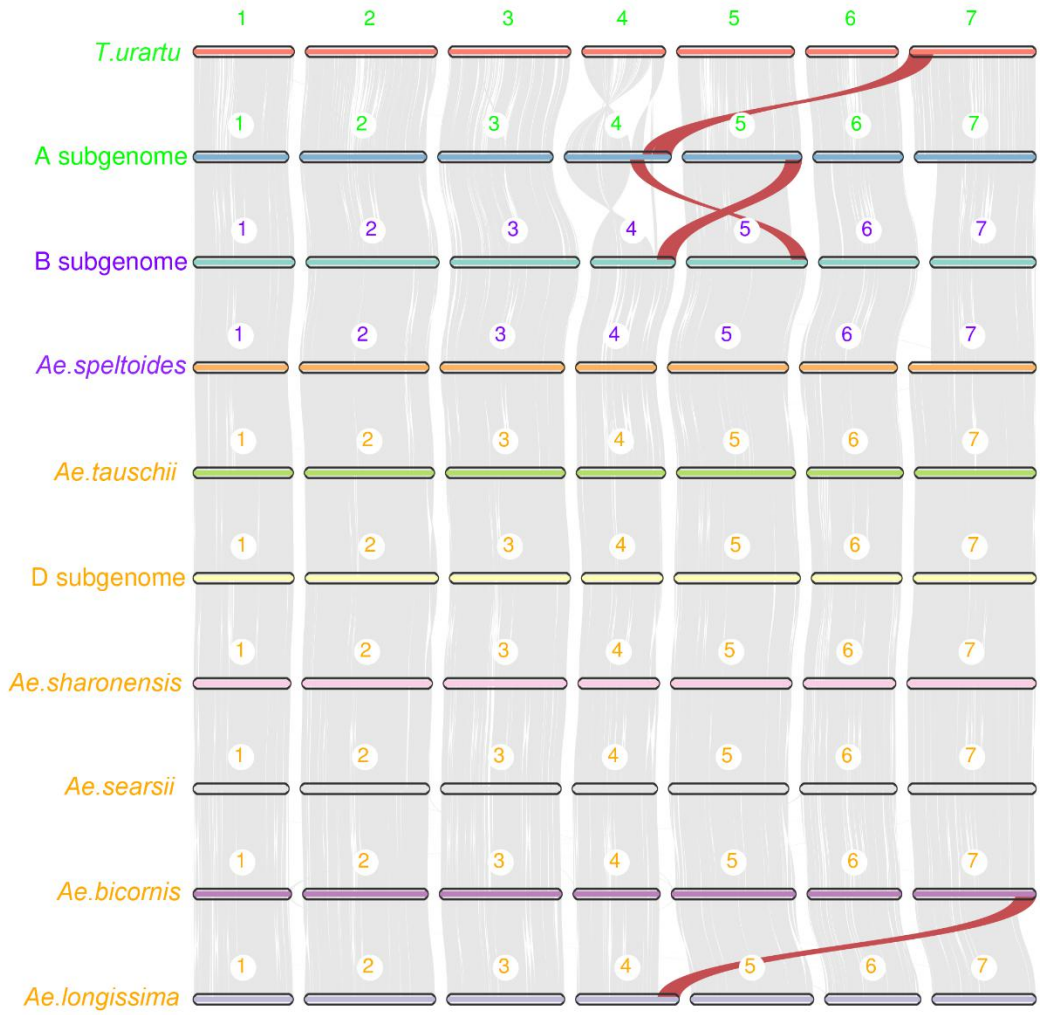
Annotation of NBS-LRR and agronomic/domestication trait-related genes

Annotation of NBS-LRR genes of the *Triticum*/*Aegilops* species were performed using NLR-Annotator pipeline (Zhang, 2020). The program PfamScan (https://www.ebi.ac.uk/Tools/pfa/pfamscan/) was used to search the candidate NBS-LRR genes against Pfam database (https://pfam.xfam.org/) to validate the integrity of NB_ARC domain (PF00931). All identified NBS-LRR genes were classified according to the position of intron located inside or outside of NB_ARC domain. In addition, we also employed the available coding sequences of agronomic/domestication trait-related gene (*i.e.*, *Q/q* gene) as query to search against to *Triticum*/*Aegilops* genomes and subgenomes using BLASTN with e-value < 10^{e-5} and hits > 40% identify and 60% coverage in length. All candidate hits were manually checked and compared with annotated protein sequences. For BLAST hits that absent in annotated protein sequences (i.e., *Btr* gene), DNA segments together with flanking regions of 1,000 bp were extracted and annotated using both ORFfinder (https://www.ncbi.nlm.nih.gov/orffinder/) and Augustus (http://bioinf.uni-greifswald.de/augustus/submission.php). Newly annotated genes were BLAST against original gene set for validation.

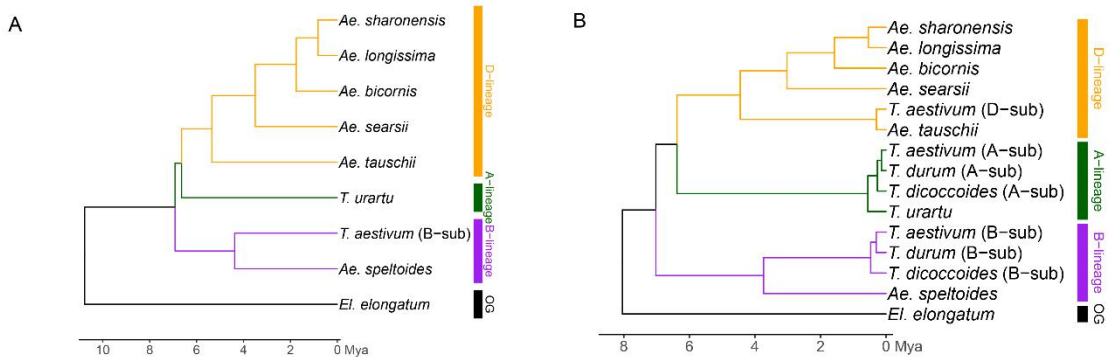
Supplementary Figures:



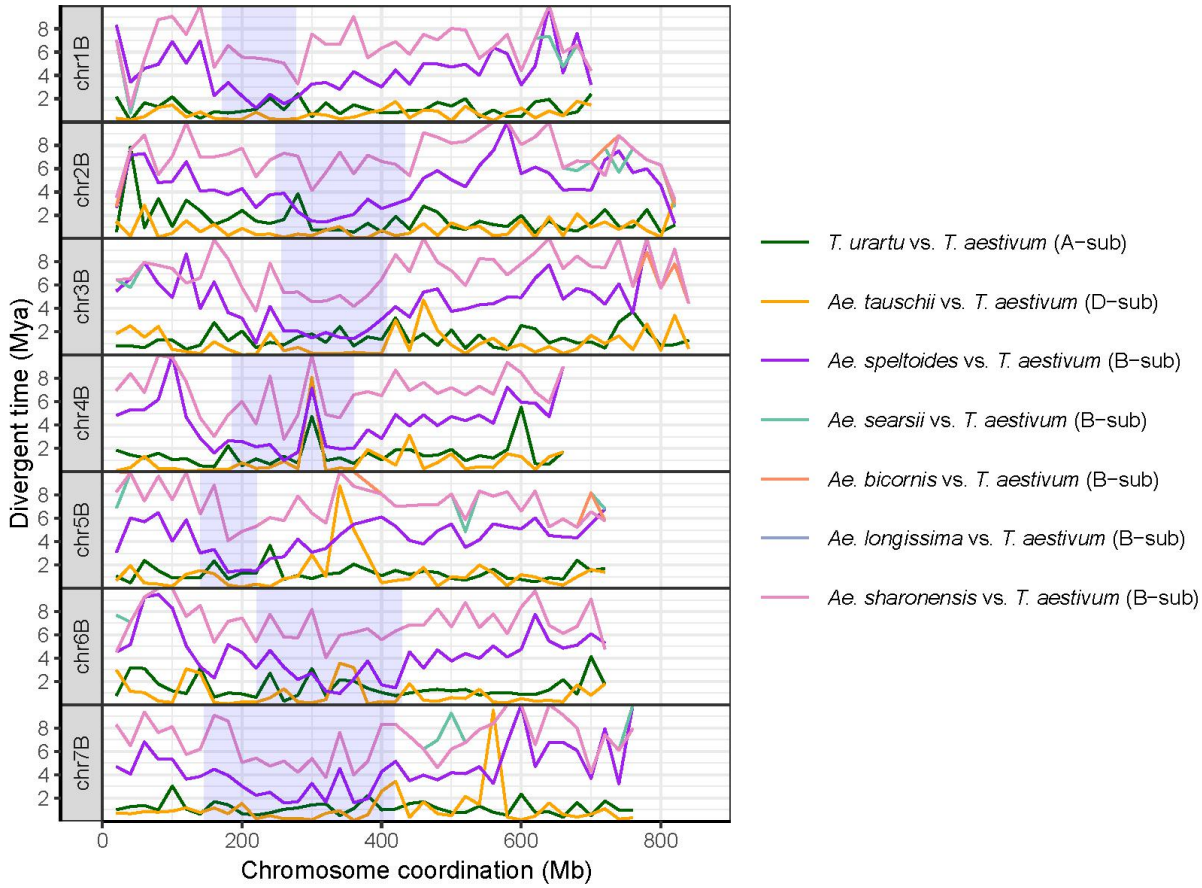
Supplementary Fig.1. Spike morphology (**A**) and fluorescence in situ hybridization (FISH) (**B**) of the five Sitopsis species. The same accessions of the five species were used to conduct *de novo* assembly.



