Reference genome assemblies reveal the origin and evolution of allohexaploid oat

Yuanying Peng (yy.peng@hotmail.com)
Sichuan Agricultural University https://orcid.org/0000-0002-3304-3301

Honghai Yan
Sichuan Agricultural University

Laichun Guo
Baicheng Academy of Agricultural Sciences

Cao Deng
DNA Stories Bioinformatics Center https://orcid.org/0000-0001-9823-9121

Lipeng Kang
Institute of Genetics and Developmental Biology https://orcid.org/0000-0003-0055-9149

Chunlong Wang
Baicheng Academy of Agricultural Sciences

Pingping Zhou
Sichuan Agricultural University

Kaiquan Yu
Sichuan Agricultural University

Xiaolong Dong
Sichuan Agricultural University

Jun Zhao
Sichuan Agricultural University

Yun Peng
Sichuan Agricultural University

Xiaomeng Liu
Sichuan Agricultural University

Di Deng
Sichuan Agricultural University

Yinghong Xu
Sichuan Agricultural University

Ying Li
Sichuan Agricultural University

Qiantao Jiang
Sichuan Agricultural University

Yan Li

Rice Research Institution https://orcid.org/0000-0002-6443-9245
Liming Wei
Baicheng Academy of Agricultural Sciences
Jirui Wang
Sichuan Agricultural University
Jian Ma
Sichuan Agricultural University
Ming Hao
Sichuan Agricultural University
Wei Li
Sichuan Agricultural University
Houyang Kang
Triticeae Research Institute, Sichuan Agricultural University
Youliang Zheng
Sichuan Agricultural University
Yuming Wei
Sichuan Agricultural University
Fei Lu
Institute of Genetics and Developmental Biology https://orcid.org/0000-0002-3596-1712
Changzhong Ren
Baicheng Academy of Agricultural Sciences

Biological Sciences - Article

Keywords: common oat, avena sativa, plant evolution, genome sequencing

Posted Date: June 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-664692/v1

Version of Record: A version of this preprint was published at Nature Genetics on July 18th, 2022. See the published version at https://doi.org/10.1038/s41588-022-01127-7.

Reference genome assemblies reveal the origin and evolution of allohexaploid oat

Yuanying Peng[1,2,3,4,10]*, Honghai Yan[1,3,10], Laichun Guo[2,4,10], Cao Deng[5,6,10], Lipeng Kang[7,8,10], Chunlong Wang[2,4], Pingping Zhou[1,3], Kaiquan Yu[3], Xiaolong Dong[3], Jun Zhao[3], Yun Peng[3], Xiaomeng Liu[3], Di Deng[3], Yinghong Xu[3], Ying Li[3], Qiantao Jiang[1,3], Yan Li[1], Liming Wei[2,4], Jirui Wang[1,3], Jian Ma[1,3], Ming Hao[1,3], Wei Li[1,3], Houyang Kang[1,3], Youliang Zheng[1,3], Yuming Wei[1,3]*, Fei Lu[7,8,9]*, Changzhong Ren[2,4,11]*

[1] State Key Laboratory of Crop Gene Exploration and Utilization in Southwest China, Sichuan Agricultural University, Chengdu, China.

[2] National Oat Improvement Center, Baicheng Academy of Agricultural Sciences, Baicheng, China.

[3] Triticeae Research Institute, Sichuan Agricultural University, Chengdu, China.

[4] China Oat and Buckwheat Research Center, Baicheng, China.

[5] The Key Laboratory of Animal Disease and Human Health of Sichuan Province, College of Veterinary Medicine, Sichuan Agricultural University, Chengdu, China.

[6] Departments of Bioinformatics, DNA Stories Bioinformatics Center, Chengdu, China

[7] State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Innovative Academy of Seed Design, Chinese Academy of Sciences, Beijing, China.

[8] University of Chinese Academy of Sciences, Beijing, China.

[9] CAS-JIC Centre of Excellence for Plant and Microbial Science (CEPAMS), Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China.

[10] These authors contributed equally to this work.

[11] Lead contact.

*Correspondence: Changzhong Ren (renchangzhong@163.com), Fei Lu (flu@genetics.ac.cn), Yuming Wei (ymwei@sicau.edu.cn), Yuanying Peng (yy.peng@hotmail.com)

Common oat (*Avena sativa*) is one of the most important cereal crops serving as a valuable source of forage and human food. While reference genomes of many important crops have been generated, such work in oat has lagged behind, primarily owing to its large, repeat-rich, polyploid genome. By using Oxford Nanopore ultralong sequencing and Hi-C technologies, we have generated the first reference-quality genome assembly of hulless common oat with a contig N50 of 93 Mb. We also assembled the genomes of diploid and tetraploid *Avena* ancestors, which enabled us to identify oat subgenome, large-scale structural rearrangements, and preferential gene loss in the C subgenome after hexaploidization. Phylogenomic analyses of cereal crops indicated that the oat lineage descended before wheat, offering oat as a unique window into the early evolution of polyploid plants. The origin and evolution of hexaploid oat is deduced from whole-genome sequencing, plastid genome and transcriptomes assemblies of numerous *Avena* species. The high-quality reference genomes of *Avena* species with different ploidies and the studies of their polyploidization history will facilitate the full use of crop gene resources and provide a reference for the molecular mechanisms underlying the polyploidization of higher plants, helping us to overcome food security challenges.

Common oat (*Avena sativa* L., 2n = 6x = 42) is the sixth most important cereal crop cultivated worldwide [1] and has long been prized by consumers largely because it is one of the richest sources of protein, fat and VB1 among all crops [2]. In addition, it is the most widely grown cool-season annual forage species, representing a major source of forage for livestock globally [3]. Polyploid plants often have significant advantages concerning biomass production, vigorousness, and robust adaptation to environmental changes and contribute to the emergence of important agronomic traits in food crops [4-8]. Therefore, crop polyploidization may play an important role in next-generation crop improvement aimed at overcoming food security challenges [8]. Many commercially important crops have been sequenced and assembled, which has improved the understanding of crop evolutionary history and the development of efficient approaches for the selection of important traits [9]. Oat research has lagged behind in this regard, primarily due to the large hexaploid oat genome [10] containing highly repetitive DNA sequences, in which two of the three subgenomes are too homologous to distinguish from each other [11]. Relatively little is currently known about the position and distribution of genes on each of the oat chromosomes and their evolution during the polyploidization events that gave rise to hexaploid species, which limits the full and effective utilization of oat germplasms.

With the recent advances made by Oxford Nanopore Technologies (ONT), the ONT system now offers ultralong sequence reads, delivering high contiguity with low assembly errors caused by long repetitive regions [12]. This technology has facilitated complete telomere-to-telomere (T2T) genome assembly in various species, including *Homo sapiens* by resolving long, complex repetitive regions [13,14]. It is very suitable for the assembly of the large, complex polyploid oat genome, with a high content of repetitive sequences and high subgenomic homology. Here, we employed the ONT

system to sequence the genome of the Sanfensan oat variety (*Avena sativa* ssp. *nuda* cv. Sanfensan, abbreviated SFS; see Methods), an ancient, important hexaploid oat landrace that originated from the diversity centre of hulless oat, together with assembling the genomes of additional diploid and tetraploid *Avena* species to address the phylogenomic relationship of cereal crops and gain insights into the evolutionary processes that established the dominant subgenome in allopolyploid oat species with large, complex genomes.

Assembly and annotation of the oat genome

Using the PromethION platform and the ultralong read sequencing strategy [13,15], we assembled the SFS genome into 329 contigs based on 1028 Gb cleaned ultralong reads (max length: 1.38 Mb), representing ~100× theoretical coverage with a contig N50 of 93.26 Mb and a maximum length of 405.55 Mb (Supplementary Tables 1-3). The longest 119 contigs contained 90% of the sequences, with the 34 longest contigs covering half the genome. The total assembly size was 10.76 Gb, including 10.44 Gb anchored to 21 pseudochromosomes using ~1.37 billion valid unique interaction pairs generated from 1296 Gb of cleaned Hi-C data (Supplementary Table 1, Extended Data Fig. 1a). The quality and contiguity of the genome assembly were assessed through alignments with the hexaploid consensus map and BUSCO pipeline. Most markers from the same linkage group were anchored to the corresponding pseudochromosome (Extended Data Fig. 2a), and 97.75% of the 1375 BUSCO genes were identified in the SFS genome assemblies, indicating a high-quality genome assembly (Extended Data Fig. 2b). A key feature of this assembly was its long-range organization, with contigs longer than 20.9 Mb representing 90% of the genome and an average of 15 contigs representing one chromosome (Table 1). The longest contig spanned 405 Mb, which

is larger than the size of rice genome [16]. This assembly represents a substantial improvement over the hexaploid oat genome sequenced by Pepsi Co. in 2020 (>1900 contigs with a contig N50 of 30.27 Mb, https://wheat.pw.usda.gov/jb/?data=/ggds/oat-ot3098-pepsico.) and recent assemblies reported for polyploid species with large, complex genomes [17,18].

Protein-coding genes were annotated using a combination of ab initio prediction and transcript evidence gathered from RNA sequences from multiple tissues based on PacBio isoform sequencing (Iso-seq) approaches and reference protein sequences from the other six closely related grass species (see Methods). In total, 120,769 protein-coding genes were identified (Table 1, Supplementary Table 4), 88.41% of which were assigned to a predicted function (Supplementary Table 5). Among these genes, the majority were retained in duplicate, largely owing to the allohexaploid nature of the genome. The C subgenome contained fewer gene loci (33181) than the A (40,085) and D (41,633) subgenomes. We also identified 59,916 noncoding RNAs, including 5,386 rRNAs, 50,536 small RNAs, 3,712 rRNAs and 283 cis-regulatory elements (Supplementary Table 6). Moreover, a total of 9.5 Gb (88.64%) of the SFS assembly was annotated as repetitive sequences (Table 1, Supplementary Table 7). Transposable element (TE)-related sequences constituted 86.2% of the total genome assembly. This TE content was higher than those previously reported for barley (80.80%) [19] and bread wheat (84.70) [9]. Long terminal repeat (LTR) retrotransposons, including Gypsy elements, unclassified retrotransposon elements and Copia elements, were the most abundant TEs and constituted 71.66% of the assembled SFS genome.

To distinguish the subgenomes accurately and clarify the polyploidization history of hexaploid oat, we also sequenced, assembled and annotated the most likely ancestral diploid species, *A. longiglumis*, and tetraploid species, *A. insularis* [20],

resulting in >60-fold genome coverage for *A. longiglumis* (218.67 Gb) and *A. insularis* (374.77 Gb). The assembled *A. longiglumis* genome was 3.74 Gb in size, and 99.20% (3.71 Gb) of its sequences were anchored on the 7 chromosomes of *A. atlantica* [21]. The tetraploid *A. insularis* genome was 7.52 Gb in size, 95.08% (7.15 Gb) of which was arranged into 14 chromosomes by Hi-C analysis (Extended Data Fig. 1b). We annotated 43,352 protein-coding genes in the *A. longiglumis* genome and 87,154 in the *A. insularis* genome. BUSCO analysis also revealed that a majority (97.53% and 98.11%, respectively) of conserved genes were identified in these two genome assemblies. Accordingly, most of the genome sequences of the *A. longiglumis* (87.82%) and *A. insularis* (87%) were composed of repetitive elements.

Based on the genomic synteny and similarity among the hexaploid, diploid, and tetraploid genomes, we successfully divided the 21 pseudochromosomes of hexaploid oat into A, C, and D subgenomes (Extended Data Fig. 3a-d). Am1 (a repeat selectively hybridized to the C subgenome) and As120a (a repeat selectively hybridized to the A subgenome) satellite repeats were overrepresented in the C and A subgenomes, respectively (Fig. 1). The subgenome assignments were further validated by mapping the molecular markers from a recently published high-density linkage map of hexaploid oat (Extended Data Fig. 2a) and by mapping the whole-genome sequencing reads from a range of *Avena* species (Extended Data Fig. 3e-f). The nomenclature system for bread wheat was adapted for naming the chromosomes of SFS (Extended Data Fig. 4).

Evolutionary position of oat among cereal crops

We next used our oat genome assemblies to assess the phylogenomic relationships among common oat and other cereal crops (Supplementary Table 8) by using 644

conserved core genes (Fig. 2a, Supplementary Fig. 1). Molecular dating estimated that the divergence of monocot cereal crops occurred ~45 million years ago (mya) and was concentrated in Poaceae. Pooideae and Oryzoideae split ~40 mya, whereas Panicoideae and Chloridoideae diverged ~33 mya. In Pooideae, the tribes Aveneae and Lolieae were indicated to be more closely related than to Triticeae. The oat lineage descended (~16 mya) earlier than that of wheat (~7 mya), indicating a longer time of evolution and offering a unique window into the early evolution of polyploid cereal crops.

A high-confidence orthologous gene set for the 43 species was obtained, 2202 gene families shared by all the genomes/subgenomes were identified, and 4387, 419, 459, 24, 152, 450, and 40 gene families were indicated to be specific to Polygonaceae, Chenopodiaceae, Chloridoideae, Panicoideae, Aveninae, and Triticinae, respectively (Fig. 2b, Supplementary Table 9). Most of these *Avena*-specific gene families were assigned to molecular binding or metabolic process functions (Supplementary Table 10). In addition, 10 of them may play roles in fat biosynthesis, including 479 genes involved in fatty acid biosynthesis. These genes might contribute to the much higher concentration of oil in oat grains relative to other cereal crops (Supplementary Table 10).

The ancestral grass karyotype (AGK) was inferred by comparing modern species via a completely automated method reconstructing the order of ancestral genes within contiguous ancestral regions and is structured into 7 protochromosomes containing 7,010 protogenes [22]. We therefore used the 7 AGK protochromosomes as an ancestral reference to investigate the chromosomal evolution of rice, wheat and oat (Fig. 2c). A total of 731 syntenic blocks (>5 genes) encompassing 19112 orthologous gene pairs were identified between oat and the 7 AGK protochromosomes (Supplementary Table

11). Essentially, the third homologous group of oat was derived from a single ancient chromosome, AGK1, (which also steadily gave rise to the complete chromosomes of rice and wheat), with the exception of a segment of AGK2 that was translocated to chromosome 3C in oat. Group six was derived from the insertion of AGK5 into AGK4. The remaining five of the seven oat homologous groups were each derived from at least three ancestral chromosomes via complex translocations (Fig. 2c). When the oat assembly was compared with the three subgenomes of common wheat, the ancestral chromosomes were found to be roughly similar, but a large number of chromosomal rearrangements occurred, leading to many structural variations and lower collinearity.

Polyploidization history and reticulate evolution

Previous phylogenetic studies aimed at identifying the progenitors of these species have often analysed only a limited set of molecular markers, yielding inconsistent results. To more accurately identify the relatives of each subgenome donor and the extant ancestral hexaploid species, we sequenced and de novo assembled transcriptomes, conducted whole-genome sequencing and assembled the complete plastid genomes of other *Avena* species. These *Avena* species represented different genomic subtypes (As, Al, Ac, Ad, Cv, and Cp) and different ploidies.