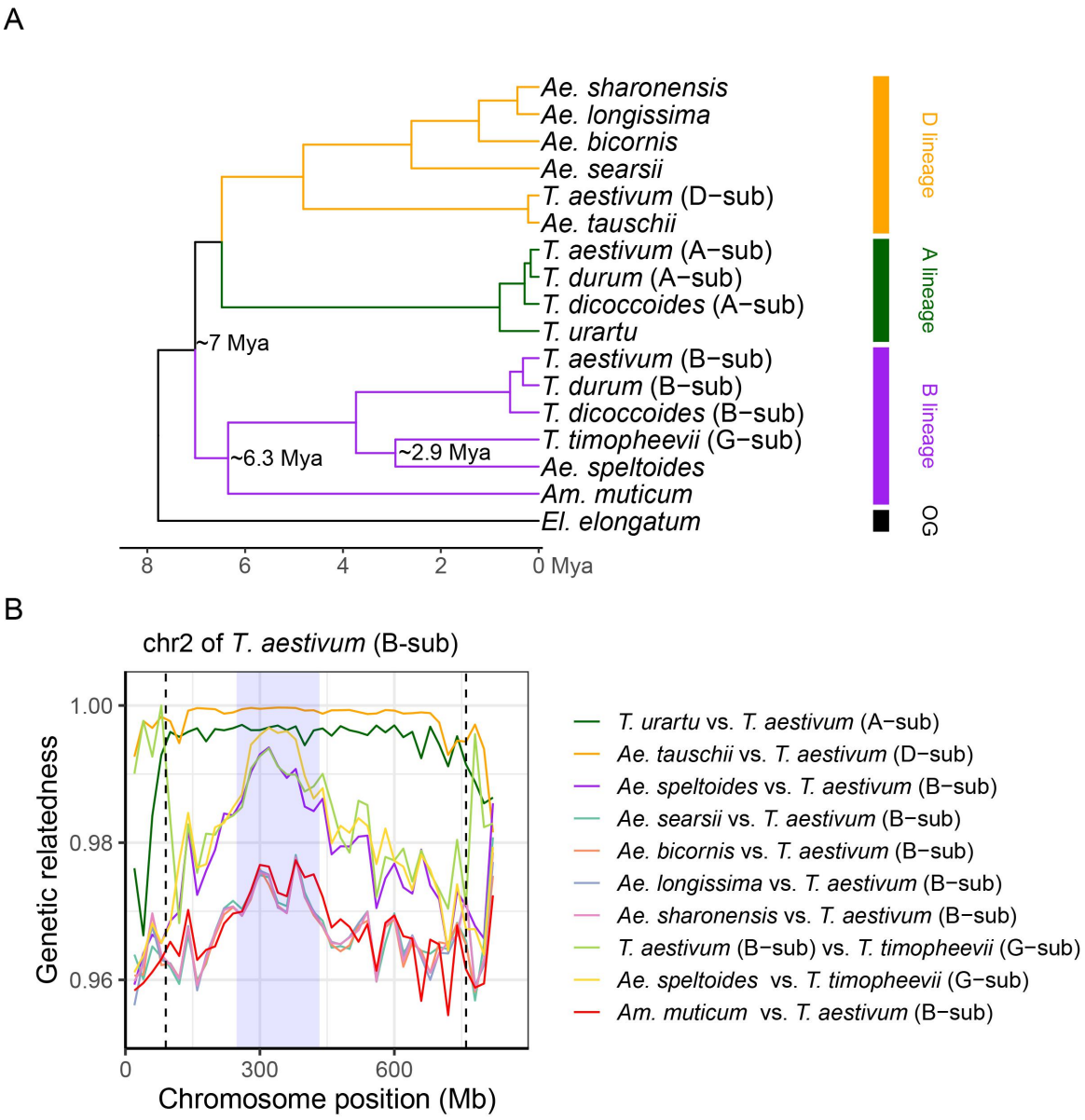
Supplementary Fig. 2. Genome collinearity analysis of the *Triticum*/*Aegilops* species complex. The two previously identified large genomic translocations (4A/5A/7B in bread wheat and $4S^{l}/7S^{l}$ in *Ae. longissima*) are highlighted by red color.



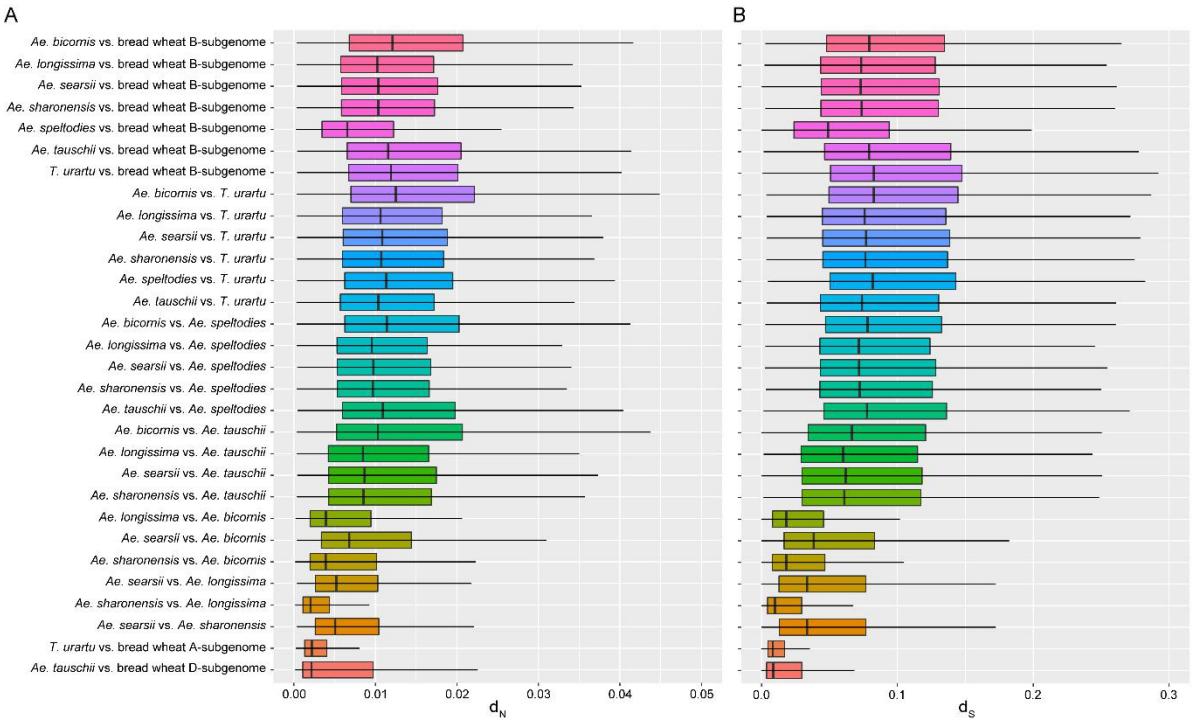
Supplementary Fig. 3. Maximum likelihood tree and divergence time of the *Triticum*/*Aegilops* species based on whole genome SNPs (A) and reduced representative genomic regions (B).



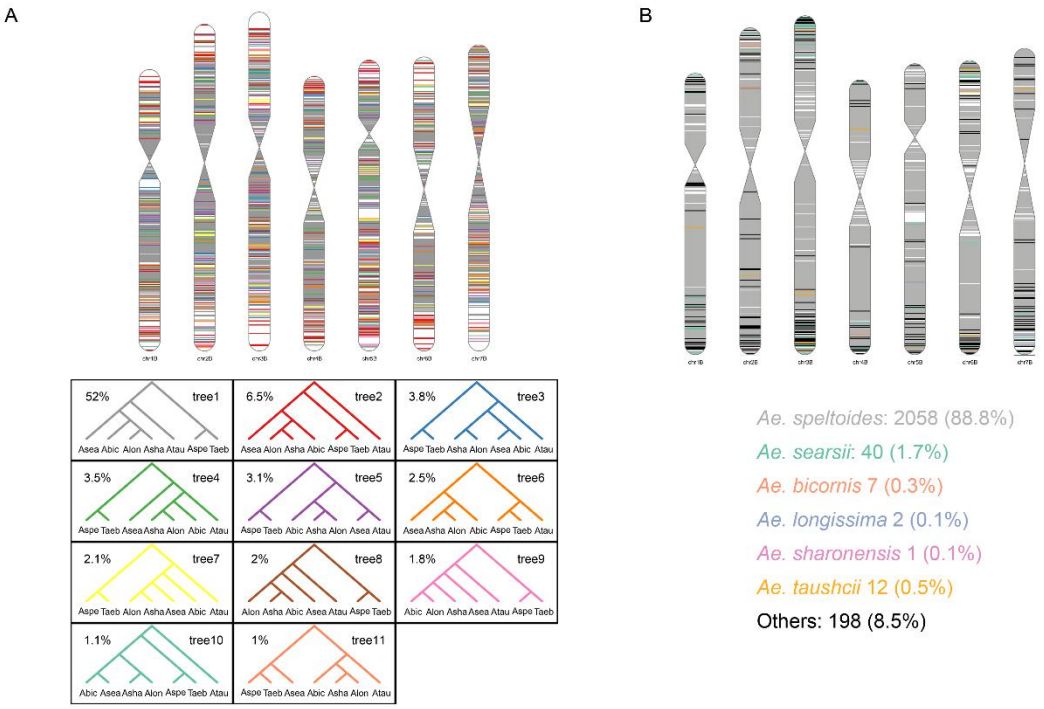
Supplementary Fig. 4. Divergence time between the seven diploid *Triticum*/*Aegilops* species and polyploid wheat B-subgenome along the seven chromosomes based on single copy orthologous genes. Note that four *Sitopsis* species (*Ae. bicornis*, *Ae. searsii*, *Ae. longissima* and *Ae. sharonensis*) show highly similar divergent pattern to the B-subgenome of bread wheat. These curves are overlapped at some genomic regions along the seven chromosomes.



Supplementary Fig. 5. **(A)** Phylogeny tree of the seven diploid *Triticum*/*Aegilops* species and bread wheat B-subgenome (cultivar Chinese Spring) and *Timopheevii* G-subgenome (an introgressed region on chromosome 2B of bread wheat cultivar "LongReach Lancer") based on the RRGRs on chromosome 2B of Chinese Spring. **(B)** Distribution of the genetic relatedness among genomes/subgenomes mentioned in **(A)** along chromosome 2B in Chinese Spring. The centromere of each chromosome is highlighted by purple color. Numbers on Y-axis are the values of genetic relatedness. Coordinates of chromosome is shown in the X-axis. Genomic region between the two dash lines represents the introgressed region from *T. timopheevii* to the bread wheat cultivar LongReach Lancer.

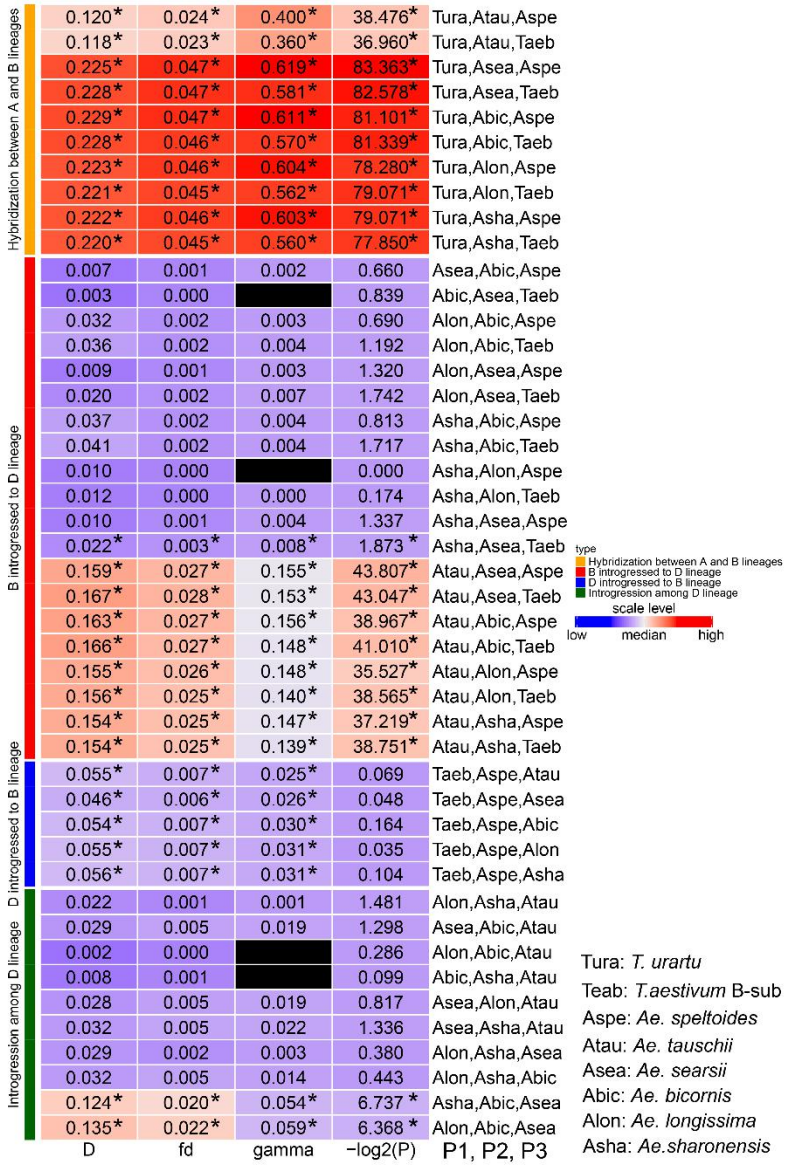


Supplementary Fig. 6. Pair-wise nonsynonymous (d_N) and synonymous (d_S) substitution rates of the diploid *Triticum*/*Aegilops* species and polyploid wheat subgenomes based on single copy orthologous genes.



Taeb: *T. aestivum* (B-sub); Aspe: *Ae. speltoides*; Asea: *Ae. searsii*; Abic: *Ae. bicornis*; Alon: *Ae. longissima*; Asha: *Ae. sharonensis*; Atau: *Ae.tauschii*

Supplementary Fig. 7. Distribution of the tree topologies of the *Triticum*/*Aegilops* species based on 2,295 reduced representative genomic regions. (**A**) Only the top 11 tree topologies were shown. Colors on the seven bread wheat B-subgenome chromosomes are the same as the tree topologies. (**B**) Distribution and percentages of the incongruent gene tree between the five Sitopsis species and their closely relatives. Colors on the seven bread wheat B-subgenome chromosomes are the same as the species names.

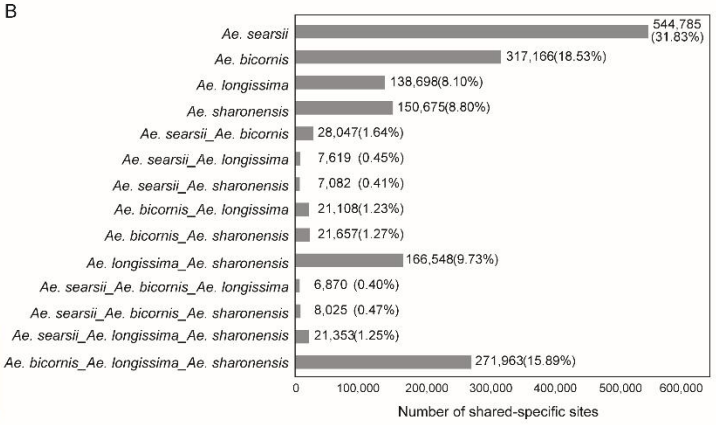


Supplementary Fig. 8. Estimates of the Patterson's *D*, *fd*, hybrid index (γ) values and *χ2* goodness of fit test of number tree topologies among the *Triticum*/*Aegilops* species complex based on the representative genomic regions. At whole genome level, *D* and *fd* statistics are used to evaluated the strengthen of the gene flow from P3 to P2 (*Elymus elongatum* as outgroup), and γ quantifies the proportion of P3 contributed to P2. *χ2* goodness of fit test is used for introgression detection based on the counts of three different slide-window trees. The ((P1, P2), P3) represents the topology of species tree, and ((P2, P3), P1) and ((P1, P3), P2) represent discordant trees in each triplet. For *D*, *fd* and γ, significant introgression signals based 1,000 bootstraps (*p*-value < 0.001) are marked as asterisks. For *χ2* goodness of fit test, the ratio between two discordant trees that is significantly deviated from 1 (*p*-value < 0.001) indicates an introgression event (marked with asterisks). The black cells represent meaningless γ values (less than 0 or larger than 1).