We sequenced 14 different *Avena* species (generating 6767.64 Gb of sequence data, Supplementary Table 1) and mapped the obtained data onto the reference SFS genome. The genomic variants of these 14 species as well as the assembled *A. longiglumis* and *A. insularis* genomes relative to the SFS sequences were called based on the mapped reads. Approximately 39~129 million SNPs were identified in these species (Supplementary Table 12). The ratios of SNP numbers in the A, C and D subgenomes

of each species showed that the C subgenome is differed greatly from the A, B and D subgenomes (Fig. 3a). These SNP-based phylogenetic analyses indicated that the species clustered according to the genome composition except for the AB genome tetraploids, which were positioned closer to different A genome diploids (Fig. 3b). Three maximum likelihood trees based on the A-, C- and D-type SNPs further implied that *A. longiglumis* was the most closely related A genome donor for hexaploids, while no extant diploids showed a particularly close relationship with the C and D subgenomes (Extended Data Fig. 5). The identity distribution of different sequenced reads clearly showed that the Al genome of the diploid species *A. longiglumis*, and the C and D subgenomes of *A. insularis* have the closest relationships with the A, C and D subgenomes of hexaploids, respectively (Fig. 3c). These results further confirmed that the diploid *A. longiglumis* and the tetraploid *A. insularis* were the most closely related extant species to hexaploid oat.

We sequenced and de novo assembled 11 transcriptomes of different diploid species (Supplementary Table 13) to further clarify the relationships between extant diploid species and cultivated hexaploid oat. The phylogenetic analyses based on 2863 nuclear genes using barley as the outgroup again provided strong support for *A. longiglumis* as the extant diploid with the closest relationship to the A subgenome of hexaploids and for no diploid being a D genome donor. However, the A subgenome of hexaploid *A. sativa* differentiated from the ancestral genome earlier than the A genome of diploid *A. longiglumis* (Fig. 2a, Fig. 3d). This finding suggests that the A genome progenitor of hexaploids was more likely the A diploid ancestor than an extant A genome diploid, which may be the reason that previous studies [23] have considered different A genome diploids to be the donors of hexaploid genomes.

To determine the maternal origin of hexaploid oat, we assembled approximately

complete plastid genomes of 136 kb for each *Avena* species. Phylogenetic analysis of 18 complete *Avena* plastid genomes together with the previous 26 *Avena* chloroplast genomes (Supplementary Tables 14-15) [24,25] showed that the C genome was undoubtedly the male parent in polyploid formation and that the D genome, rather than the A genome, was the maternal donor of hexaploid oat. The evolutionary order of the different A genome subtypes was Ac-Ad-Al-As (Fig 3e, Supplementary Fig. 2). The relatively low collinearity between C genome diploid species and the C subgenome of *Avena* polyploids was consistent with the nuclear-cytoplasmic interaction hypothesis suggesting that the paternally inherited genome of an allopolyploid is usually more prone to genetic changes than the maternally derived genome [26].

Our phylogenetic analyses revealed a common ancestor of A and D genome diploids, indicating that the D genome ancestor is showed a closer relationship to the A genome and may be extinct. The C and A/D lineages diverged from the ancestor of the genus *Avena* ~17 mya, followed by the A genome subtypes, and the D genome split from a common ancestor ~9 mya (Fig. 3d, 3f). Cultivated oat initially originated from an allotetraploidy event between a paternal C genome and a maternal D genome diploid (~2 mya, Supplementary Fig. 4), followed by hybridization between a paternal A genome diploid progenitor (closely related to *A. longiglumis*) and maternal CD genome tetraploid (closely related to *A. insularis*) to form the extant ACD genome hexaploids (~0.5 mya, Supplementary Table 16, Supplementary Fig. 3). Our findings clarify the evolutionary history of oats based on various lines of evidence (Fig. 3f) and have broad implications for the understanding of genome function and cultivar improvement in oat.

Subgenome structure, content, and response of subgenomes to polyploidization

Synteny analyses between the subgenomes of hexaploid oat revealed a much lower level of conservation in oat (62.70% of genes organized in collinear blocks, 75727/120769) than in bread wheat (72.57%, 77292/106508). The A and D subgenomes of oat showed a higher synteny than the C and A or D subgenomes, as indicated by the numbers of collinear genes preserved between the three subgenomes (Extended Data Fig. 6a). Such low levels of conservation between the C and D subgenomes have also been observed in *A. insularis* (Extended Data Fig. 6b), supporting the hypothesis that both the A and D subgenomes were derived from a common diploid lineage that diverged from the C subgenome. The low levels of conservation between the hexaploid subgenomes, particularly between C and A or D, may reflect the preexisting differences in the subgenomes before polyploidization but could also be attributed to chromosomal rearrangements after polyploidization. To explore broad-scale structural evolution after polyploidization, we performed a whole-genome synteny comparison, which revealed low collinearity between the sequenced diploid C and tetraploid C subgenomes, especially for chromosome7C (Fig. 4a), suggesting that large chromosomal rearrangements might have accumulated during evolution. We also found evidence of large chromosomal rearrangements during hexaploidization (Fig. 4a, Extended Data Fig. 7a, b). For instance, one C to A introgression and four C to D introgressions (Fig. 4c, Extended Data Fig. 7d) occurred. These large translocations were confirmed by FISH assays (Fig. 4d, Extended Data Fig. 7e). Moreover, the structural rearrangements in oat appeared biased among the three subgenomes (Fig. 4b): most of them occurred between subgenomes A and D, whereas fewer occurred between subgenomes A and C and between subgenomes D and C. For instance, half of the tetraploid 6D chromosome was transferred to

hexaploid 6A, 7A and 2D chromosomes, resulting in a much smaller hexaploid 6D chromosome. These results may support the hypothesis that the existence of two well-conserved homologous genomes would facilitate inter-subgenome recombination and rearrangement after polyploidization, as observed in a previous study [27].

The C subgenome (3.94 Gb) is 20% larger than the A (3.28 Gb) and D (3.22 Gb) subgenomes in hexaploid (Table 1). TE abundance accounted for almost all of the size differences between the subgenomes (Fig. 1, Extended Fig. 8a). We further found that C genomes contained twice as many full-length LTRs as A and D genomes (Extended Data Fig. 8b), and that the average length of LTRs was shorter in the C subgenome than in the A and D subgenomes (Extended Data Fig. 8c). The amplification of full-length LTRs in C genomes occurred over a relatively long period, peaking approximately 1.8 mya, whereas LTRs were inserted into A and D genomes within the last 1 mya (Extended Data Fig. 8d), which indicates that the C genome has a longer evolutionary history than the A and D genomes. The C subgenome contained a smaller number of gene loci (33181; 29%) than the A (40,085, 35%) and D (41,633, 36%) subgenomes in hexaploid (Table 1), whereas the tetraploid (89,995) showed almost the same gene number as the A (43,477) and C (46,925) diploids together (Table 1), suggesting that a remarkable gene loss occurred in the C subgenome after hexaploidization (33,181 in hexaploid vs 43,243 in tetraploid). These results were supported by the gene loss ratio in the subgenomes of *A. insularis* and SFS (Extended Data Fig. 9a, b). We further found that more gene families underwent decreases in hexaploid C subgenomes than in A and D subgenomes (Extended Data Fig. 9c), and that gene losses mainly affected genes with expanded families, rather than singleton genes (Extended Data Fig. 9d, e). Accordingly, the hexaploid C subgenome contained more pseudogenes (29297) than the A (26141) and D (27262) subgenomes (Table 1,

295 Extended Data Fig. 9f). Together, these results suggest preferential gene loss or
296 pseudogenization in the C subgenome after hexaploidization.

297 Subgenome dominance is a striking whole-genome feature common to polyploids,
298 presumably associated with biased fractionation [28]. Genes located in the dominant
299 subgenome always present a lower nonsynonymous ratio and higher expression levels
300 than orthologues in other subgenomes[29,30]. To test whether subgenome dominance has
301 occurred in hexaploid oat, we first compared nonsynonymous (*Ka*) with synonymous
302 *(Ks*) substitution rates between (sub)genomes using 2,767 one-to-one orthologous
303 gene sets among the polyploid oat and the progenitors, and the results suggested that
304 an accelerated evolution rate exists in polyploids relative to their diploid progenitors
305 (Supplementary Fig. 5). In SFS, the C subgenome presented a significantly higher
306 *Ka/Ks* value (0.37) than the A (0.17) and D (0.25) subgenomes (Fig. 4e), suggesting
307 that the C subgenome is subject to more relaxed selective constraints than the other
308 two subgenomes. We further examined the expression of 13,774 single-copy genes in
309 the three subgenomes, referred to as triads, in diverse organs and under six different
310 abiotic stresses. We found that most genes were expressed in a subgenome-specific
311 manner (Fig. 4f). Pairwise comparisons of the homologous gene triads revealed that a
312 similar number of gene pairs were preferentially transcribed in the A and D
313 subgenomes; however, the C subgenome had a significantly lower number of
314 preferentially transcribed gene pairs than the A (P=1.3E-3) and D (P=6.1E-3)
315 subgenomes (Fig. 4g, Supplementary Tables 17, 18). TE abundance plays an
316 important role in gene expression bias because TE insertion can have deleterious
317 effects on genes leading to lower expression [31]. In oat, we found that the C
318 subgenome contained more TEs and showed higher overall TE densities near genes
319 than the A and D subgenomes; these characteristics are negatively correlated with

gene expression (Extended Data Fig. 10), supporting the hypothesis that subgenome expression dominance is influenced by TE-density differences between subgenomes as observed in other allopolyploids [7,32]. All of these results revealed a biased gene fractionation in the hexaploid C subgenome, which might be closely related to the high global TE amount and TE density near genes, and the relaxation of selection pressure of the C subgenome.

In conclusion, the three reference-quality genomes presented here constitute important community resources for cereal genomics and provide new insights into the evolutionary position of oat among cereal crops. This is the most comprehensive molecular phylogenetic analysis of the genus *Avena* conducted to date, because it included samples representing all extant *Avena* genomes and the greatest number (and number of different types) of molecular markers evaluated thus far. The model for the chronological formation of polyploid oat has been clarified, and the investigation of oat subgenomes evolution during polyploidization events offers a unique window into the early evolution of polyploid crops, which will accelerate improvements of oat and help to make oat a useful model for studying polyploid genome evolution.

1 FAOSTATS. Food and agriculture organization of the united nations--statistics division. http://www.fao.org (2021).

2 Fu, J. *et al.* Concise review: Coarse cereals exert multiple beneficial effects on human health. *Food chem* **325**, 126761, doi:10.1016/j.foodchem.2020.126761 (2020).

3 Fraser, J. & McCartney, D. *Fodder oats in North America*. in Fodder oats: A World Overview (eds Suttie, J. M & Reynolds, S. G.) 19-35 (FAO, 2004).

4 Comai, L. The advantages and disadvantages of being polyploid. *Nat Rev*

Genet **6**, 836-846, doi:10.1038/nrg1711 (2005).

5 Fang, Z. & Morrell, P. L. Domestication: Polyploidy boosts domestication. *Nat Plants* **2**, 16116, doi:10.1038/nplants.2016.116 (2016).

6 Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat Rev Genet* **18**, 411-424, doi:10.1038/nrg.2017.26 (2017).

7 Edger, P. P. *et al.* Origin and evolution of the octoploid strawberry genome. *Nat Genet* **51**, 541-547, doi:10.1038/s41588-019-0356-4 (2019).

8 Yu, H. *et al.* A route to de novo domestication of wild allotetraploid rice. *Cell* **184**, 1156-1170. e14, doi:10.1016/j.cell.2021.01.013 (2021).

9 Appels, R. *et al.* Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191, doi:10.1126/science.aar7191 (2018).

10 Yan, H. *et al*. Genome size variation in the genus *Avena*. *Genome* **59**, 209-220, doi:10.1139/gen-2015-0132 (2016).

11 Jellen, E., Gill, B. & TS, C. Genomic in situ hybridization differentiates between A/D- and C-genome chromatin and detects intergenomic translocations in polyploid oat species (genus *Avena*). *Genome* **37**, 613-618, doi:10.1139/g94-087 (1994).

12 Lang, D. *et al.* Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. *GigaScience* **9**, doi:10.1093/gigascience/giaa123 (2020).

13 Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79-84, doi:10.1038/s41586-020-2547-7 (2020).

14 Logsdon, G. A. *et al.* The structure, function and evolution of a complete

human chromosome 8. *Nature* **593**, 101-107, doi:10.1038/s41586-021-03420-7 (2021).

15 Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**, 338-345, doi:10.1038/nbt.4060 (2018).

16 Sasaki, T. & International Rice Genome Sequencing, P. The map-based sequence of the rice genome. *Nature* **436**, 793-800, doi:10.1038/nature03895 (2005).

17 Lovell, J. T. *et al.* Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature* **590**, 438-444, doi:10.1038/s41586-020-03127-1 (2021).

18 Huang, G. *et al.* Genome sequence of *Gossypium herbaceum* and genome updates of *Gossypium arboreum* and *Gossypium hirsutum* provide insights into cotton A-genome evolution. *Nat Genet* **52**, 516-524, doi:10.1038/s41588-020-0607-4 (2020).

19 Mascher, M. *et al.* A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**, 427-433, doi: 10.1038/nature22043 (2017).

20 Yan, H. *et al.* High-density marker profiling confirms ancestral genomes of *Avena* species and identifies D-genome chromosomes of hexaploid oat. *Theor Appl Genet* **129**, 2133-2149, doi:10.1007/s00122-016-2762-7 (2016).

21 Maughan, P. J. *et al.* Genomic insights from the first chromosome-scale assemblies of oat (*Avena* spp.) diploid species. *BMC Biol* **17**, 92, doi:10.1186/s12915-019-0712-y (2019).

22 Murat, F., Armero, A., Pont, C., Klopp, C. & Salse, J. Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat Genet* **49**, 490-496, doi:10.1038/ng.3813 (2017).

23 Peng, Y. *et al*. Phylogenetic inferences in *Avena* based on analysis of *FL intron2* sequences. *Theor Appl Genet* **121**, 985-1000, doi:10.1007/s00122-010-1367-9 (2010).

24 Fu, Y. B. Oat evolution revealed in the maternal lineages of 25 *Avena* species. *Sci Rep* **8**, 4252, doi:10.1038/s41598-018-22478-4 (2018).

25 Liu, Q. *et al.* Comparative chloroplast genome analyses of *Avena*: insights into evolutionary dynamics and phylogeny. *BMC Plant Biol* **20**, 406, doi:10.1186/s12870-020-02621-y (2020).

26 Gill, B. S. & Friebe, B. Nucleocytoplasmic interaction hypothesis of genome evolution and speciation in polyploid plants revisited: polyploid species-specific chromosomal polymorphisms in wheat. in *Polyploid and Hybrid Genomics* (eds Chen, Z. J & Birchler, J. A) 213-221, doi:10.1002/9781118552872.ch13 (Wiley, 2013).

27 Zhuang, W. *et al.* The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat Genet* **51**, 865-876, doi:10.1038/s41588-019-0402-2 (2019).