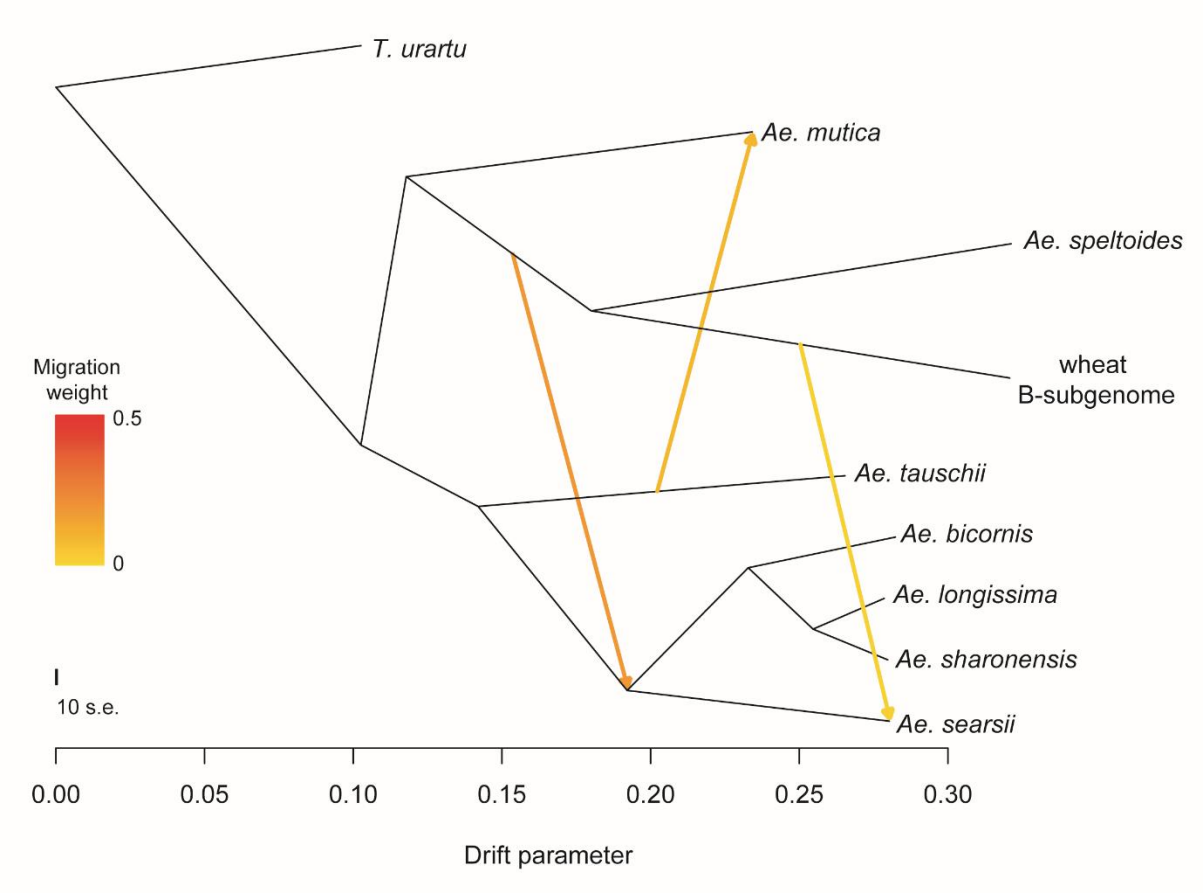
A

		T. urartu (A)	wheat B-subgenome (B)	*Ae. speltoides* (Aspe)	Ancestor of B and Aspe	*Ae. tauschii*
Specific sites shared with n *Sitopsis* species	4	44,580	20,749	20,665	61,104	232,995
	3	25,165	18,271	17,940	25,451	55,624
	2	13,565	14,992	12,957	14,505	19,513
	1	45,960	47,908	36,694	40,150	62,771
	Total	129,270	101,920	88,256	141,210	370,903
	Pro.	1.31%	1.04%	0.90%	1.44%	3.77%
Specific sites shared with *Ae. tauschii*	Num	87,963	25,380	25,788	29,235	N
	Pro.	0.89%	0.26%	0.26%	0.30%	N

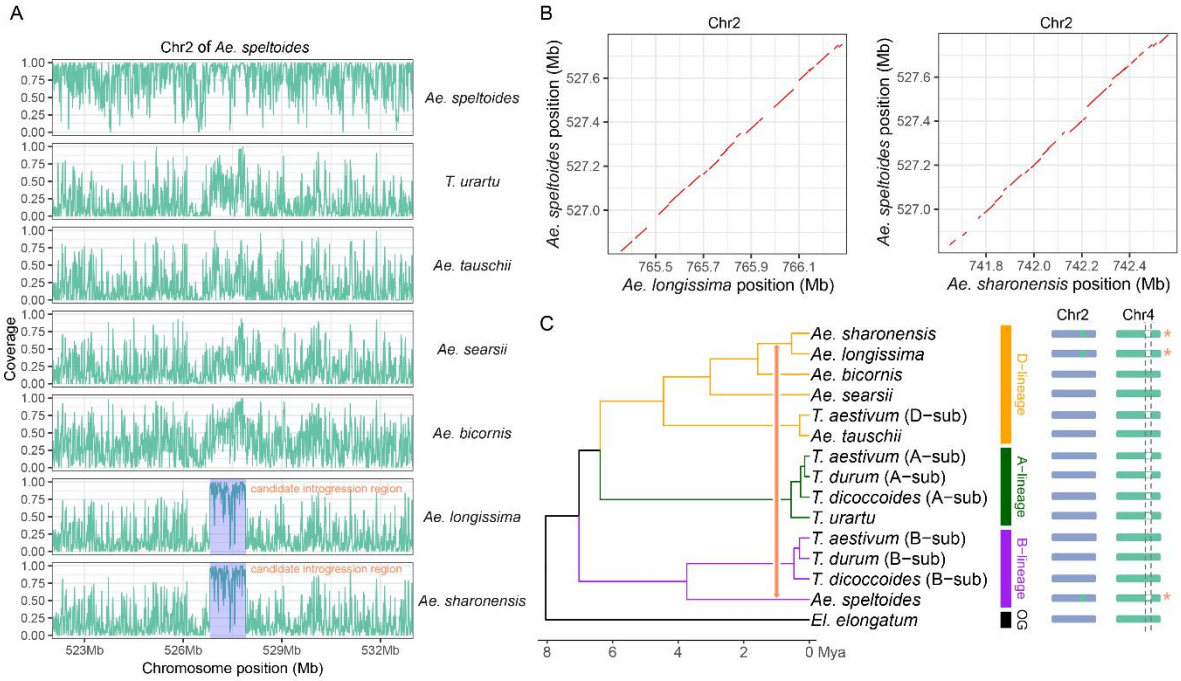
B



Supplementary Fig. 9. Percentages of the putative introgressed sites (**A**) and shared species-specific SNPs (**B**) among the *Triticum*/*Aegilops* species. Ancestor of B and Aspe indicates the most recent common ancestor of the bread wheat B-subgenome and *Ae. speltoides*.

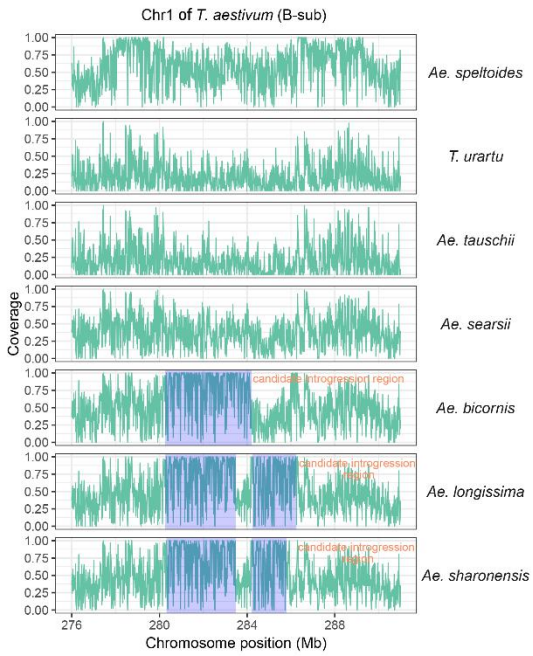


Supplementary Fig. 10. Allele frequency-based migration inference of the seven diploid *Triticum*/*Aegilops* species and bread wheat B-subgenome using Treemix based on whole genome SNPs. Arrows with color from orange to red indicate the high and low possibility migration events. Only the top three migration events are shown. Bar on the bottom represents the branch length of maximum likelihood tree.



Supplementary Fig. 11. Putative introgressed genomic regions between the *Ae. speltoides* and the D-lineage Sitopsis species based on the inter-specific sequence similarity through whole-genome alignment and mapping coverage of Illumina short reads. (**A**) Putative introgressed genomic region identified between both the *Ae. longissima* and *Ae. sharonensis* and *Ae. speltoides*. This genomic region was identified through genome-wide inter-specific sequence alignment. X indicates the coordinate on *Ae. speltoides* chromosome 2. Y axis represents the total reads of the seven diploid species mapped onto the *Ae. speltoides* reference genome. Only the *Ae. longissima* and *Ae. sharonensis* show high read coverage at this genomic region. (**B**) Dotplots indicate the high collinear genomic region on chromosome 2 between *Ae. speltoides* and each of the *Ae. longissima* and *Ae. sharonensis*. (**C**) The genomic region on chromosome 2 (green color) was translocated from chromosome 4 (white color). The same genomic region shared in the *Triticum*/*Aegilops* species is most likely through inter-specific genetic introgression.

A

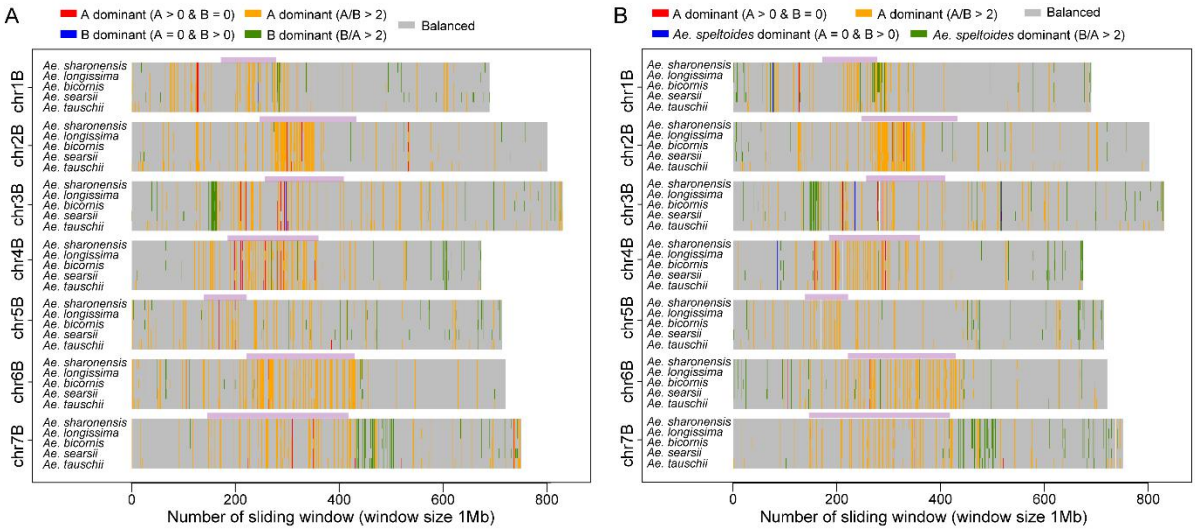


B

Introgression events among Sitopsis species
(0. not occurred; 1. occurred)

Event id	*Ae. searsii*	*Ae. bicornis*	*Ae. longissima*	*Ae. sharonensis*
1	0	1	1	1
2	1	0	0	0
3	1	1	0	0
4	1	0	0	0
5	1	1	1	1
6	1	1	1	1
7	1	1	0	0
8	1	1	1	1
9	1	0	0	0

Supplementary Fig. 12. Putative introgressed genomic regions between bread wheat B-subgenome and the four D-lineage Sitopsis species based on the inter-specific sequence similarity through whole-genome alignment and mapping coverage of Illumina short reads. On the left, X indicates the coordinate on B-subgenome chromosome 1. Y axis represents the total reads of the seven diploid species mapped onto the B-subgenome reference genome. On the right, the red number 1 (highlighted by black color) represents the putative introgression event between bread wheat B-subgenome and the four D-lineage Sitopsis species.

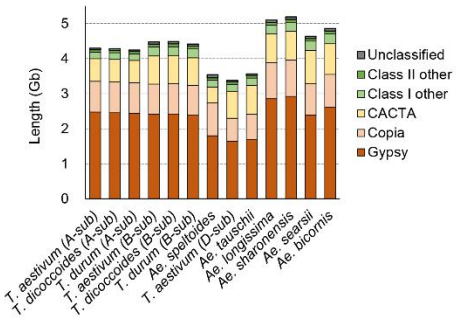


Supplementary Fig. 13. Distribution patterns of the A- and B-lineage specific SNPs in the five D-lineage species. The A-dominant genomic regions are defined as the total number of A-lineage specific SNPs are >2 fold higher than B-lineage specific SNPs (or no B-lineage specific SNPs). The reverse pattern is defined as B-dominant genomic region. The remaining genomic regions that are neither A- nor B-lineage dominant are defined as balanced genomic regions.

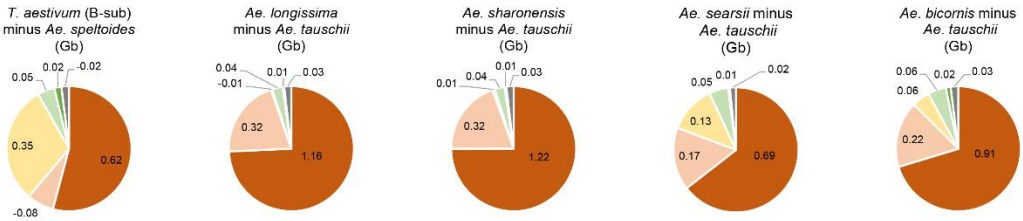
A

Species name	Length (Gb)							
	Gypsy	*Copia*	*CACTA*	Class I other	Class II other	Unclassified	All repeats	Genome size
T. aestivum (A-sub)	2.48	0.88	0.64	0.19	0.07	0.04	4.30	4.93
T. dicoccoides (A-sub)	2.46	0.88	0.64	0.19	0.07	0.04	4.28	4.90
T. durum (A-sub)	2.44	0.87	0.64	0.19	0.07	0.04	4.25	4.86
T. aestivum (B-sub)	2.41	0.86	0.80	0.25	0.08	0.07	4.48	5.18
T. dicoccoides (B-sub)	2.42	0.86	0.81	0.25	0.08	0.06	4.49	5.18
T. durum (B-sub)	2.39	0.85	0.79	0.25	0.08	0.06	4.42	5.11
Ae. speltoides	1.80	0.94	0.45	0.20	0.06	0.09	3.54	4.11
T. aestivum (D-sub)	1.64	0.66	0.77	0.20	0.07	0.04	3.38	3.95
Ae. tauschii	1.70	0.71	0.82	0.21	0.07	0.05	3.56	4.22
Ae. longissima	2.86	1.03	0.81	0.26	0.08	0.07	5.11	5.80
Ae. sharonensis	2.92	1.03	0.83	0.26	0.08	0.08	5.19	5.89
Ae. searsii	2.39	0.89	0.95	0.27	0.08	0.06	4.64	5.34
Ae. bicornis	2.61	0.94	0.88	0.27	0.09	0.07	4.86	5.64