28 Alger, E. I. & Edger, P. P. One subgenome to rule them all: underlying mechanisms of subgenome dominance. *Curr Opin Plant Biol* **54**, 108-113, doi:10.1016/j.pbi.2020.03.004 (2020).

29 Parkin, I. A. P. *et al.* Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. *Genome Biol* **15**, R77, doi:10.1186/gb-2014-15-6-r77 (2014).

30 Liu, S. *et al.* The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nat Commun* **5**, 3930, doi:10.1038/ncomms4930 (2014).

31 Hollister, J. D. & Gaut, B. S. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res* **19**, 1419-1428, doi:10.1101/gr.091678.109 (2009).

32 VanBuren, R. *et al.* Exceptional subgenome stability and functional divergence in the allotetraploid Ethiopian cereal teff. *Nat Commun* **11**, 884, doi:10.1038/s41467-020-14724-z (2020).

Table 1 | Assembly statistics for diploid, tetraploid and hexaploid *Avena* species.

	A. longiglumis CN 58138 (2n = 2x = 14)	*A. insularis* CN 108634 (2n = 4x = 28)	*A. sativa* ssp. *nuda* cv. Sanfensan (2n = 6x = 42)
Illumina (Gb)	204.68	451.89	649.68
ONT (Gb)	268.74	481.39	-
ONT ultralong (Gb)	-	-	1260.30
HiC (Gb)	-	816.93	1312.83
IsoSeq (Gb)	25.74	49.94	81.14
Total assembly size (bp)	3,736,548,545	7,519,018,440	10,757,433,345
Longest contigs (bp)	29,014,927	36,557,065	405,550,188
Number of contigs	956	1,924	326
N50 contig length (bp)	7,297,603	7,836,599	93,262,735
L50 contig count	160	297	34
N90 contig length (bp)	2,123,884	2,126,637	20,933,943
L90 contig count	523	1,003	119
Number of contigs per (sub)genome	A: 943		A: 84
		C: 1,019	C: 155
		D: 848	D: 77
	Total: 943	T: 1,867	T: 31
Sequences assigned per (sub)genome (bp)	A: 3,708,832,268		A: 3,279,788,166
		C: 4,020,068,809	C: 3,942,742,583
		D: 3,131,824,267	D: 3,215,854,646
	T: 3,708,832,268	T: 7,151,893,076	T: 10,438,385,395
Genome completeness (BUSCO)	97.53	98.11	97.75
Repeats (%)	87.09%	87.11%	87.12%
Protein-coding genes	A: 43,477		A: 40,085
		C: 46,925	C: 33,181
		D: 43,477	D: 41,633
	T: 43,477	T: 89,995	T: 120,769
Pseudogenes	A: 14,058		A: 23,013
		C: 20,847	C: 26,510
		D: 17,766	D: 24,009
	T: 14,058	T: 38,613	T: 73,532

Note: A: A subgenome, C: C subgenome, D: D subgenome, T: Total.

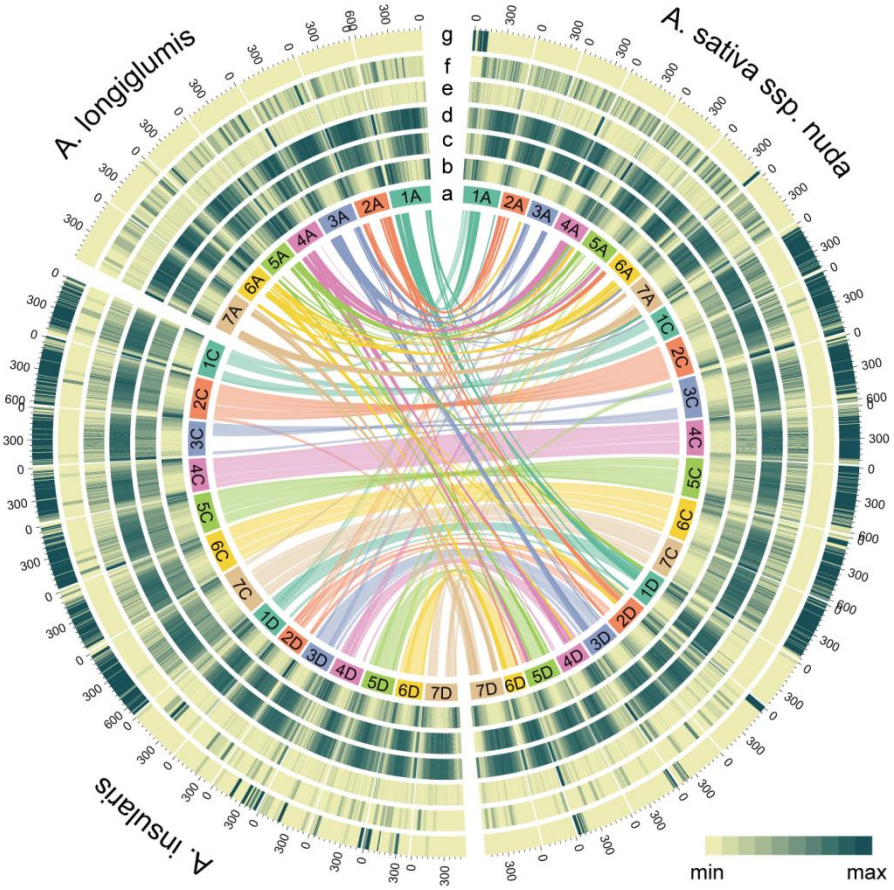


Fig. 1 | Genomic features of oat. Circos display of important features of the assembled diploid, tetraploid and hexaploid *Avena* species genomes. **a**, Chromosome names and sizes. **b**, Gene density. **c**, Long terminal repeat (LTR) retrotransposon density. **d**, Tandem repeat (TR) density. **e**, *k*-mer frequencies. **f**, The A genome-specific repeat As120a. **g**, The C genome-specific repeat Am1. Links between syntenic genes are shown in different colours.

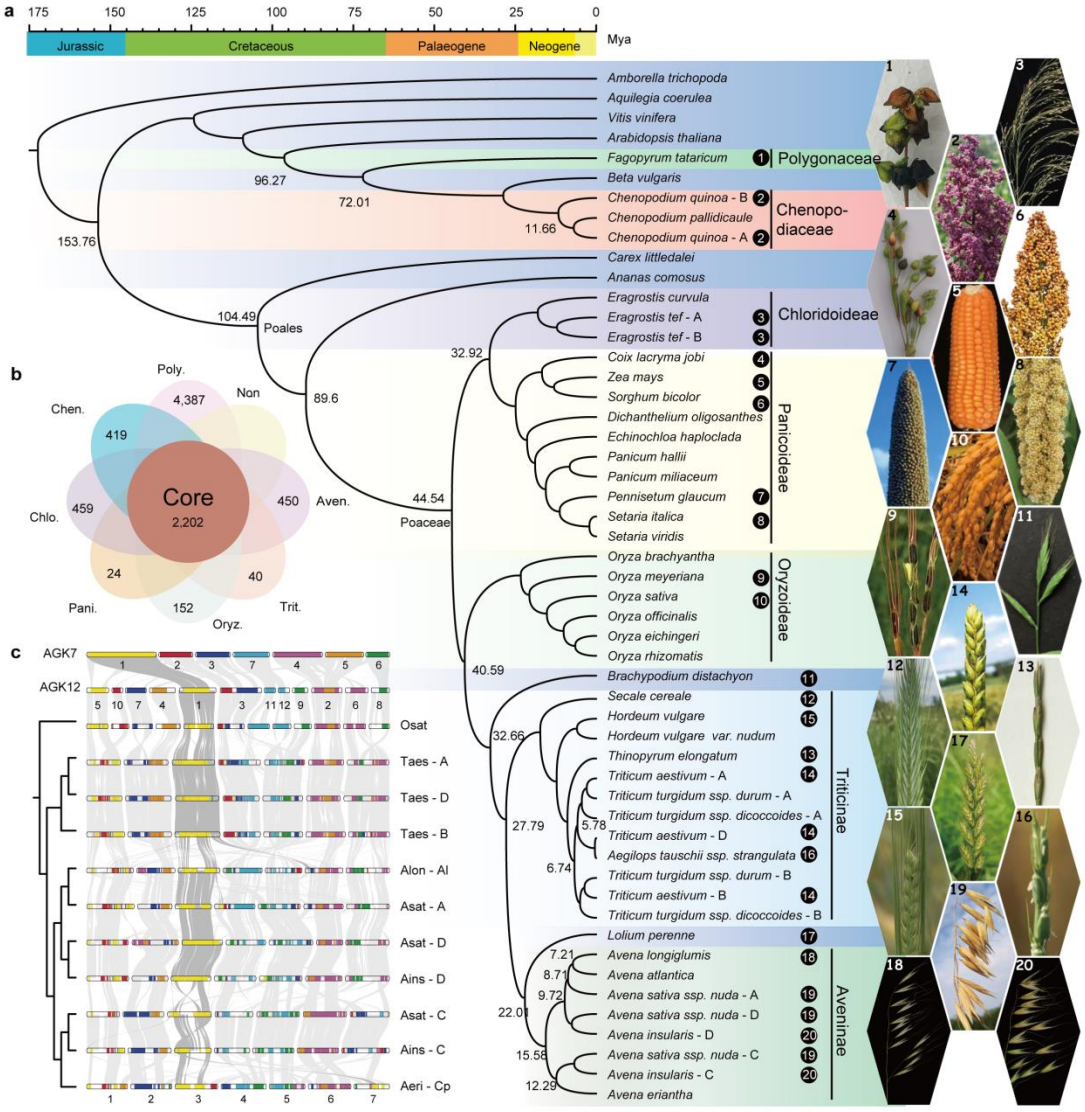


Fig. 2 | Phylogenomic relationships of cereal crops. a, Phylogeny and time scale of 43 plant species, including 33 assembled cereal crops and panicles of the 20 sampled cereal crops. The number in the upper left corner of each panicle image corresponds to the number following the species name in the phylogenetic tree. **b**, Numbers of gene families shared among and unique to Polygonaceae, Chenopodiaceae, Chloridoideae, Panicoideae, Aveninae and Triticinae. **c**, *Avena* chromosome evolution according to the AGK model and chromosomal synteny between the three subgenomes of oat and related cereal crops. The genomes used for collinearity analysis are marked with an abbreviation of the full name in the phylogenetic tree.

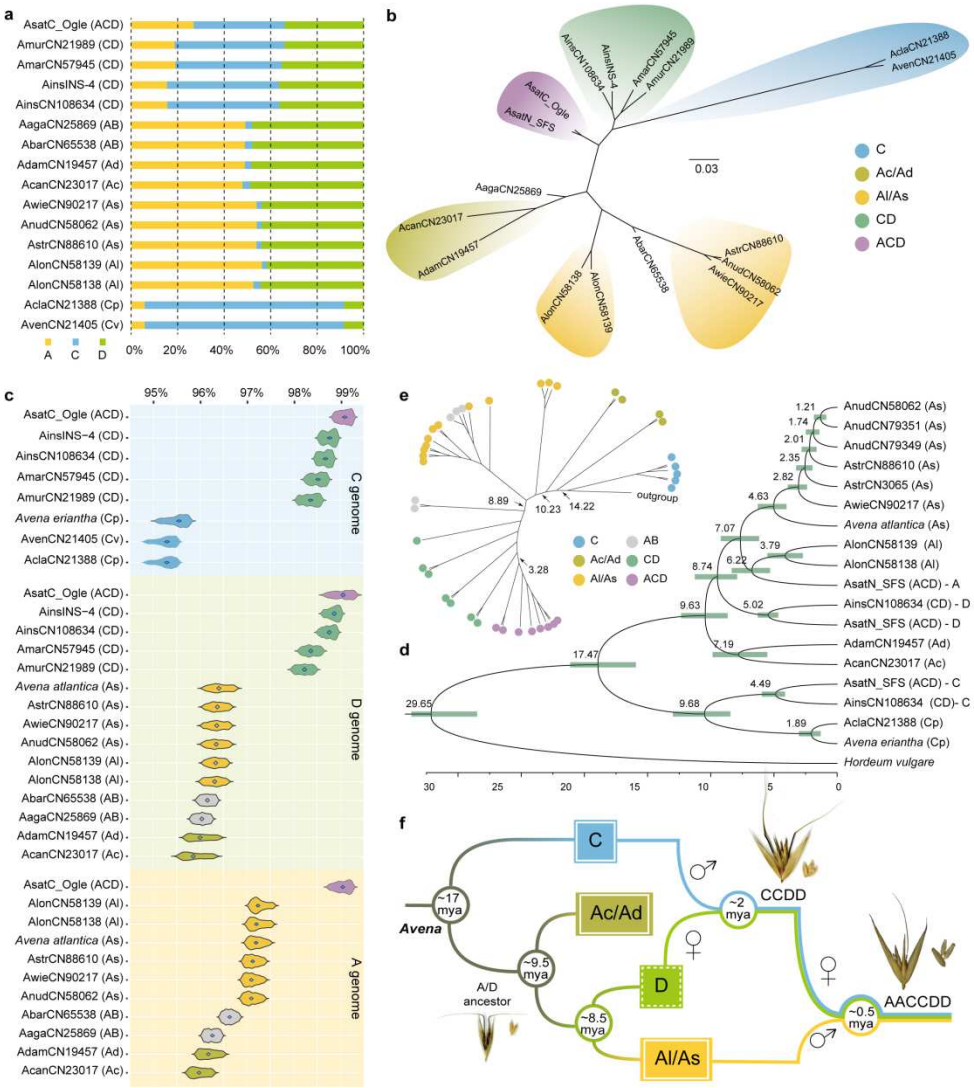


Fig. 3 | Phylogeny and polyploidization of *Avena* species. a, Comparable numbers of A-, C- and D-type SNPs for each taxon. **b**, Maximum likelihood tree generated from the SNPs. **c**, Identity distribution of different sequenced reads on different subgenomes of hexaploids. **d**, Phylogenetic analyses based on 2,863 nuclear genes using barley as an outgroup. **e**, Phylogenetic studies of complete plastid *Avena* genomes. **f**, Model of the phylogenetic history of oat (*Avena sativa*; AACCDD)