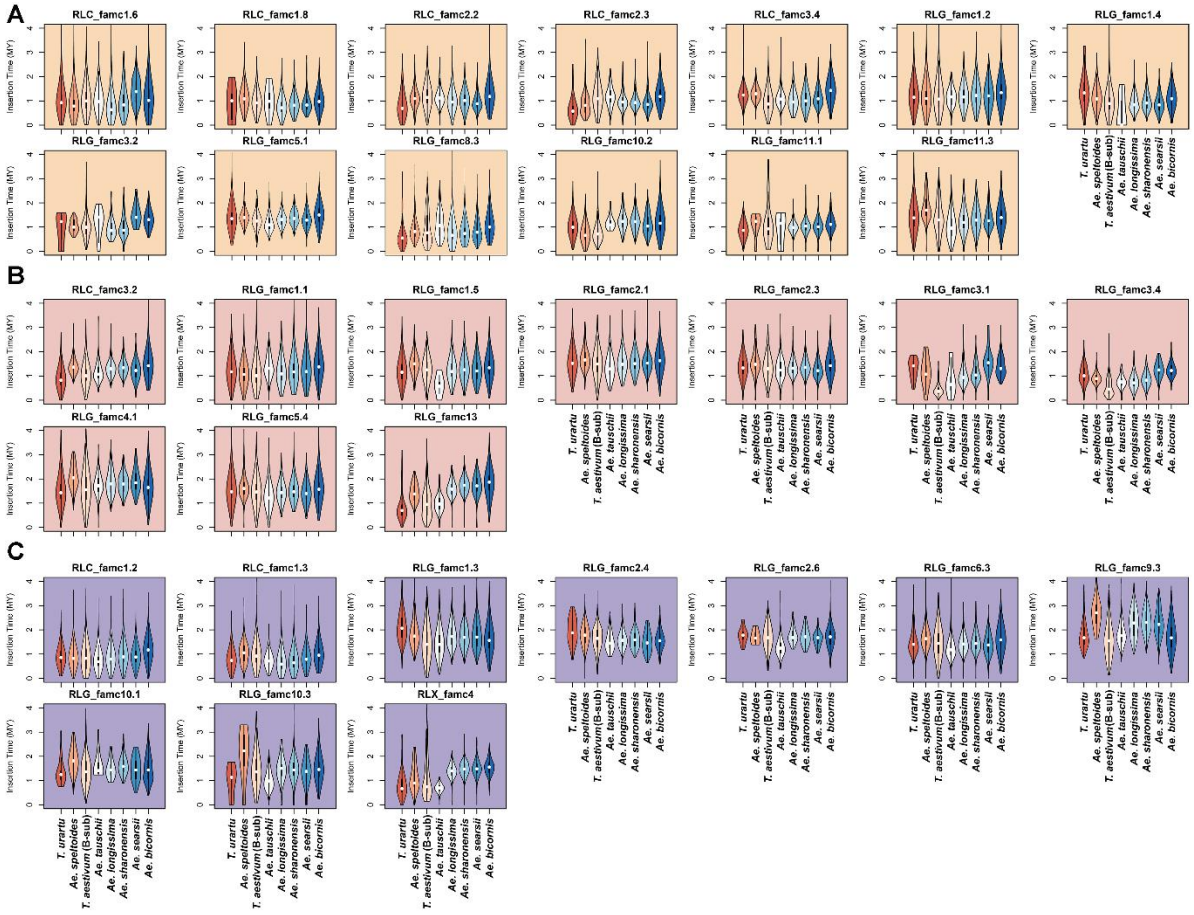
B



C



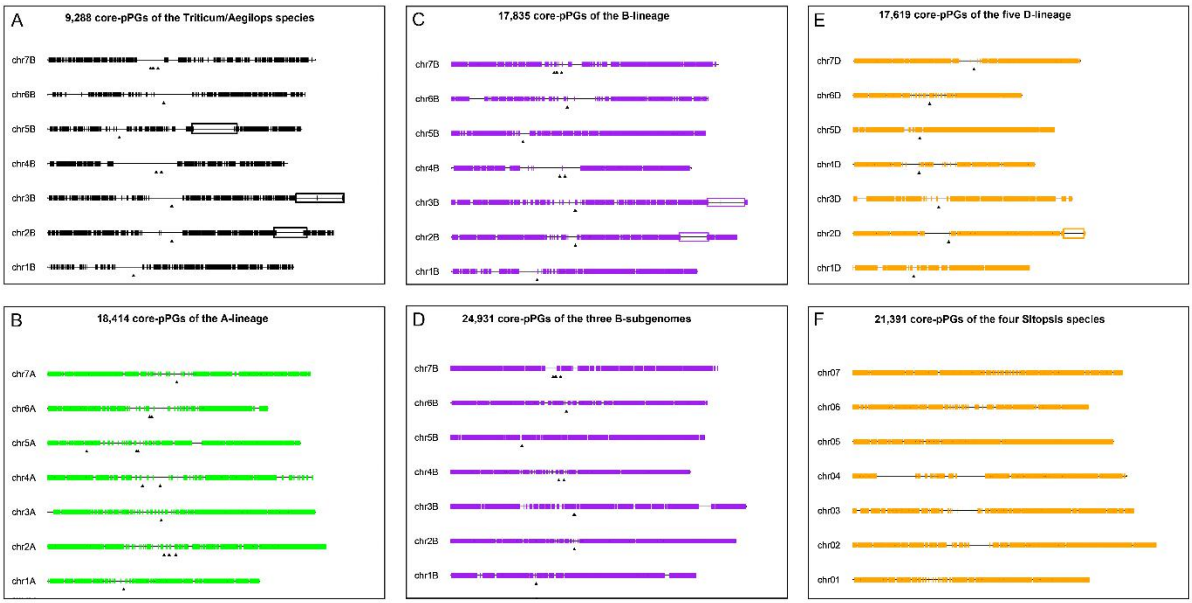
Supplementary Fig. 14. Genome-wide feature of the repetitive sequence in the diploid species and polyploid wheat subgenomes. (**A**-**B**) Total length of each type of the repetitive sequence. (**C**) Difference in total length of the repetitive sequence in the B-lineage (between *Ae. speltoides* and bread wheat B-subgenome) and D-lineage (between *Ae. tauschii* and the four D-lineage Sitopsis species).



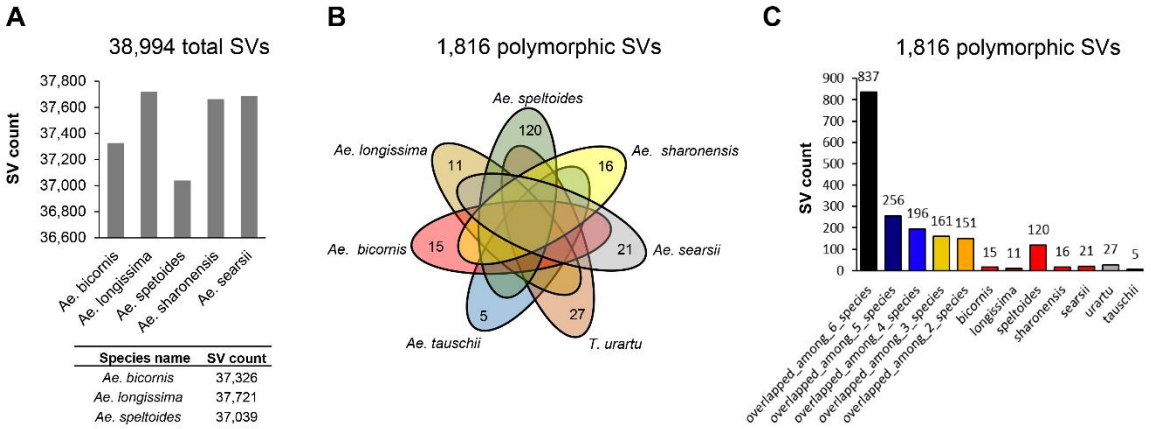
316

317 **Supplementary Fig. 15**. Insertion times of the transposon subfamilies that amplified specifically in D-

318 (orange) and B-lineages (purple) (**A** and **C**) and shared between the two lineages (red) (**B**).

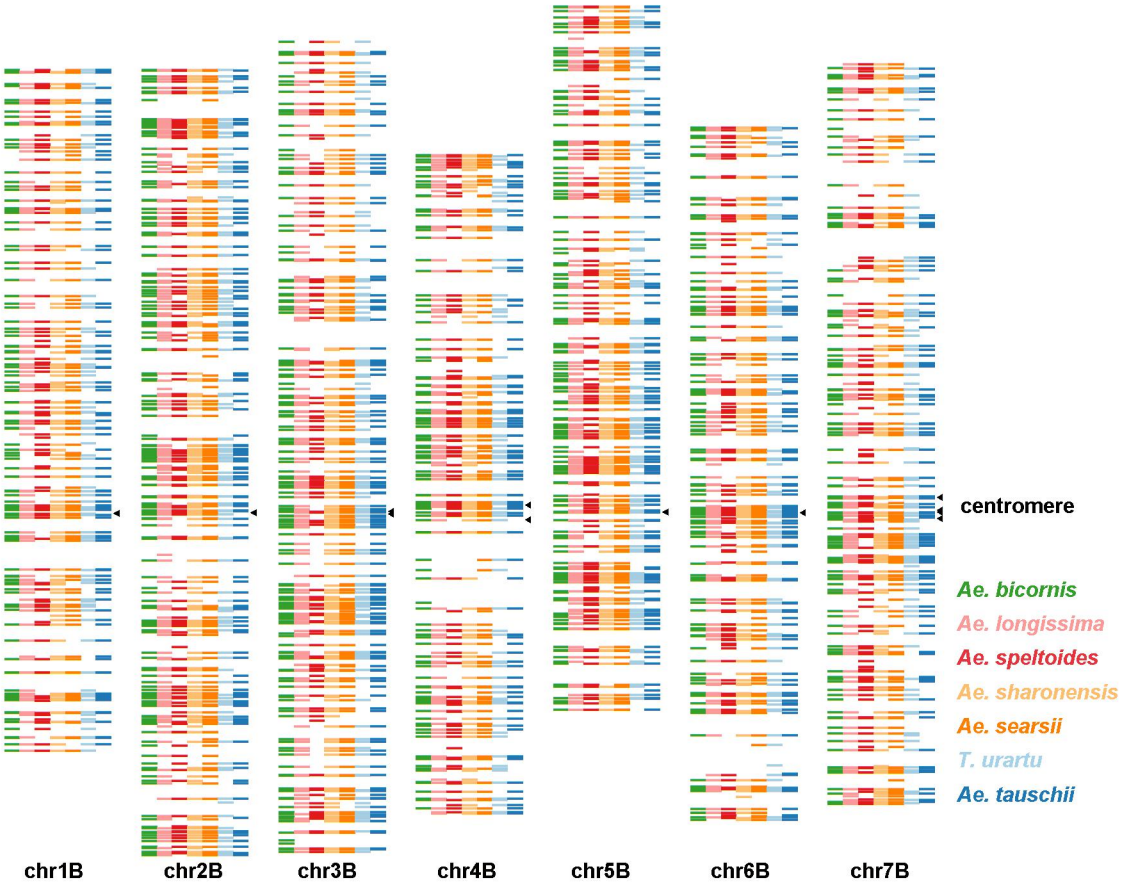


Supplementary Fig. 16. Distribution patterns of the core putative proto-gene (pPG) in all the *Triticum*/*Aegilops* species complex (black) (**A**), A-lineage (green) (**B**), B-lineage (purple) (**C**), B-subgenome (purple) (**D**), D-lineage (orange) (**E**), and D-lineage Sitopsis species (orange) (**F**). Identification of the core pPG in the D-lineage Sitopsis species were defined based on the *Ae. longissima* reference genome, all the other core pPG were defined based on the three subgenomes of bread wheat. Non-conserved genomic regions in the B-, D- and all *Triticum*/*Aegilops* species are shown in purple, orange and black boxes. The black triangles indicate the centromeric regions of the bread wheat A-, B- and D-subgenomes.

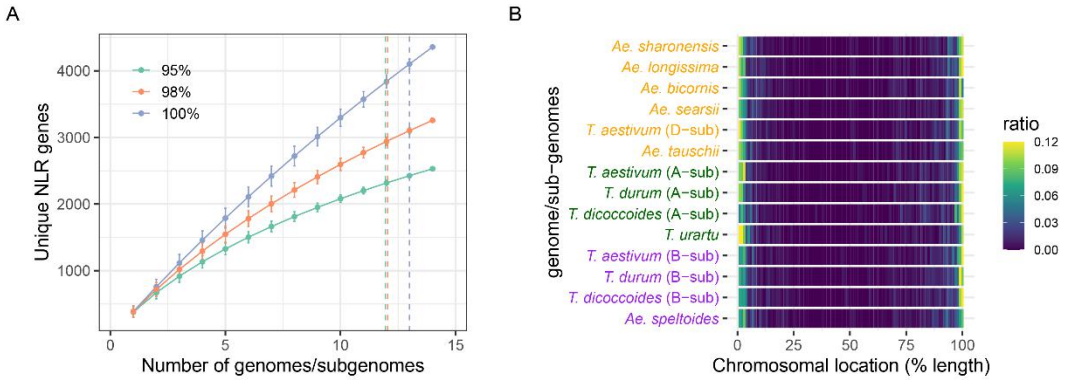


Species name	SV count
Ae. bicornis	37,326
Ae. longissima	37,721
Ae. speltoides	37,039
Ae. sharonensis	37,663
Ae. searsii	37,689
T. urartu	21,272
Ae. tauschii	22,789

Supplementary Fig. 17. Pan-genomic analyses of the genome structural variations (SVs) of the seven diploid *Triticum*/*Aegilops* species. (**A**) Total numbers of SVs identified in the seven diploid species. (**B**-**C**) Intersection analysis of the 18,16 polymorphic SVs among the seven diploid species.