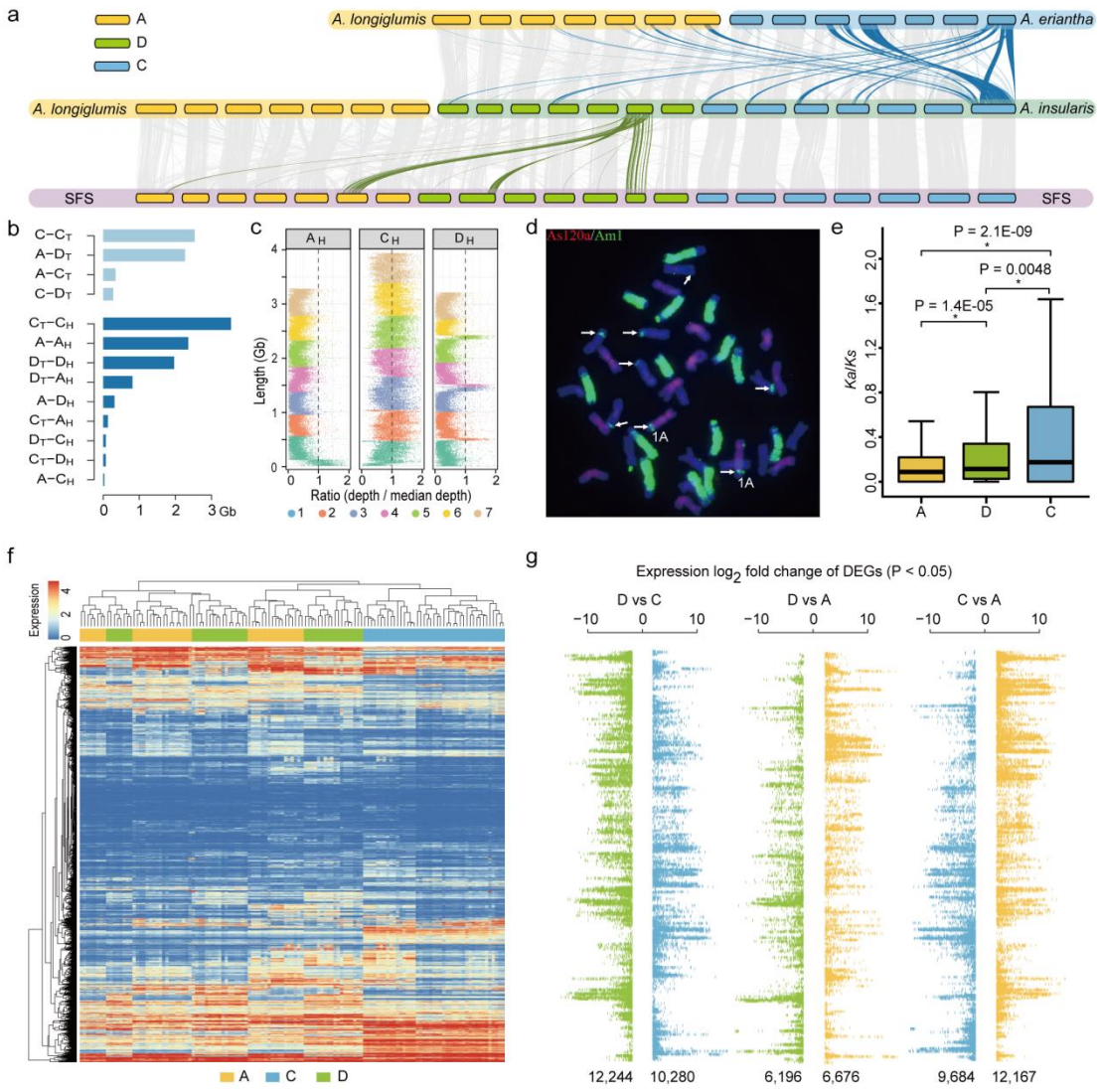


Fig. 4 | Genomic structures of hexaploid oat and its relatives. a, Different colours of lines depict synteny between subgenomes of SFS and putative tetraploid and diploid ancestors. Thick lines connecting chromosomes 7C and 6D depict observed large-scale chromosome rearrangements. **b**, Length of individual subgenome DNA sequences inherited from ancestral genomes. Subscript letter represents subgenomes of hexaploid (H) or tetraploid (T). **c**, Reads of the C genome diploid mapped to the SFS reference reveal at least four large C to D genomic exchanges and one C to A genomic exchange after polyploidization. **d**, FISH using the C genome-specific repeat as a probe confirms the C to D and C to A translocations. Red and green represent FISH signals detected with an A-specific repeat (As120a) or a C-specific repeat (Am1) as probes. White arrows indicate C to D or C to A subgenome translocations. **e**,

Comparison of *Ka/Ks* value distributions between three subgenomes of SFS. The central line for each box plot indicates the median. The top and bottom edges of the box indicate the 25th and 75th percentiles, and the whiskers extend 1.5 times the interquartile range beyond the edges of the box. The asterisks represent significant differences (* Wilcoxon rank-sum test, $P<0.05$). **f**, Two-dimensional hierarchical cluster analysis of expression among single-copy homologous oat genes compared with organ-specific gene expression. **g**, Analysis of log2-fold changes in pairwise gene expression between homologous genes.

Methods

Plant materials

The hexaploid species SFS (2n=6x=42, AACCDD), the diploid species *A. longiglumis* (accession CN 58139, 2n=2x=14, AlAl) and the tetraploid species *A. insularis* (accession CN 108634, 2n=4x=28, CCDD) were selected for genome sequencing and assembly. SFS is a traditional hulless oat landrace that has a long cultivation history in Shanxi, China, which is thought to be the region of origin of hulless oat varieties. It has also been widely used as a parental line in hulless oat breeding programs, while *A. insularis* and *A. longiglumis* have been considered the most likely tetraploid and diploid ancestors of hexaploid oat [20].

Genome sequencing and assembly

Illumina HiSeq X-Ten or MGISEQ2000 platform was used to generate short paired-end reads for SFS, *A. insularis* and *A. longiglumis*. For Illumina sequencing, paired-end libraries were constructed by using the TruSeq Nano DNA HT Sample preparation kit (Illumina, USA) according to the manufacturer's instructions. For MGI sequencing, method described in Supplementary Note was used for paired-end libraries construction. A total of 649.7 Gb, 451.9 Gb and 204.7 Gb of raw reads were generated for SFS, *A. insularis* and *A. longiglumis*, respectively (Supplementary Table 1). Oxford Nanopore Technology (ONT) platforms were used to generate long reads for these three taxa. Considering the large and complex genome of hexaploid oat, the ONT ultralong reads were generated for the SFS, whereas regular ONT long reads were generated for the genomes of *A. longiglumis* and *A. insularis*. For ONT ultralong reads library construction, high-molecular-weight genomic DNA molecules were extracted from young leaves using the SDS method without the purification step to maintain the length of the DNA. Then, 8-10 μg of high-quality DNA was

size-selected (>50 kb) and processed using the Ligation Sequencing 1D kit (SQK-LSK109, Oxford Nanopore Technologies, UK) according to the manufacturer's instructions. For regular Nanopore library construction, 3-4 μg of high-molecular-weight DNA was size selected (>20 kb) and ligated using the SQK-LSK109 kit with the recommended protocol. All the prepared libraries were sequenced on the PromethION platform at the Genome Center of Grandomics (Wuhan, China). A total of 1260.3 Gb, 481.4 Gb and 268.7 Gb of raw ONT (ultra-)long reads for SFS, *A. insularis* and *A. longiglumis* were produced from 71, 7 and 8 libraries, covering approximately 100-, 60- and 60- fold of their genomes, respectively (Supplementary Table 2).

The raw ONT long reads were subjected to self-correction using the NextCorrect module implemented in NextDenovo (https://github.com/Nextomics/NextDenovo) (v2.0-beta.1). The corrected reads were then assembled into contigs by using NextDenovo (Supplementary Table 3). To improve the accuracy of the assembled contigs, a two-step polishing strategy was applied: the corrected Nanopore reads were first used for initial polishing with Racon [33] (three rounds), and the highly accurate short reads were then used to further correct the assemblies with NextPolish (four rounds).

We employed Hi-C technology to obtain chromosome-level genome assemblies of SFS and *A. insularis*. Hi-C libraries were created from tender leaves of SFS and *A. insularis*. In brief, tender leaves were fixed with formaldehyde, and the cross-linked DNA was then isolated, purified, and digested with *DpnII*. Sticky ends were filled and marked with biotin-14-dATP; the resulting blunt-end fragments were ligated to form chimaeric junctions, physically sheared, and enriched for fragments with a size of 300-600 bp. Chimaeric fragments representing the original cross-linked long distance

524 physical interactions were then processed to obtain paired-end sequencing libraries,
525 which were sequenced on the Illumina platform. A total of 1312.8 Gb and 816.9 Gb of
526 raw Hi-C data were generated for SFS and *A. insularis*. The polished contigs of SFS
527 and A. insularis were further clustered, ordered, anchored to pseudochromosomes
528 using Hi-C data with the LACHESIS program [34], whereas RaGOO [35] with the default
529 parameters was employed to anchor the contigs of *A. longiglumis* to seven
530 pseudochromosomes with the previously published As genome of diploid *A. atlantica*
531 as the reference. Marker sequences from the oat consensus genetic linkage map that
532 was derived from ten populations [36] were aligned to the final SFS assemblies to
533 evaluate the consistency between the Hi-C map and the genetic map (Extended Data
534 Fig. 2). The completeness of the assembled genomes was evaluated with the BUSCO
535 (v3.1.0) [37] pipeline.

536 **Subgenome assignment**

537 A reference-guided strategy based on subgenome homology was used to distinguish
538 the subgenomes of *A. insularis* and SFS. The *A. longiglumis* reference genome was
539 split into 100 bp markers and mapped onto the hexaploid genome assemblies. Unique
540 mapped markers were retained (Extended Data Fig. 3a). A syntenic block was defined
541 based on the presence of at least five synteny markers. The chromosomes with the
542 highest homology to *A. longiglumis* were assigned to the A subgenomes, and the
543 chromosomes with the second-highest homology were assigned to the D subgenomes
544 because of the high homology between the *Avena* A and D genomes according to
545 previous studies; the remaining chromosomes with the lowest homology were
546 accordingly assigned to the C subgenomes (Extended Data Fig. 3c). The
547 chromosomes of *A. insularis* were aligned with those of SFS and assigned to C and D
548 subgenomes (Extended Data Fig. 3b, d). The subgenome assignments were further

549 validated by two independent approaches. First, trimmed short reads from *A.*
550 *longiglumis* and *A. insularis* were mapped to the reference SFS genome, and the depth
551 coverage of paired reads from these two species was quantified (Extended Data Fig.
552 3e-f). Second, the abundance and distribution of two types of DNA repeats, As120a
553 and Am1, in the *A. insularis* reference and SFS genomes were investigated; these
554 repeat types have been reported to be overrepresented in the *Avena* A and C genomes,
555 respectively (Fig. 1).

556 **Repeat and protein-coding gene annotation**

557 Each of the whole-genomes was searched for repetitive sequences including tandem
558 repeats (TRs) and transposable elements (TEs). TRs were identified by using GMATA
559 [38] and Tandem Repeats Finder [39]. A species-specific de novo repeat library was
560 constructed using MITE-Hunter [40], LTR_FINDER [41] and RepeatModeler
561 (https://github.com/Dfam-consortium/RepeatModeler). Then RepeatMasker [42] was
562 adapted to search for TEs in the reference genome against Repbase [43] and the
563 species-specific *de novo* repeat library.

564 Protein coding genes were predicted using an evidence-based annotation workflow
565 by integrating evidence from transcriptomic data, homologue searches and ab initio
566 prediction. Transcriptomic data were generated by performing PacBio Iso-seq using
567 total RNA isolated from mixed organs. Raw reads were processed with IsoSeq3,
568 LIMA, and REFINE to identify full-length, nonchimeric circular consensus sequences
569 (CCSs). The resulting high-quality CCSs were mapped onto the reference genome for
570 de-redundancy. The nonredundant isoforms were then used to determine the locations
571 of potential intron-exon boundaries using GeneMarkST [44]. Protein sequences from *A.*
572 *atlantica*, *A. eriantha*, *Brachypodium distachyon*, *Hordeum vulgare*, *Oryza sativa*, and
573 *Triticum aestivum* were used as protein evidence. *Ab initio* gene prediction was

performed using GeneMark-ET [45] and AUGUSTUS [46] with two rounds of iterative training. All gene predictions were integrated using the recommended settings of EVidenceModeler (EVM) [47] after removing transposable element-related genes, pseudogenes and noncoding genes using TransposonPSI [48] with the default settings.

Noncoding RNAs (ncRNAs), including microRNAs, small nuclear RNAs, rRNAs and regulatory elements, were identified using the Infernal [49] program to search against the Rfam database [50]. The rRNAs, tRNAs and miRNAs were identified using RNAmmer [51], tRNAscan-SE [52], and miRanda v3.0, respectively. Functional assignments of the predicted protein-coding genes were obtained with BLAST by aligning the coding regions to sequences in public protein databases, including the NCBI nonredundant (NR) protein, Kyoto Encyclopedia of Genes and Genomes (KEGG), Eukaryotic Orthologous Groups of proteins (KOG), Gene Ontology (GO) and SwissProt databases. The putative domains and GO terms of the predicted genes were identified using the InterProScan [53] program with the default settings.

Phylogenomic analyses of cereal crops

Foughty-three plant species that containing the main cereal crops with high-quality reference genomes were downloaded to infer the phylogenetic relationships of cereal crops. For this purpose, single-copy gene families were identified using reciprocal best hit (RBH)-based methods. In brief, each proteome of 42 species (51 subgenomes, Supplementary Table 8) was subjected to BLAST searches against *Amborella trichopoda* sequences according to an E-value $\leq 1e^{-5}$. The RBHs in each pair were obtained and the gene families present in all the subgenomes were retained. Conserved CDS alignments were extracted by using Gblocks [54] and were concatenated to generate a supermatrix. 4DTV sites were extracted from this supermatrix and subject to analysis with RAxML (v.8.2.0) [55] to generate the

599 maximum likelihood tree with the GTR+I+Γ model. Divergence times were estimated
600 using the MCMCTree program in the PAML4.7 package [56].
601 To explore the evolution of gene families in these cereal crops, orthologous groups
602 were constructed with OrthoFinder2 (v2.2.7) [57] using default settings based on the
603 all-vs-all BLASTP results of the 52 subgenomes. The genes that were exclusively
604 found in each subfamily (>60% species presented) were identified. Significantly
605 overrepresented GO terms in each group were identified using the R package
606 "topGO".
607 To investigate the chromosomal evolution of the *Avena* genomes, representative
608 species from five subfamilies in Poaceae with chromosome-level genome assemblies
609 were selected. The synteny between these extant genomes and the reconstructed
610 AGKs [22] was identified using MCScanX [58] with the default settings, and the identified
611 syntenic blocks were then used to deduce the homologous relationship between AGK
612 marker genes and the protein sequences of *Avena* and the related cereal crop species.
613 The collinearity between species was identified and plotted using MCScanX (python
614 version).
615 **The evolution and allopolyploidization history of oat**
616 For whole-genome sequencing, 14 other *Avena* accessions were chosen, including six
617 A genome diploids: *A. canariensis* (CN 23017, Ac), *A. damascena* (CN 19457, Ad), *A.*
618 *longiglumis* (CN58138, Al), *A. strigosa* (CN 88610, As), *A. wiestii* (CN 90217, As)
619 and *A. nuda* (CN 58062, As); two C genome diploids: *A. clauda* (CN 21388, Cp) and
620 *A. ventricosa* (CN 21405, Cv); two AB genome tetraploids: *A. barbata* (CN 65538)
621 and *A. agadiriana* (CN 25869); three CD genome tetraploids: *A. insularis* (INS-4), *A.*
622 *maroccana* (CN 57945), *A. murphyi* (CN 21989); and one ACD genome hexaploid:
623 *A. sativa* (Ogle). These accessions represent all genome types found among extant

624 *Avena* species. All sequencing was performed with an Illumina HiSeq X-Ten
625 instrument, using 400 bp paired-end libraries.
626 Raw reads of the 14 newly sequenced as well as *A. longiglumis* and *A. insularis*
627 were trimmed using Trimmomatic v 0.40 [59] and mapped onto the reference SFS
628 genome using BWA [60]. The genomic variants were called based on the mapped short
629 paired-end reads using the call function of BCFtools [61], and the obtained variants were
630 further filtered. The qualified SNPs were used as the RAxML input for constructing
631 the phylogenetic relationships using maximum likelihood and 100 bootstrap replicates.
632 The identity between SFS and each sequenced accession was calculated by mapping
633 ~1X clean reads of each accession to the reference SFS genome.
634 All sequenced diploid accessions were further subjected to transcriptome
635 sequencing. For this purpose, RNA was isolated from each accession independently
636 from a combination of sample from seven tissues/conditions, including seedlings (two
637 weeks old), flag leaves at the booting (Zodoks 45) and heading (Zodoks 58) stages,
638 panicles at the booting (Zodoks 45), heading (Zodoks 50, 58) and grain dough
639 (Zodoks 83) stages. Then, the seven types of RNA samples were mixed in equal
640 amounts and sequenced in paired-end, 150-bp reads on an MGI system. The raw reads
641 were cleaned with Trimmomatic. De novo assembly was performed using Trinity with
642 the default settings. Coding sequences (CDSs) were predicted using TransDecoder [62].
643 Each proteome from these diploid species and five *Avena* reference genomes was
644 subjected to BLAST searches against *Hordeum vulgare* according to an E-value ≤
645 1e-5. The RBHs in each pair were obtained and the gene families present in all the
646 evaluated species were retained. Phylogenetic tree inference and divergence time
647 estimation were conducted using the same methods described above.
648 The chloroplast genomes of all sequenced taxa were assembled using the clean

649 short reads with NOVOPlasty [63]. Another 26 *Avena* chloroplast genomes previously
650 published by Fu *et al*. [24] and Liu *et al*. [25] were also downloaded (Supplementary Table
651 15). Multiple sequence alignments were performed using MUSCLE [64], and the
652 identified informative sites were used for phylogenetic tree construction using
653 RAxML with 100 bootstrap replicates under the GTR+I+Γ evolutionary model, where
654 the chloroplast genome sequence from *Triticum aestivum* was used as the outgroup.
655 Divergence times were estimated according to independent rates and JC69 models
656 using the MCMCTree program in the PAML4.7 package.

657 **Synteny and comparative genomics**

658 The subgenome synteny between SFS and *A. insularis* was analysed by plotting the
659 positions of homologous pairs in the subgenome pairs within the context of 21 and 14
660 chromosomes using the Circos [65] software (Extended Data Fig. 6a, b). The
661 chromosome painting of SFS chromosomes with *A. insularis* and *A. longiglumis*, as
662 well those of *A. insularis* chromosomes with *A. longiglumis* and *A. eriantha* were
663 performed by aligning genomic chunks (100 bp markers) to their potential ancestors
664 using BWA with the default settings. The uniquely mapped markers were retained. We
665 then processed the markers on each chromosome by requiring at least five consecutive
666 markers supporting homology to the same chromosome. We consolidated each group
667 of five consecutive potential markers as one confirmed block. These confirmed blocks
668 with less than 2 Mb were further consolidated as superblocks. The total lengths of the
669 subgenomes of *A. insularis* and SFS inherited from the ancestral genomes were
670 calculated by summing the lengths of superblocks.

671 The interchromosomal exchanges between *A. insularis* and SFS after
672 polyploidization was also analysed by individually mapping reads from *A.*
673 *longiglumis* and *A. eriantha* to the *A. insularis* reference genome and reads from *A.*

longiglumis, *A. eriantha* and *A. insularis* to the SFS reference genome. The single-base depth coverage of the properly paired reads obtained from the *A. longiglumis*, *A. eriantha* and *A. insularis* mapping was calculated using the Mosdepth program. The median depth within a sliding window (window size: 1 Mb, step size: 0.5 Mb) was calculated and plotted along with the chromosomes of the reference genome (Fig. 4c, Extended Data Fig. 7a-d).

FISH analysis

Major interchromosomal exchanges between the C and D subgenomes of *A. insularis* and SFS were detected by using FISH technology with the C genome-specific repeat Am1 as the probe. Metaphase chromosome preparation [66] and FISH analysis [67] were performed as previously described (Fig. 4d, Extended Data Fig. 7e).

***Ka/Ks* analysis**

One-to-one orthologous gene sets among the genome assemblies for *Hordeum vulgare*, the A and C diploid progenitors, *A. longiglumis* and *A. eriantha*, and the subgenomes of *A. insularis* and SFS were identified using OrthoFinder2 [57]. A total of 2,767 orthologous gene sets were obtained and then used for nonsynonymous (*Ka*) and synonymous (*Ks*) rate calculations. For this purpose, the homoeologous gene pair list was used as input and the protein sequences from each gene pair were aligned using MUSCLE [64]. PAL2NAL [68] was used to convert the peptide alignment to a nucleotide alignment and *Ka* and *Ks* values were computed between gene pairs by using Codeml from PAML4.7 in free-ration mode. All estimates with $Ks<0.01$ were excluded from the analysis. The significance of the differences in *Ka/Ks* values between genomes (subgenomes) was estimated using the Wilcoxon rank-sum test for nonnormal distributions in R.

Full-length long terminal repeat (LTR) retrotransposon analysis

Full-length LTRs were identified using LTR_FINDER and classified using hierarchical methods (Supplementary Note) (Extended Data Fig. 8b). To estimate the insertion times of full-length LTRs, the 5'- and 3'-LTR sequences of the full-length LTRs were aligned and used to calculate divergence using distmat [69]. The insertion times were estimated with the formula T=K/2r, where r represents the neutral mutation rate, which is 1.3×10^{-8} mutations per site per year [70] (Extended Data Fig. 8d).

Gene family size comparisons between the SFS subgenomes and the A and C diploid genomes

To evaluate the relationships between gene family sizes in the hexaploid subgenomes and the diploid genomes of *A. longiglumis* and *A. eriantha*, gene family were identified through OrthoFinder2 [57]. The gene family sizes were compared between each diploid genome and each of the SFS subgenomes. The scatter dots and regression lines were plotted (Extended Data Figure 9a, b).

Gene loss and retention

Orthologues between *A. eriantha* and the C subgenome of *A. insularis* were identified using RBH-based methods. A sliding window approach with a window size of 100 genes and a step size of 10 genes by using the *A. eriantha* genome as the reference was employed to calculate the percentage of retained genes in the C and D subgenomes individually and in *A. insularis* (Extended Data Fig. 9e). The gene retention rates of the subgenomes of SFS were also estimated and plotted using the same methods (Extended Data Fig. 9f).

Gene expression analysis

For gene expression analyses, RNA from SFS samples from seven tissues/conditions

(described above) of SFS was individually isolated and sequenced with three biological replicates using an MGISEQ2000 sequencer. In addition, two-week-old seedlings of SFS were exposed to six different abiotic stresses: heat, cold, drought, waterlogging, alkaline and salt. RNA was isolated from seedlings after stress treatment for one week and sequenced with three biological replicates.

Gene expression levels in each sample were quantified using the HiSAT2-HTSeq pipeline. Differentially expressed genes (DEGs) between stress and control conditions were detected with the edgeR software package [71] (FDR < 0.05 and |log2-fold change (FC)| > 0.5). To analyse differences in the expression patterns of homoeologous genes, we undertook an initial analysis of expression variation among strict single-copy homoelogous triplets. To this end, we used MCScanX [58] to detect syntenic blocks, and 13,744 orthologous triads with a single gene copy per subgenome were identified. Triplet expression vectors were created by concatenating the observed gene expression values for the A, C, and D homoeologs. Triplets that expressed at least one homolog across the sampled tissues were summarized in a triplet expression matrix. The expression values of the triplet expression matrix were log10 transformed ($\log_{10}$(TPM+1)), and the matrix was subjected to two-dimensional hierarchical clustering using “hclust” implemented in R with the “average” correlation distance and clustering. Heatmap visualization was performed using the heatmap.2 command from the R package gplots.

33 Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**, 737-746, doi:10.1101/gr.214270.116 (2017).

34 Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* **31**, 1119-1125,

748 doi:10.1038/nbt.2727 (2013).
749 35 Alonge, M. et al. RaGOO: fast and accurate reference-guided scaffolding of
750 draft genomes. Genome Biol 20, 224. doi:10.1186/s13059-019-1829-6 (2019).
751 36 Bekele, W. A., Wight, C. P., Chao, S., Howarth, C. J. & Tinker, N. A.
752 Haplotype-based genotyping-by-sequencing in oat genome research. *Plant*
753 *Biotechnol J* **16**, 1452-1463, doi: 10.1111/pbi.12888 (2018).
754 37 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov,
755 E. M. BUSCO: assessing genome assembly and annotation completeness with
756 single-copy orthologs. *Bioinformatics* **31**, 3210-3212,
757 doi:10.1093/bioinformatics/btv351 (2015).
758 38 Wang, X. & Wang, L. GMATA: An integrated software package for
759 genome-scale SSR mining, marker development and viewing. *Front Plant Sci*
760 **7**, doi:10.3389/fpls.2016.01350 (2016).
761 39 Benson, G. Tandem repeats finder: a program to analyze DNA sequences.
762 *Nucleic Acids Res* **27**, 573-580, doi:10.1093/nar/27.2.573 (1999).
763 40 Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature
764 inverted-repeat transposable elements from genomic sequences. *Nucleic Acids*
765 *Res* **38**, e199-e199, doi:10.1093/nar/gkq862 (2010).
766 41 Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of
767 full-length LTR retrotransposons. *Nucleic Acids Res* **35**, W265-W268,
768 doi:10.1093/nar/gkm286 (2007).
769 42 Bedell, J. A., Korf, I. & Gish, W. MaskerAid: a performance enhancement to
770 RepeatMasker. *Bioinformatics* **16**, 1040-1041,
771 doi:10.1093/bioinformatics/16.11.1040 (2000).
772 43 Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements.

Cytogenet Genome Res **110**, 462-467, doi:10.1159/000084979 (2005).

44 Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res* **43**, e78, doi:10.1093/nar/gkv227 (2015).

45 Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res* **42**, e119-e119, doi:10.1093/nar/gku557 (2014).

46 Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637-644, doi:10.1093/bioinformatics/btn013 (2008).

47 Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7, doi:10.1186/gb-2008-9-1-r7 (2008).

48 Urasaki, N. *et al.* Draft genome sequence of bitter gourd (*Momordica charantia*), a vegetable and medicinal plant in tropical and subtropical regions. *DNA Res* **24**, 51-58, doi:10.1093/dnares/dsw047 (2017).

49 Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935, doi:10.1093/bioinformatics/btt509 (2013).

50 Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**, D121-D124, doi:10.1093/nar/gki081 (2005).

51 Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**, 3100-3108, doi:10.1093/nar/gkm160 (2007).

52 Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection

798 of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964,
799 doi:10.1093/nar/25.5.955 (1997).
800 53 Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic*
801 *Acids Res* **37**, D211-D215, doi:10.1093/nar/gkn785 (2009).
802 54 Talavera, G. & Castresana, J. Improvement of phylogenies after removing
803 divergent and ambiguously aligned blocks from protein sequence alignments.
804 *Syst Biol* **56**, 564-577, doi:10.1080/10635150701472164 (2007).
805 55 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and
806 post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313,
807 doi:10.1093/bioinformatics/btu033 (2014).
808 56 Arenas, M., Sánchez-Cobos, A. & Bastolla, U. Maximum-likelihood
809 phylogenetic inference with selection on protein folding stability. *Mol Biol*
810 *Evol* **32**, 2195-2207, doi:10.1093/molbev/msv085 (2015).
811 57 Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for
812 comparative genomics. *Genome Biol* **20**, 238, doi:10.1186/s13059-019-1832-y
813 (2019).
814 58 Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of
815 gene synteny and collinearity. *Nucleic Acids Res* **40**, e49,
816 doi:10.1093/nar/gkr1293 (2012).
817 59 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for
818 Illumina sequence data. *Bioinformatics* **30**, 2114-2120,
819 doi:10.1093/bioinformatics/btu170 (2014).
820 60 Li, H. & Durbin, R. Fast and accurate long-read alignment with
821 Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595,
822 doi:10.1093/bioinformatics/btp698 (2010).

823 61 Li, H. *et al.* The sequence alignment/map format and SAMtools.
824 *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
825 62 Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq
826 using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**,
827 1494-1512, doi:10.1038/nprot.2013.084 (2013).
828 63 Dierckxsens, N., Mardulyn, P. & Smits, G. NOVOPlasty: de novo assembly of
829 organelle genomes from whole genome data. *Nucleic Acids Res* **45**, e18,
830 doi:10.1093/nar/gkw955 (2017).
831 64 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and
832 high throughput. *Nucleic Acids Res* **32**, 1792-1797, doi:10.1093/nar/gkh340
833 (2004).
834 65 Krzywinski, M. *et al.* Circos: an information aesthetic for comparative
835 genomics. *Genome Res* **19**, 1639-1645, doi:10.1101/gr.092759.109 (2009).
836 66 Yan, H. *et al.* New evidence confirming the CD genomic constitutions of the
837 tetraploid *Avena* species in the section Pachycarpa Baum. *PloS One* **16**,
838 e0240703, doi:10.1371/journal.pone.0240703 (2021).
839 67 Fu, S. *et al.* Oligonucleotide probes for ND-FISH analysis to identify rye and
840 wheat chromosomes. *Sci Rep* **5**, 10552, doi:10.1038/srep10552 (2015).
841 68 Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein
842 sequence alignments into the corresponding codon alignments. *Nucleic Acids*
843 *Res* **34**, W609-612, doi:10.1093/nar/gkl315 (2006).
844 69 Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular
845 biology open software suite. *Trends Genet* **16**, 276-277,
846 doi:10.1016/s0168-9525(00)02024-2 (2000).
847 70 Wicker, T. *et al*. Impact of transposable elements on genome structure and

evolution in bread wheat. *Genome Biol* **19**, 103, doi:10.1186/s13059-018-1479-0 (2018).

71 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2009).

Acknowledgements This research was supported by China Agriculture Research System of MOF and MARA (CARS07), the National Natural Science Foundation of China (32072025, 31801430), the Science and Technology Development Program of Jilin Province (20200402034NC) and Talent Fund Project of Jilin Province. The authors thank Mawsheng Chern, Department of Plant Pathology and the Genome Center, University of California, Davis, California, USA, for improving the writing of this article. We also thank Agriculture & Agri-Food Canada (AAFC) and Dr. Eric N. Jellen, Brigham Young University, Provo, Utah, USA, for providing the *Avena* materials.