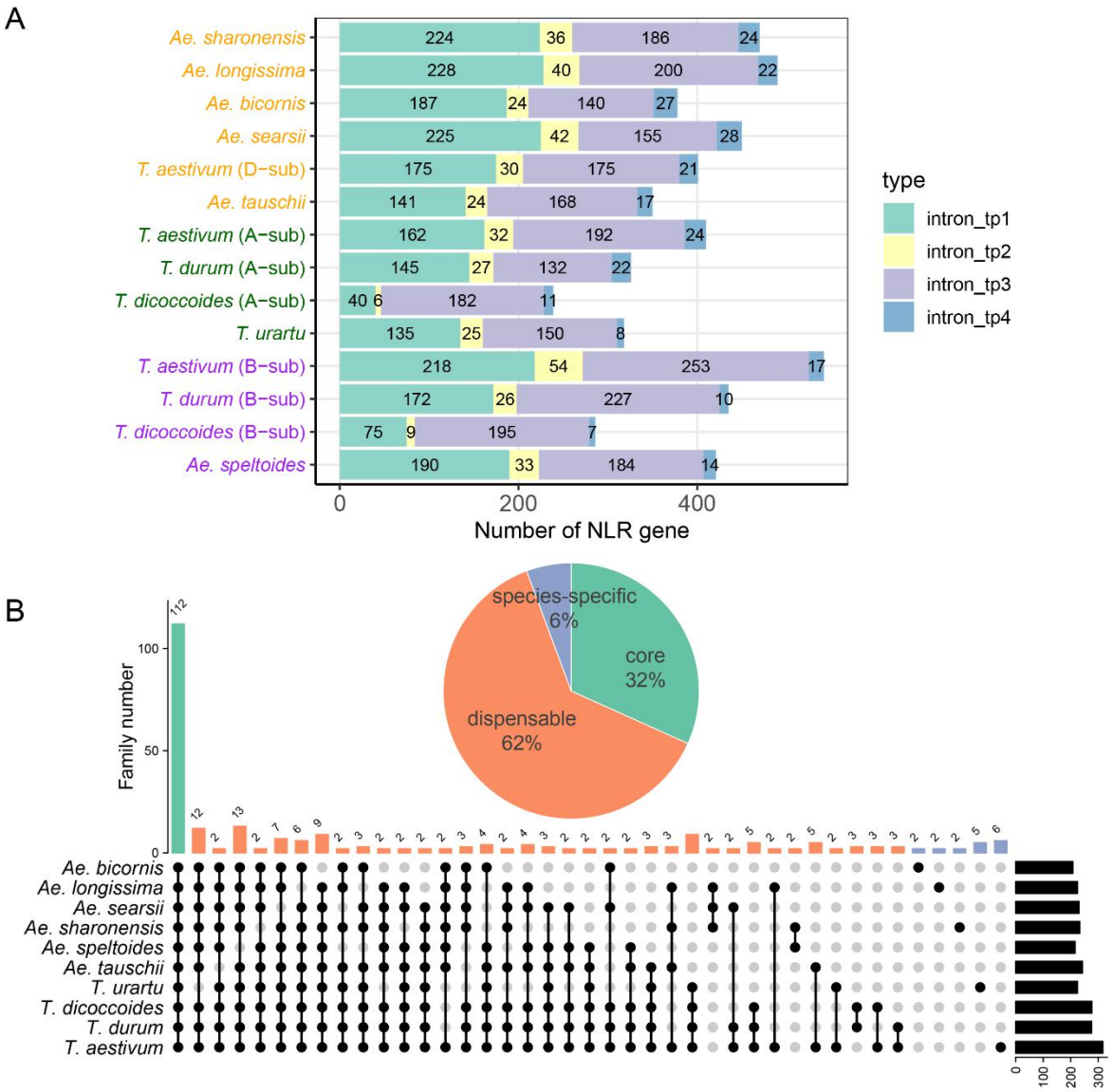
Supplementary Fig. 18. Genome-wide distribution of the 1,816 shared polymorphic SVs in the seven diploid species on the seven chromosomes.



336

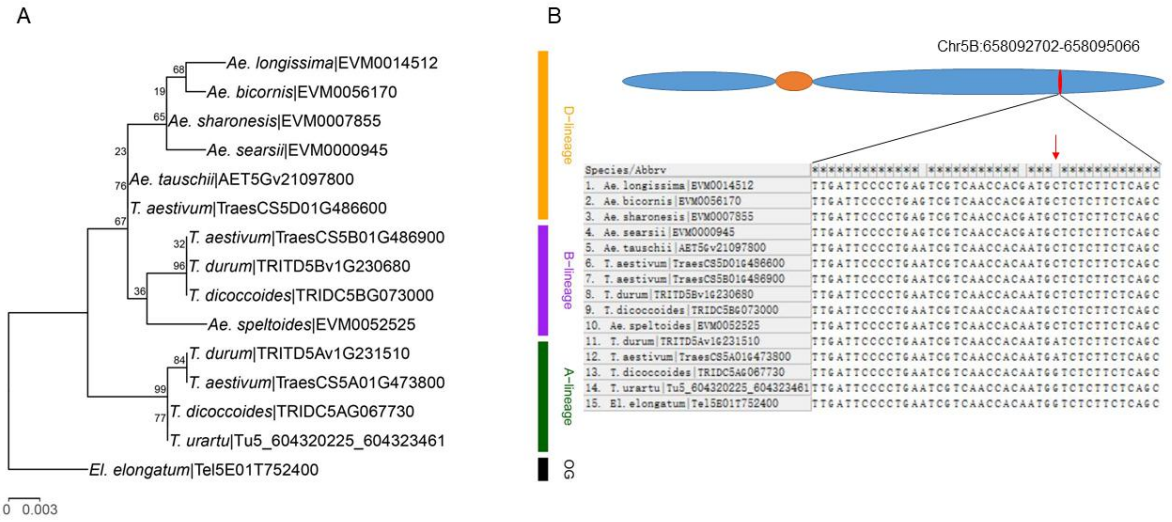
337 **Supplementary Fig. 19**. Identification of nucleotide-binding and leucine-rich repeat (NBS-LRR) gene
338 in the *Triticum*/*Aegilops* species. (**A**) Numbers of unique NBS-LRR genes identified in the
339 *Triticum*/*Aegilops* species based on 95%, 98% and 100% genetic identity. X and Y-axes indicate the
340 numbers of genomes and NBS-LRR genes. (**B**) Genome-wide distribution of the of the NBS-LRR genes
341 in the *Triticum*/*Aegilops* species.

342

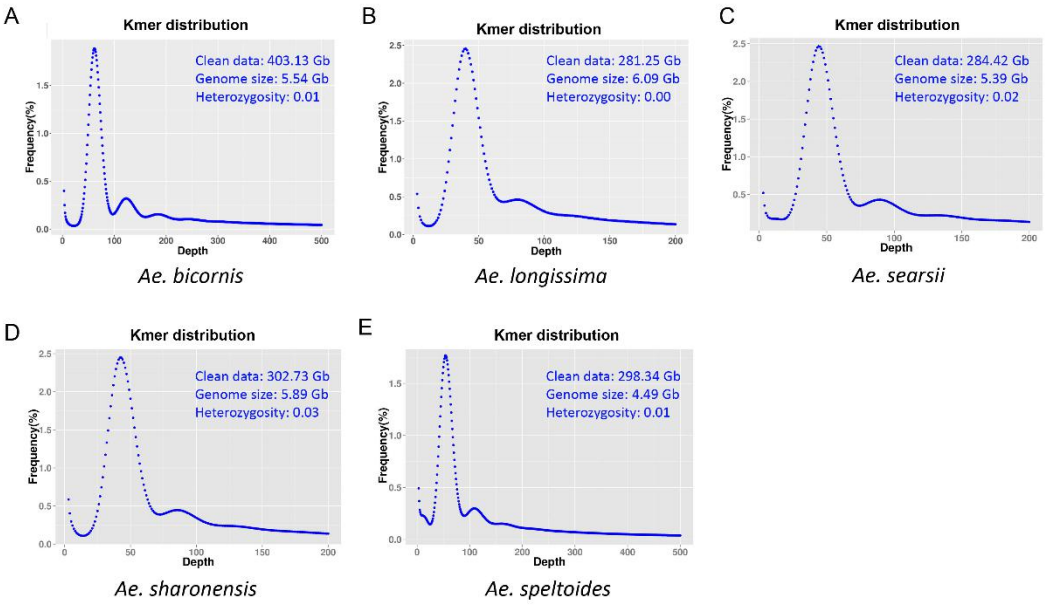


343

344 **Supplementary Fig. 20**. (**A**) Total NBS-LRR genes identified in the diploid species and polyploid wheat
345 subgenomes. Tuiquoise, yellow, orchid and blue colors indicate the different types of NBS-LRR genes.
346 (**B**) Intersection analysis of the NBS-LRR genes in the *Triticum*/*Aegilops* species.