Author contributions C.R., Yuanying Peng, F.L. and Y.W. conceived the study. C.R., Yuanying Peng and H.Y. provided funding. C.R., Yuanying Peng, H.Y., L.G., P.Z., C.W., J.Z., Yun Peng, D.D. and L.W. collected and prepared the tissue samples for sequencing. Yuanying Peng, C.D., H.Y., L.K. and F.L. led the bioinformatics analyses. P.Z., K.Y., C.D., Y.X. and Yun Peng conducted transcriptome sequencing and analysis. K.Y and X.D. constructed database. X.L. and Ying Li conducted the FISH validation of chromosome translocation. J.M., M.H. and Yan Li collected the pictures of spikes. C.D., Yuanying Peng and H.Y. developed the figures. Yuanying Peng, H.Y., L.G., and C.D. drafted the manuscript, Q.J., J.W., L.W., W.L., H.K., Y.W., Y.Z. contributed to the writing. Yuanying Peng, H.Y., L.G., C.D. and L.K. contributed equally.

Competing interests The authors declare no competing interests.

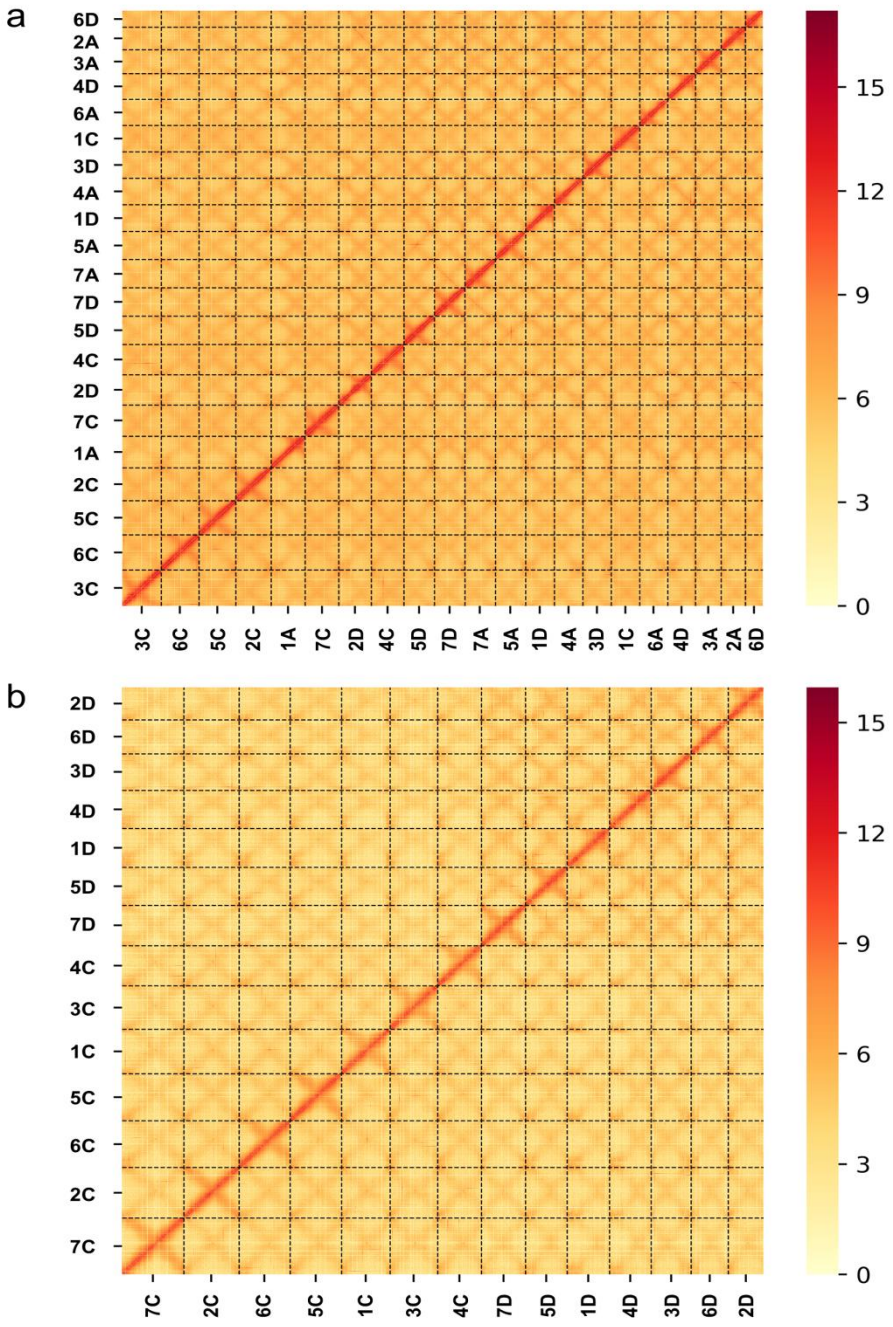
Additional Information

Supplementary Information is available for this paper.

Correspondence and requests for materials should be addressed to R.C. (renchangzhong@163.com) or Yuanying Peng (yy.peng@hotmail.com).

Data availability

The genome assemblies and sequence data for *A. sativa* ssp. *nuda* cv. Sanfensan, *A. insularis* (CN 108634) and *A. longiglumis* (CN 58139) were deposited at NCBI under BioProject codes PRJNA 727473, PRJNA731599 and PRJNA716144, respectively. Sanfensan genome assembly (SAMN19770945), ONT data (SAMN19021785), Hi-C data (SAMN19340419), NGS data (SAMN19582572), Iso-seq data (SAMN19581880) and RNA-seq data (SAMN19582573, SAMN19582574); *A.insularis* genome assembly (SAMN19771048), ONT data (SAMN19291344), Hi-C data (SAMN19312172), NGS data (SAMN19579880) and Iso-seq data (SAMN19581879); *A.longiglumis* genome assembly (SAMN19771099), ONT data (SAMN18395928), NGS data (SAMN19523931) and Iso-seq data (SAMN19581877). All raw data for the other 14 deep-sequenced accessions including 8 diploids, 5 tetraploids and 1 hexaploid are available under projected numbers can be found in Supplementary Table 1. All other data are available from the corresponding author upon reasonable request.



Extended Data Fig. 1 | Hi-C contact map of *Avena* polyploid species. Abundant intrachromosomal contacts were observed. Chromosomes are sorted by size; interchromosomal contacts were also found but with a much lower intensity. **a**, The 21 chromosomes of hexaploid SFS. **b**, The 14 chromosomes of tetraploid *A. insularis*.

a

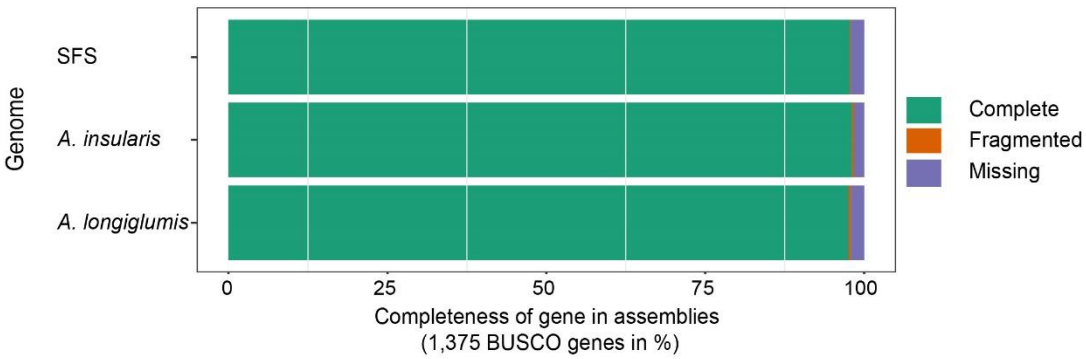
A. insularis

Linkage Group	18	33	23	20	24	5	12	28	13	15	9	3	17	11	1	8	19	21	6	4	2
Total	326	180	209	456	348	317	343	276	454	569	643	745	715	550	694	505	286	552	437	325	559
1C	248	0	63	1	1	4	3	246	0	11	9	8	5	3	26	2	2	2	3	1	1
2C	1	15	1	5	0	14	1	0	403	6	3	8	4	17	1	12	5	2	134	1	2
3C	2	0	9	11	1	1	1	0	2	384	4	1	4	15	1	1	3	1	3	0	1
4C	2	0	1	31	7	2	2	0	2	2	538	8	4	3	2	0	0	14	0	0	8
5C	1	0	2	6	32	7	18	4	1	2	11	677	2	4	0	0	1	2	10	0	23
6C	14	4	0	0	1	24	0	0	1	0	9	6	647	2	1	14	0	3	1	16	7
7C	1	2	1	12	11	69	16	3	3	1	2	10	5	475	7	5	0	0	8	0	6
1D	28	9	9	50	64	10	74	8	4	2	3	0	2	8	592	5	17	4	1	4	4
2D	5	86	5	4	1	28	5	1	12	1	2	1	0	5	4	188	0	4	3	3	5
3D	0	2	67	25	16	7	13	0	2	13	2	1	1	1	6	4	236	1	3	0	6
4D	1	4	4	278	7	9	3	0	1	4	23	3	0	2	7	3	6	501	13	2	13
5D	0	2	38	7	166	9	4	6	2	132	3	6	0	2	8	4	2	1	228	3	6
6D	20	3	3	8	3	124	10	1	1	3	3	1	18	1	11	249	0	0	2	287	7
7D	1	49	3	7	5	4	188	1	0	1	5	9	9	3	7	8	0	7	1	2	459

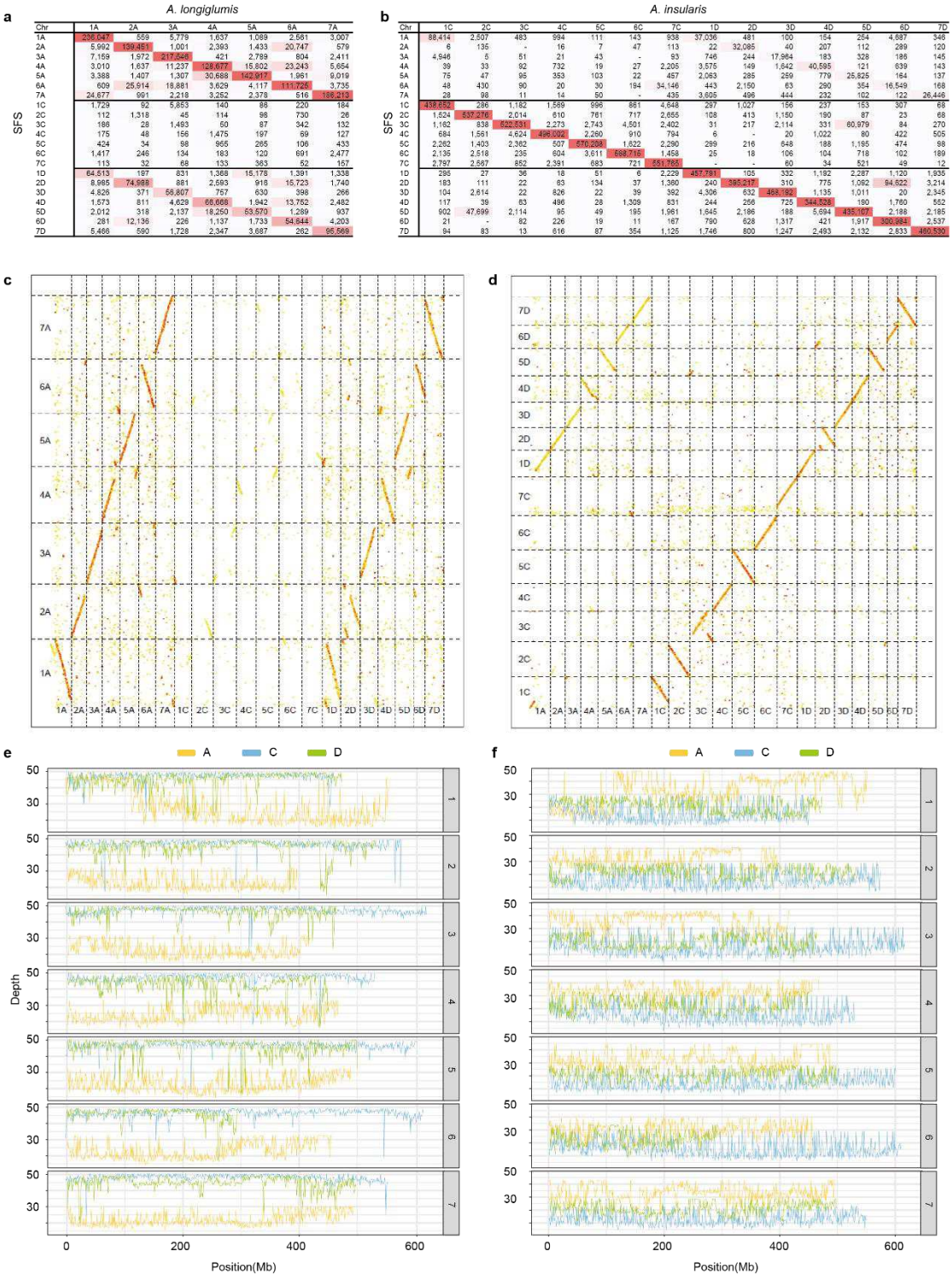
SFS

Linkage Group	18	33	23	20	24	5	12	28	13	15	9	3	17	11	1	8	19	21	6	4	2
Total	359	261	316	638	457	432	491	301	477	598	699	779	751	575	777	548	302	620	481	341	599
1A	320	1	1	3	1	0	1	15	2	0	10	2	2	2	126	34	3	1	0	0	6
2A	0	147	0	1	0	1	1	0	3	1	0	0	0	0	1	12	0	0	0	0	1
3A	0	0	258	1	2	2	2	11	0	8	1	0	0	0	1	2	4	0	0	0	2
4A	0	1	0	517	2	3	3	0	1	3	8	0	0	3	14	2	2	54	2	1	2
5A	0	0	1	3	272	4	1	0	1	0	1	4	0	1	9	0	0	1	51	0	2
6A	0	0	0	0	0	345	1	1	3	0	0	1	6	21	3	3	0	2	5	45	0
7A	1	0	2	1	1	2	397	0	0	0	0	5	0	0	18	0	6	0	0	0	30
1C	2	0	25	0	0	0	0	262	0	4	13	9	0	0	1	1	0	1	2	0	1
2C	1	4	1	0	0	3	0	0	442	1	3	4	2	0	0	3	0	1	0	0	0
3C	0	0	7	5	0	0	1	3	1	564	2	2	2	19	0	1	18	2	0	0	2
4C	0	0	0	11	1	0	1	0	0	4	598	1	0	2	2	0	0	6	0	0	3
5C	3	0	0	0	11	4	8	3	0	0	2	735	0	0	0	0	0	3	4	0	16
6C	5	2	0	0	0	9	0	0	1	2	1	4	710	2	0	9	0	0	0	13	3
7C	0	0	1	2	5	0	6	0	0	0	0	0	1	292	0	1	0	0	3	0	2
1D	7	2	0	14	24	0	15	2	1	2	1	0	0	3	585	3	12	1	0	0	0
2D	9	13	1	1	0	6	2	0	10	0	0	3	1	1	4	469	0	0	0	0	2
3D	0	0	16	5	1	1	6	0	0	3	1	0	1	0	3	1	256	1	0	0	2
4D	1	2	1	66	1	2	0	0	1	2	24	1	1	0	0	1	0	535	0	1	1
5D	8	0	0	0	123	4	1	3	2	1	0	2	0	5	0	1	0	0	412	1	1
6D	0	0	0	0	0	13	1	0	1	2	0	0	5	0	0	1	0	0	0	279	3
7D	1	13	0	2	1	0	42	0	0	0	0	2	6	2	4	4	0	1	1	0	472

b



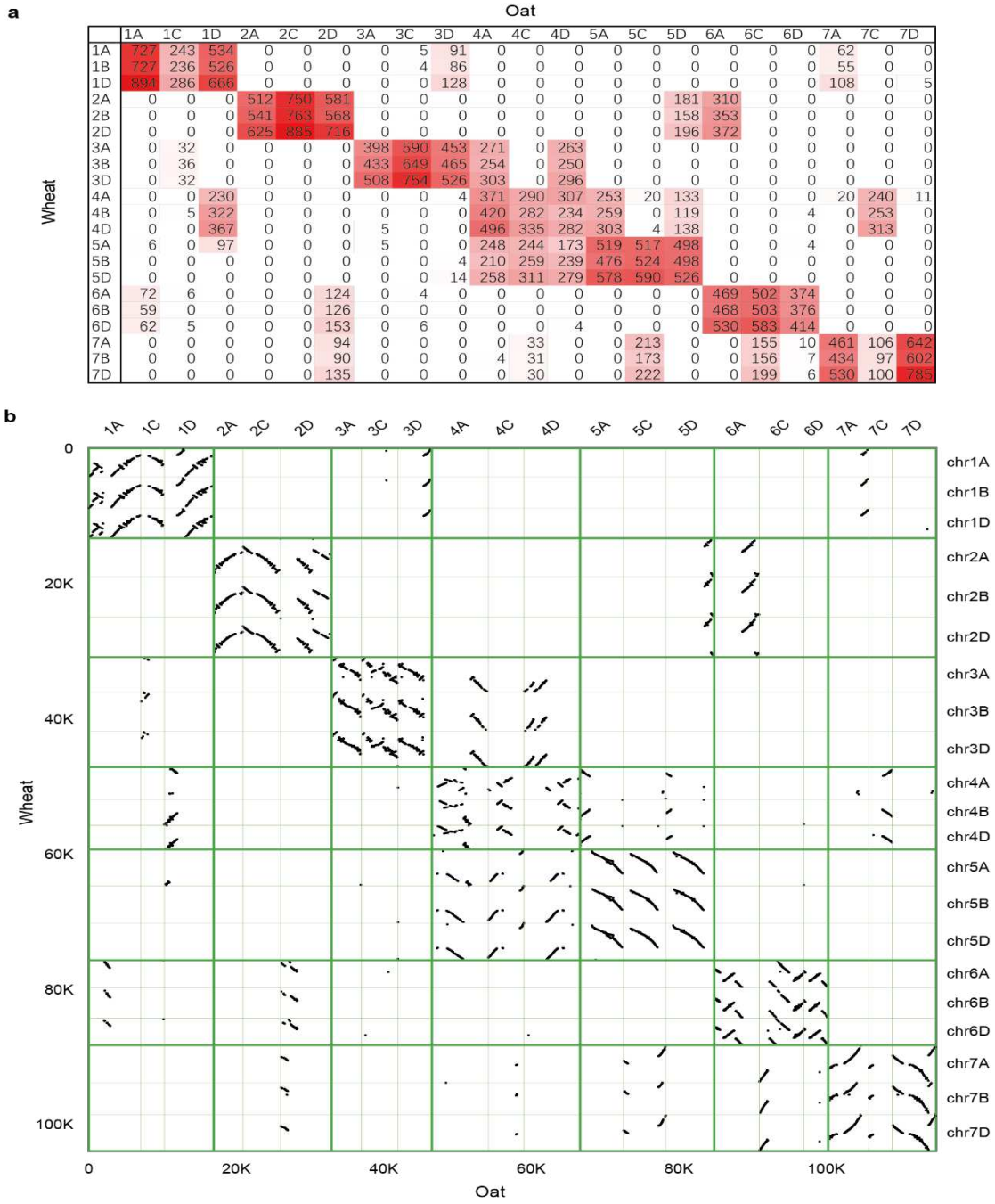
Extended Data Fig. 2 | Assembly validation. a, The final assemblies were evaluated using the consensus hexaploid linkage map. Markers from each of the linkage groups of the hexaploid consensus linkage map anchored to the assembled SFS (top) and *A. insularis* (bottom) chromosomes. **b**, Completeness of the three assembled genomes as assessed by BUSCO. Bar charts show the percentages of 1,375 highly conserved plant BUSCO genes that are completely present, fragmented or missing in the assembly.



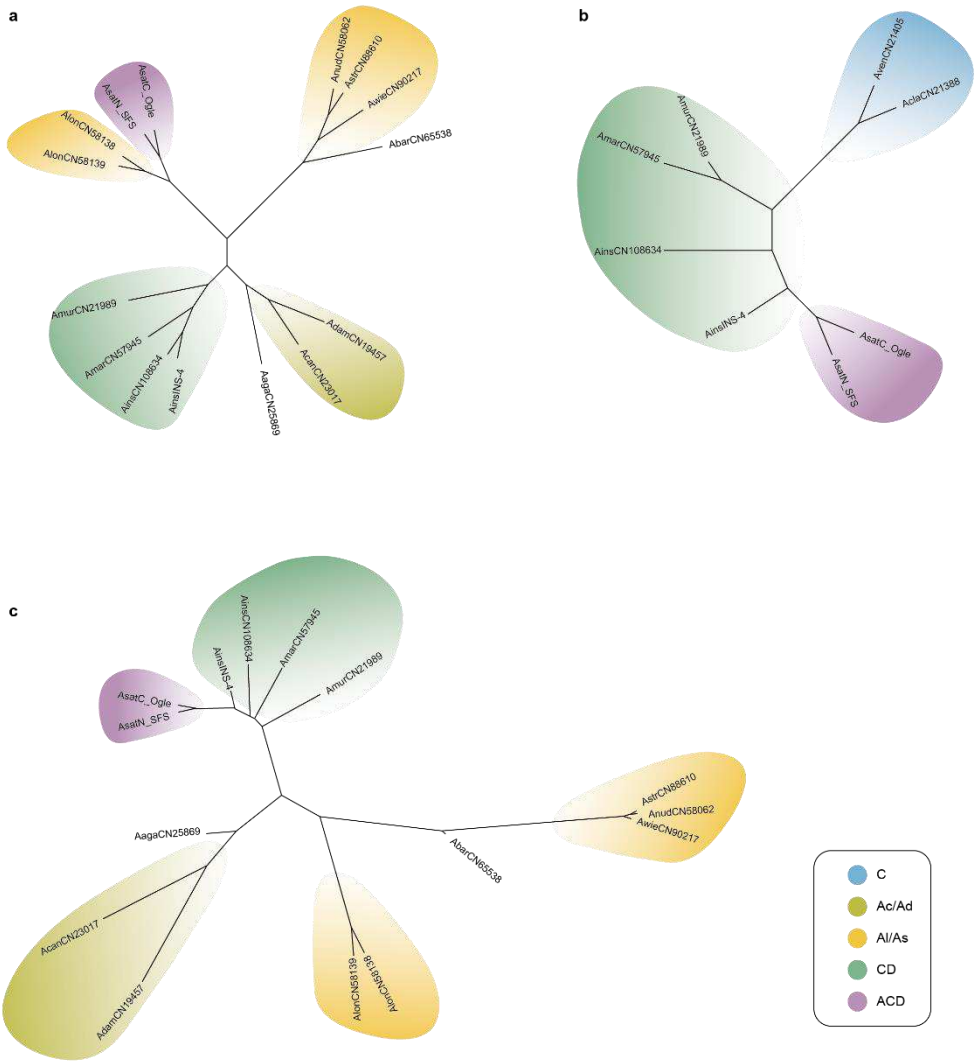
908

909 **Extended Data Fig. 3 | Subgenome assignment of hexaploid oat.** The genome
910 sequences of *A. longiglumis* and *A. insularis* were divided into 100 bp markers and
911 then aligned with the reference SFS genome sequence. Unique mapped markers were
912 retained. **a,** The greatest numbers of markers shared by the chromosomes of *A.*
913 *longiglumis* and SFS are highlighted. **b**, The greatest number of markers shared by the

914 chromosomes of *A. insularis* and SFS are highlighted. **c,** Dot plots display the degrees
915 of synteny between *A. longiglumis* and SFS. **d,** Dot plots display the degrees of
916 synteny between *A. insularis* and SFS. Each dot represents a syntenic block of five or
917 more 100 bp markers. The distance between each pair of adjacent markers is less than
918 200 kb. **e-f,** Coverage depth obtained along the SFS chromosomes after mapping
919 Illumina sequence reads from *A. longiglumis* (e) and *A. insularis* (f) to the reference
920 genome of SFS.

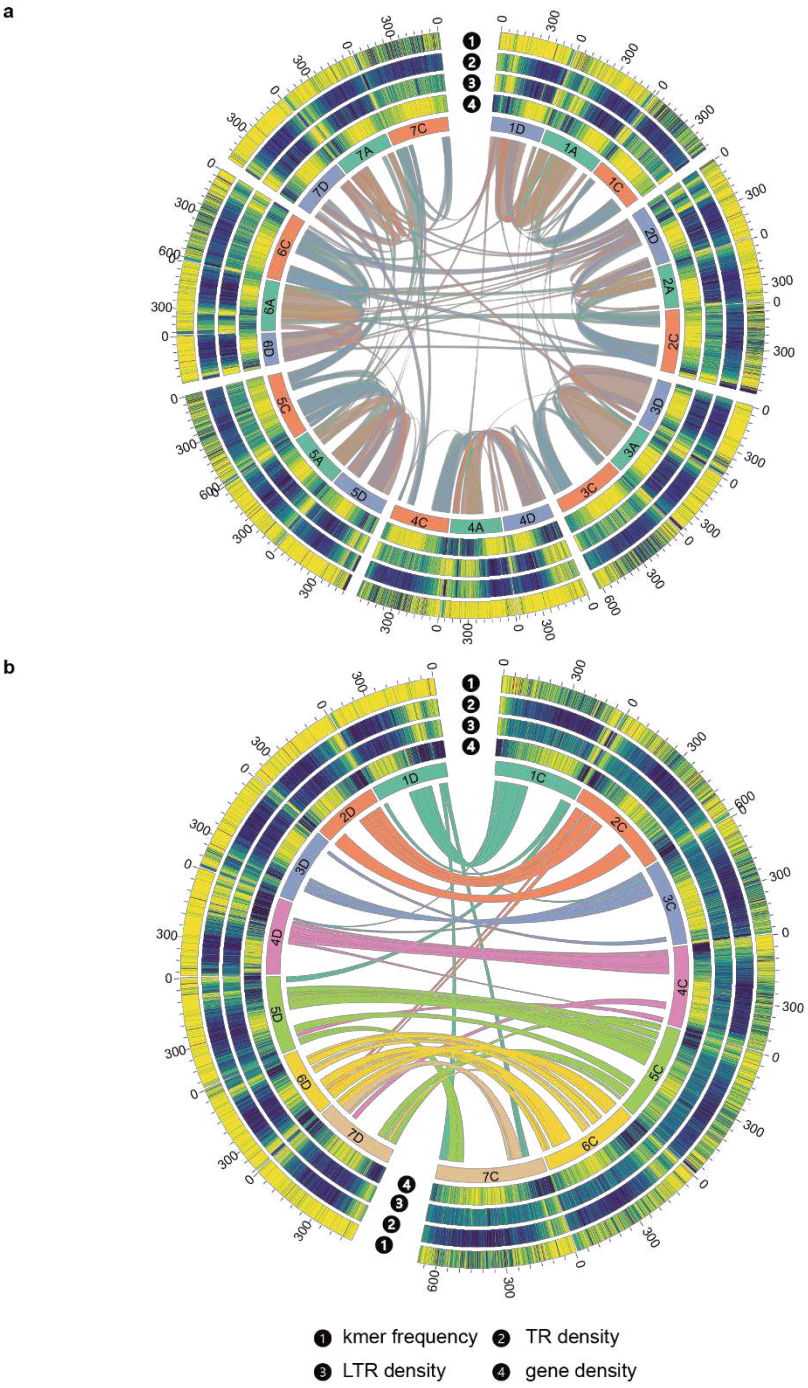


Extended Data Fig. 4 | Standard nomenclature for SFS chromosomes. The assignment of SFS chromosomes was performed based on synteny with the chromosomes of bread wheat to maintain consistency with future assemblies. Chromosome names specify homologous groups (1 to 7) and subgenomes (A/C/D). **a**, The matrix shows the number of conserved protein-coding genes between each pair of SFS and bread wheat chromosomes. **b**, Dot plots display the degrees of synteny between SFS and bread wheat. Each dot represents a syntenic block of five or more genes.



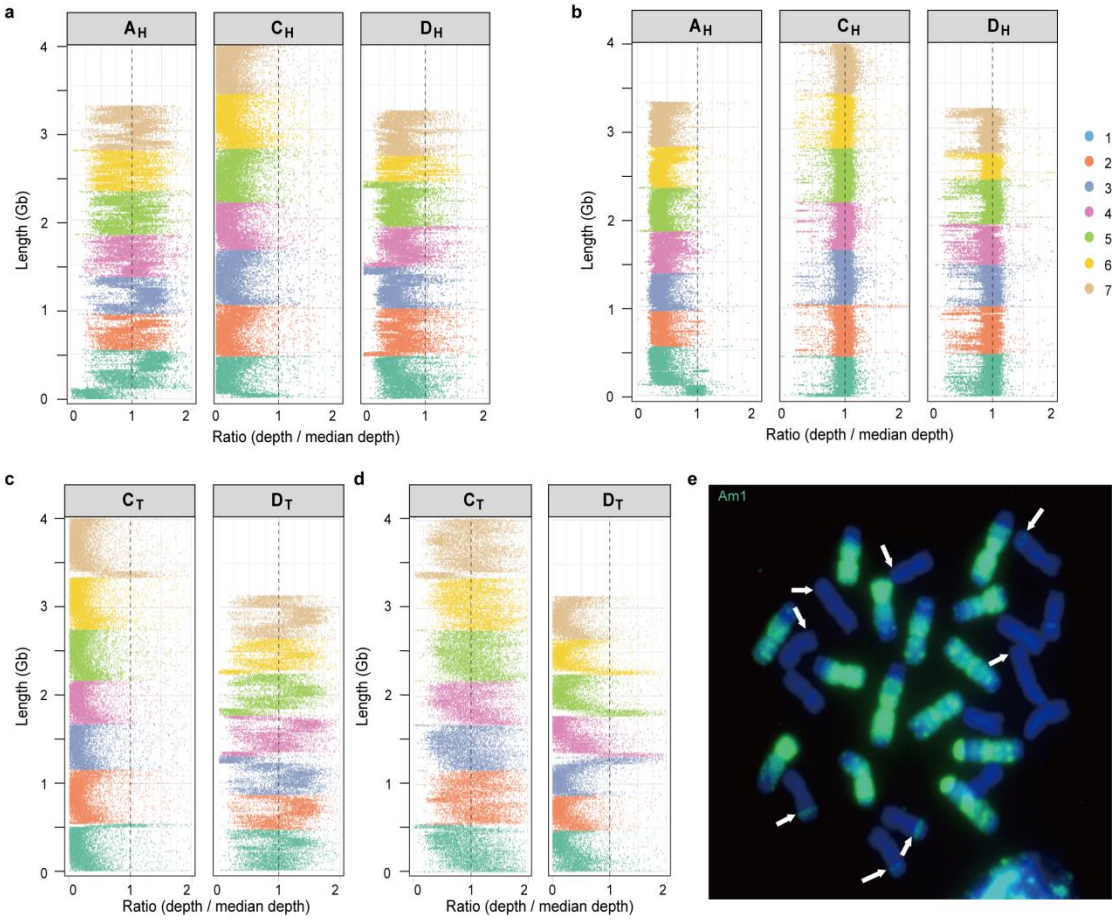
930

931 **Extended Data Fig. 5 | Phylogenetic analyses of different *Avena* genomes**. Clean
932 short paired reads from all the sequenced *Avena* taxa were aligned to the subgenomes
933 of SFS individually, and variants were called and filtered using bcftools. The resulting
934 SNPs based on three subgenomes of SFS were used for phylogenetic tree construction.
935 **a**, Phylogenetic tree generated from the SNPs of the A subgenome. **b**, Phylogenetic
936 tree generated from the SNPs of the C subgenome. **c**, Phylogenetic tree generated
937 from the SNPs of the D subgenome.