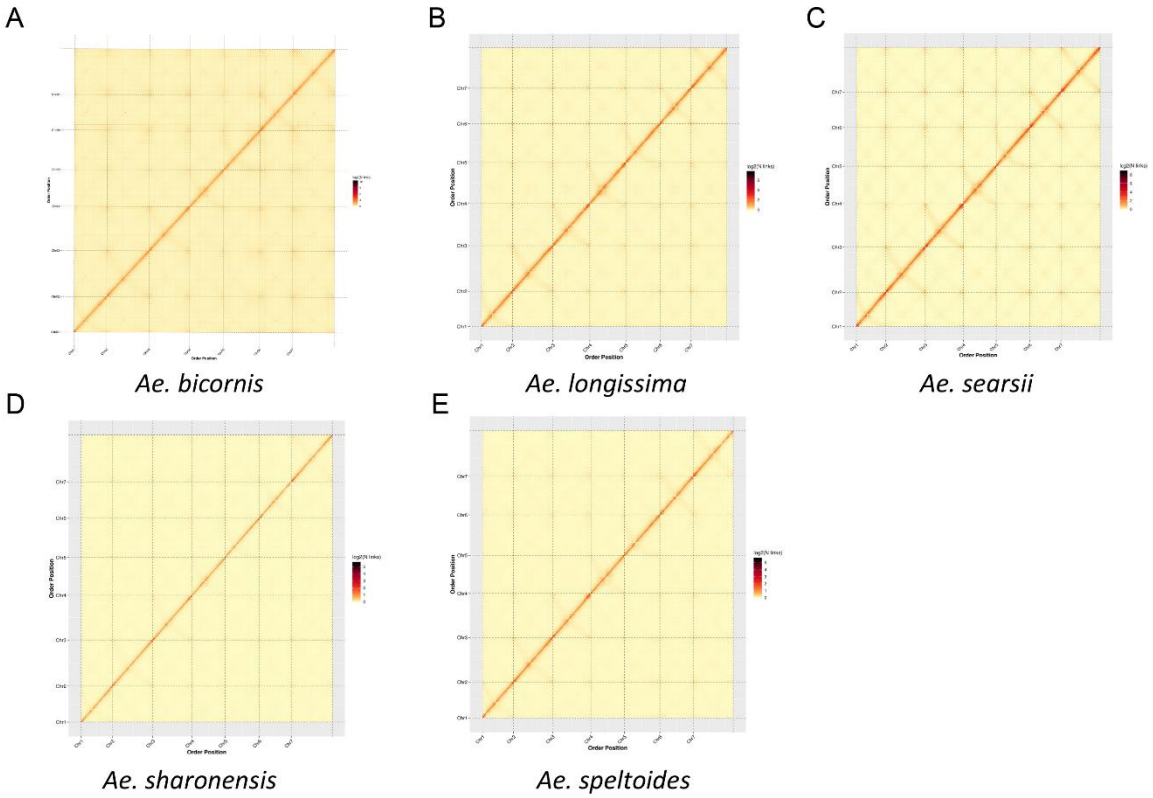


Supplementary Fig. 21. Phylogenetic tree (**A**) and domestication allele (**B**) of the *Q*/*q* gene in in the *Triticum*/*Aegilops* species. The red arrow indicates the physical position of the key mutation of *Q* (I_{329}) and *q* (L_{329} and V_{329}) alleles. The *Q*/*q* gene was not annotated in the public data of *T. urartu (*Ling *et al*., 2017*)*. We then reannotated an unknown protein that shows high sequence similarity to the *Q*/*q* gene of the other relative diploid and polyploid wheat species (see details in **Supplementary Notes**).



353

354 **Supplementary Fig. 22**. Genome survey of the five Sitopsis species based on Illumina short reads.



355

356 **Supplementary Fig. 23**. Genome-wide analysis of chromatin interactions at 500-kb resolution of the

357 five Sitopsis species.

358

Appels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B., Rogers, J., Pozniak, C.J., Choulet, F., Distelfeld, A., and Poland, J. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science **361**.

Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S., Gundlach, H., Hale, I., Mascher, M., Spannagl, M., and Wiebe, K. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science **357**:93-97.

Blanco, E., Parra, G., and Guigó, R. (2007). Using geneid to identify genes. Current Protocols in Bioinformatics **18**:4.3. 1-4.3. 28.

Blischak, P.D., Chifman, J., Wolfe, A.D., and Kubatko, L. (2018). HyDe: a Python package for genome-scale hybridization detection. Systematic Biology **67**:821-829.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. Journal of Molecular Biology **268**:78-94.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nature Biotechnology **31**:1119-1125.

Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M., and Buell, C.R. (2006). Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC Genomics **7**:1-17.

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Molecular Biology and Evolution **17**:540-552.

Emms, D., and Kelly, S.L. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology **20**:238.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biology **9**:1-22.

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., and Quesneville, H. (2014). PASTEC: an automatic transposable element classification tool. PloS One **9**:e91929.

Keilwagen, J., Wenk, M., Erickson, J.L., Schattat, M.H., Grau, J., and Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. Nucleic Acids Research **44**:e89-e89.

Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology **37**:907-915.

Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., and Rao, B.S. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biology **5**:1-28.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research **27**:722-736.

Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics **5**:1-9.

Kurtz, S., Phillippy, A.M., Delcher, A.L., Smoot, M.E., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome Biology **5**:1-9.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34**:3094-3100.

Ling, H., Ma, B., Shi, X., Liu, H., Dong, L., Sun, H., Cao, Y., Gao, Q., Zheng, S., and Li, Y. (2018). Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. Nature **557**:424-428.

Luo, M., Gu, Y.Q., Puiu, D., Wang, H., Twardziok, S., Deal, K.R., Huo, N., Zhu, T., Wang, L., and Wang, Y. (2017). Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. Nature **551**:498-502.

Maccaferri, M., Harris, N.S., Twardziok, S., Pasam, R.K., Gundlach, H., Spannagl, M., Ormanbekova, D., Lux, T., Prade, V.M., and Milner, S.G. (2019). Durum wheat genome highlights past domestication signatures and future improvement targets. Nature Genetics **51**:885-895.

Majoros, W.H., Pertea, M., and Salzberg, S.L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics **20**:2878-2879.

Marcussen, T., Sandve, S.R., Heier, L., Spannagl, M., Pfeifer, M., Jakobsen, K.S., Wulff, B.B.H., Steuernagel, B., Mayer, K.F.X., and Olsen, O. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. Science **345**:1250092-1250092.

Martin, S.H., Davey, J.W., and Jiggins, C.D. (2015). Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. Molecular Biology and Evolution **32**:244-257.

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S., Wicker, T., Radchuk, V., Dockter, C., Hedley, P.E., and Russell, J. (2017). A chromosome conformation capture ordered sequence of the barley genome. Nature **544**:427-433.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics **20**:289-290.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23**:1061-1067.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology **33**:290-295.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics **21**:i351-i358.

Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nature Methods **17**:155-158.

Sela, I., Ashkenazy, H., Katoh, K., and Pupko, T. (2015). GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Research **43**.

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

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**:1312-1313.

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics **19**:ii215-ii225.

Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evolution **4**.

Sukumaran, J., and Holder, M.T. (2010). DendroPy: a Python library for phylogenetic computing. Bioinformatics **26**:1569-1571.

Suvorov, A., Scornavacca, C., Fujimoto, M.S., Bodily, P., Clement, M., Crandall, K.A., Whiting, M.F., Schrider, D.R., and Bybee, S.M. (2020). Deep ancestral introgression shapes evolutionary history of dragonflies and damselflies. bioRxiv.

Tang, S., Lomsadze, A., and Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. Nucleic Acids Research **43**:e78-e78.

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome Research **27**:737-746.

Vogel, J.P., Garvin, D.F., Mockler, T.C., Schmutz, J., Rokhsar, D.S., Bevan, M., Barry, K., Lucas, S., Harmonsmith, M., and Lail, K. (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature **463**:763-768.

Vurture, Gregory, W., Sedlazeck, Fritz, J., Nattestad, Maria, Underwood, Charles, J., Han, F., and Gurtowski (2017). GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., and Young, S.K. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One **9**:e112963.

Wang, H., Sun, S., Ge, W., Zhao, L., Hou, B., Wang, K., Lyu, Z., Chen, L., Xu, S., Guo, J., et al. (2020). Horizontal gene transfer of *Fhb7* from fungus underlies *Fusarium* head blight resistance in wheat. Science **368**:eaba5435. 10.1126/science.aba5435.

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Research **35**:W265-W268. 10.1093/nar/gkm286.

Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.Y. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution **8**:28-36.

Zhang, W. (2020). NLR-Annotator: A tool for de novo annotation of intracellular immune receptor repertoire. Plant physiology **183**:418-420.