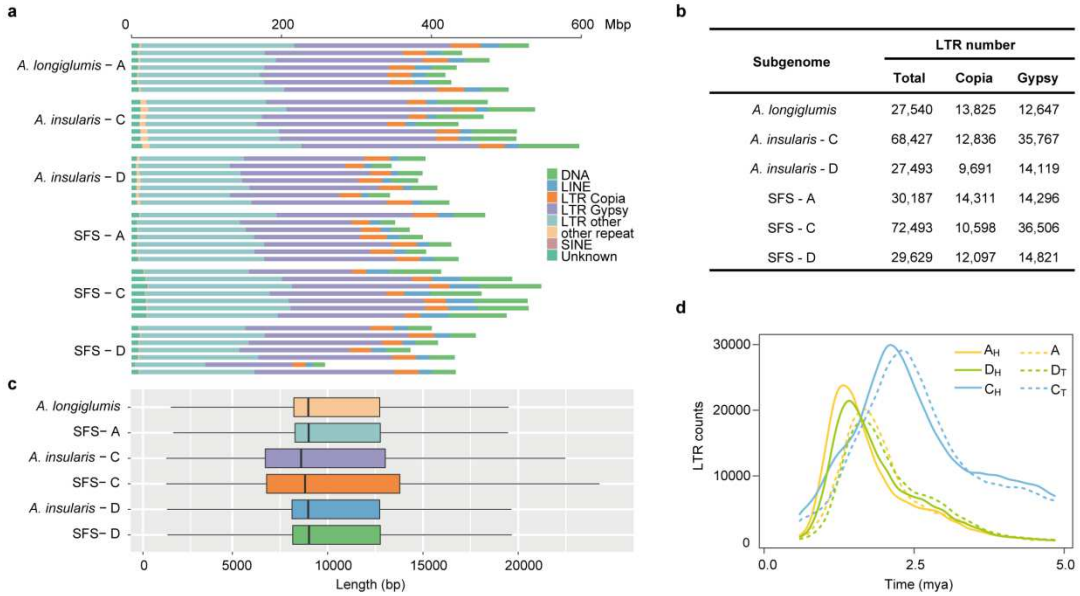
938

939 **Extended Data Fig. 6 | Structural and conserved synteny landscape of the *A.*
940 *insularis* and SFS subgenomes**. Homoeologous gene pairs in syntenic blocks are
941 linked. The rings depict the 31-mer distribution along each chromosome (1), the
942 density of tandem repeat (TR) (2), the density of long terminal repeats (LTRs) (3), and
943 the density of protein-coding genes (4). **a**, SFS. **b**, *A. insularis*.



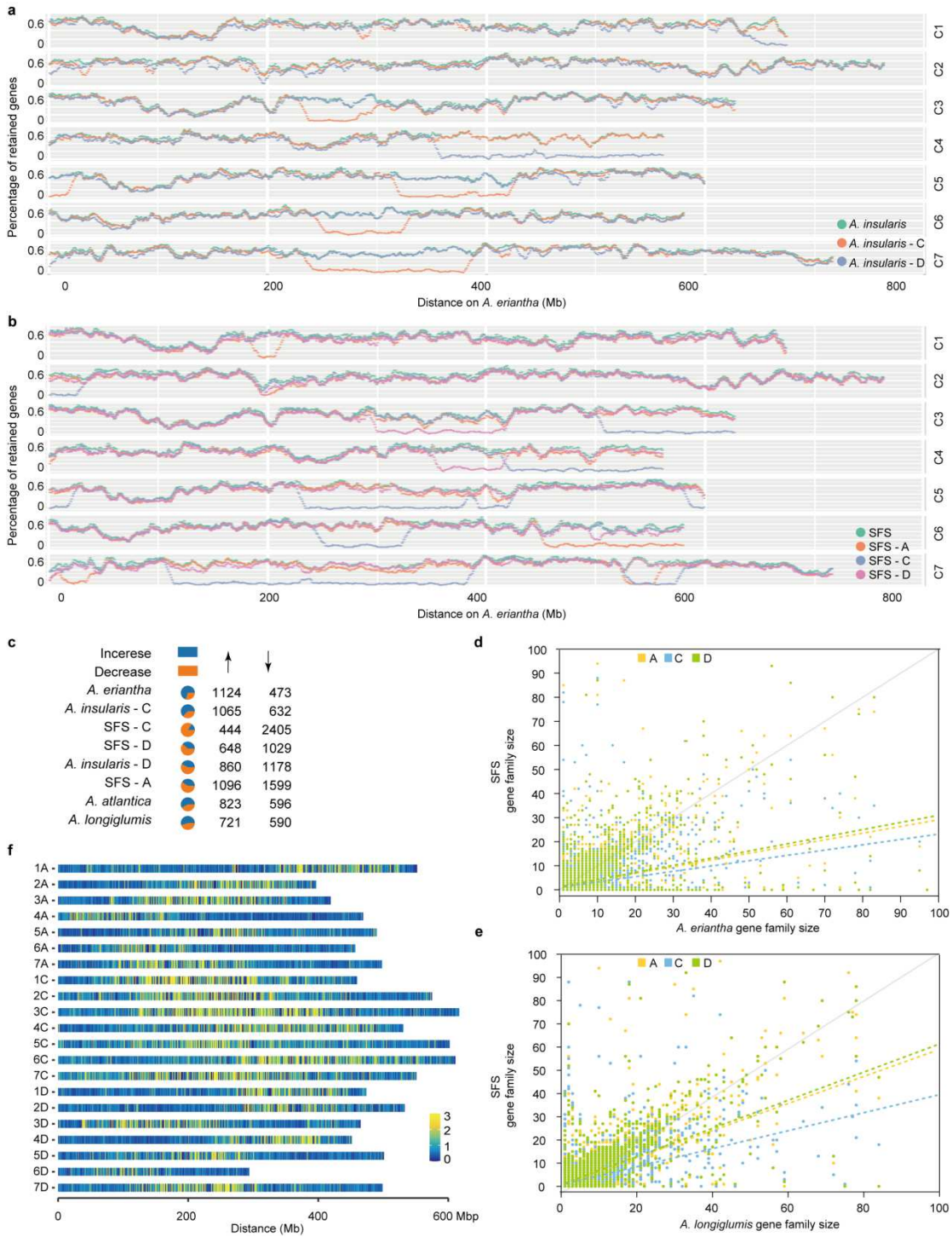
944

945 **Extended Data Fig. 7 | Intergenomic exchanges between the oat subgenomes. a**,
946 Reads of the A genome diploid mapped to the reference SFS genome. **b**, Reads of the
947 CD genome tetraploid mapped to the reference SFS genome. **c**, Reads of the A
948 genome diploid mapped to the reference *A. insularis* genome. **d**, Reads of the C
949 genome diploid mapped to the reference *A. insularis* genome reveal at least four large
950 C to D genomic exchanges. **e**, The major C to D translocations (indicated by white
951 arrows) in *A. insularis* were confirmed using FISH technology with the C
952 genome-specific repeat Am1 (green signals) as the probe.



Subgenome	LTR number		
	Total	Copia	Gypsy
A. longiglumis	27,540	13,825	12,647
A. insularis - C	68,427	12,836	35,767
A. insularis - D	27,493	9,691	14,119
SFS - A	30,187	14,311	14,296
SFS - C	72,493	10,598	36,506
SFS - D	29,629	12,097	14,821

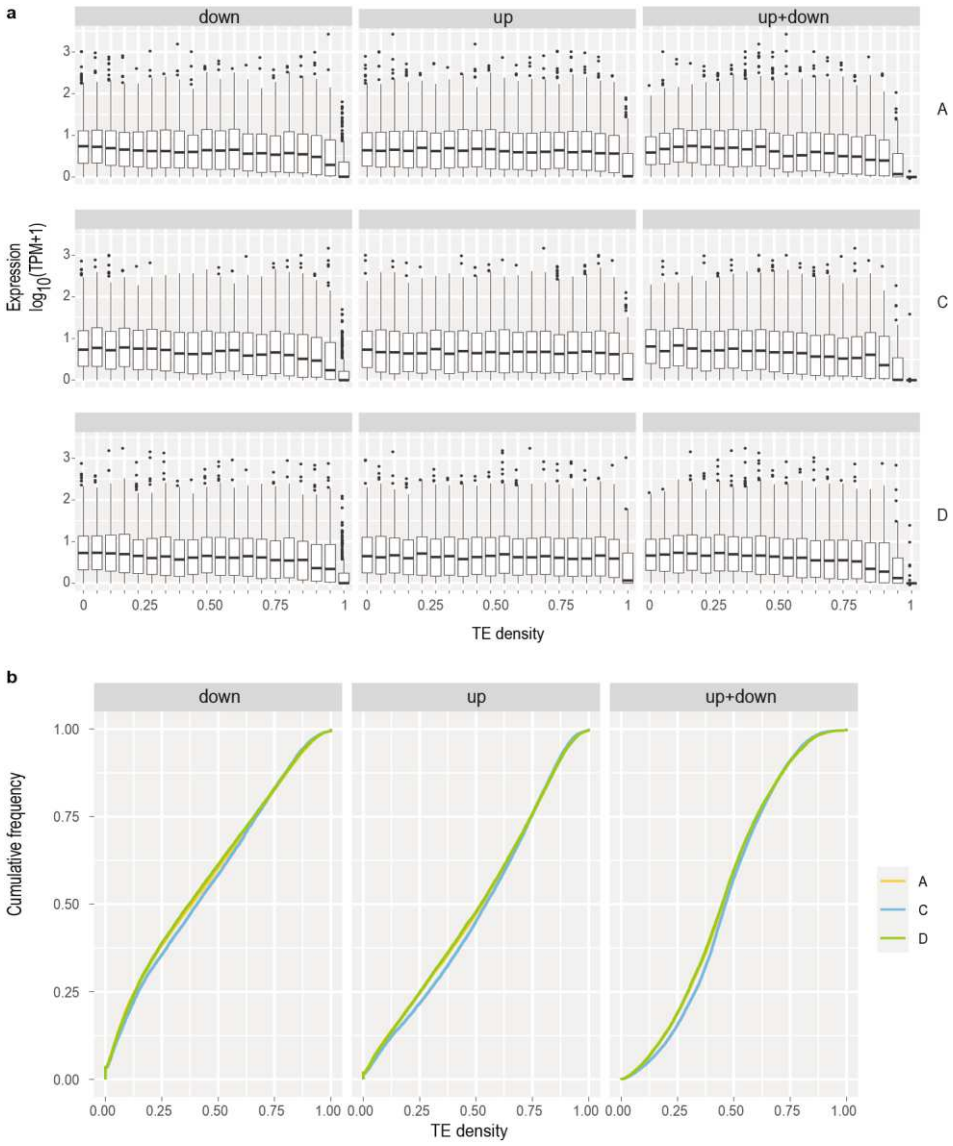
Extended Data Figure 8 | Analysis of TEs in *Avena* genomes. a, Genomic constituents of the subgenomes of SFS in comparison with those of the subgenomes of *A. insularis*, and the Al (*A. longiglumis*), As (*A. atlantica*), and Cp (*A. eriantha*) diploid genomes. **b**, Numbers of full-length long terminal repeats (LTRs) identified with LTR_FINDER in the subgenomes of SFS and *A. insularis*, as well as in the Al, As, and Cp genomes. **c**, Average sequence length of LTRs. **d**, Insertion times of full-length LTRs in *Avena* genomes.



962

963 **Extended Data Fig. 9 | Gene conservation and subgenome fractionation patterns**.
964 **a**, A sliding window approach with a window size of 100 syntenic genes and a step
965 size of 10 syntenic genes was used to show the percentages of retained genes in
966 subgenome C (orange), subgenome D (blue), and both (green) of *A. insularis* using
967 the *A. eriantha* (Cp) genome as a reference. **b**, A sliding window approach with a

window size of 100 syntenic genes and a step size of 10 syntenic genes was used to show the percentage of retained genes in subgenome A (orange), subgenome C (blue), subgenome D (pink), and all three (green) of SFS using the *A. eriantha* (Cp) genome as a reference. **c**, Number of expanded and contracted gene families in the subgenomes of SFS and *A. insularis* and the putative diploid progenitor genomes of *A. longiglumis* and *A. eriantha*. **d**, Relationships between gene family sizes in diploid *A. eriantha* and each subgenome of SFS. **e**, Relationship between gene family sizes in diploid *A. longiglumis* and each subgenome of SFS. The black line shows a 1:1 gene copy number relationship for SFS, *A. eriantha*, and *A. longiglumis*, and coloured lines show the regression fits for the observed gene family size in the subgenomes of SFS (colours: green, A subgenome; orange, D subgenome; blue, C subgenome). **f**, Distribution of the identified pseudogenes along the SFS chromosomes.



Extended Data Fig. 10 | Impact of TE density on gene expression in SFS. **a**, Gene expression is negatively correlated with TE density in SFS. Genes were grouped into twenty bins (5% increments in TE density). The mean log_{10}-transformed gene expression of each bin is shown by a boxplot. The central line in each box plot indicates the median. The top and bottom edges of the box indicate the first and third quartiles and the whiskers extended 1.5 times the interquartile range beyond the edges of the box. Black dots represent outliers. **b**, TE density was calculated in 5 kb up- and 5 kb downstream windows from the gene. TE densities near genes in the C subgenome are highest relative to homoeologs in the A and D subgenomes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables12720210627.xlsx
- SupplementaryInformation20210627.pdf
- SupplementaryInformation20210627.pdf