Supplementary Notes

Contents

Supplementary Notes

1. Genome sequencing and assembly

1.1. Plant material

The hulless hexaploid oat (*Avena sativa* L. ssp. *nuda*, 2n=6x=42, AACCDD) landrace cv. Sanfensan (abbreviated as SFS), the diploid species *A. longiglumis* (accession CN 58139, 2n=2x=14, AlAl) and the tetraploid species *A. insularis* (accession 108634, 2n=4x=28, CCDD) were chosen for whole-genome sequencing. Sanfensan is a traditional hulless variety that has a long cultivation history in Shanxi, China, which has been assumed to be the region of origin of hulless oat. *A. longiglumis* and *A. insularis* have been assumed to be the extant diploid and tetraploid species most closely related to hexaploid oat (**Supplementary Table 1**).

1.2. Short-read sequencing

High-quality genomic DNA was isolated from fresh leaf tissue using the Qiagen DNeasy Plant Mini Kit. Two sequencing platforms, Illumina HiSeq Xten (Illumina, USA) and MGISEQ2000 (BGI, China), were used for genome sequencing. Illumina sequencing libraries were prepared using the TruSeq Nano DNA HT Sample preparation kit (Illumina, USA) following the manufacturer's recommendations. MGI libraries were constructed as follows. In brief, 1-1.5 μg of genomic DNA was randomly fragmented with a Covaris instrument. Then, fragments with sizes between 200-400 bp were selected using an Agencourt AMPure XP-Medium kit, followed by end repair, 3' adenylated and adapter ligation. After PCR enrichment, the PCR products were recovered using the AxyPrep Mag PCR clean-up Kit. The double-stranded PCR products were heat denatured and circularized using the splint oligo sequence. Single-strand circular DNA (ssCir DNA) was formatted as the final library and qualified according to QC procedures. The qualified libraries were sequenced on the Illumina HiSeq X-Ten or MGISEQ2000 platform at the Genome Center of Grandomics (Wuhan, China) (**Supplementary Table 1**).

1.3. Nanopore sequencing

The Oxford Nanopore Technologies (ONT) system was used to sequence all three oat genomes in this study. The ONT ultralong strategy was selected for the whole genome sequencing of the hexaploid species SFS because of its large, complex genome. For this purpose, approximately 8-10 μg of gDNA was size-selected (>50 kb) with the SageHLS HMW library system (Sage Science, USA) and processed using the Ligation sequencing 1D kit (SQK-LSK109, Oxford Nanopore Technologies, UK) according to the manufacturer's instructions. For the diploid and tetraploid samples, ONT-Regular was used for genome sequencing. A total of 3-4 μg DNA per sample was used as input material for the ONT library preparation. After a sample was qualified, size-selection of long DNA fragments was performed using the PippinHT system (Sage Science, USA). The ends of DNA fragments were repaired and A-ligation reactions were conducted with a NEBNext Ultra II End Repair/dA-tailing Kit (Cat# E7546). The adapter provided in the SQK-LSK109 kit (Oxford Nanopore Technologies, UK) was used for the subsequent ligation reaction, and the DNA library was measured by a Qubit® 4.0 Fluorometer (Invitrogen, USA). DNA libraries with approximately 800 ng and 700 ng inserts were constructed for ONT ultralong and ONT-Regula sequencing, respectively, and were sequenced on the PromethION platform (Oxford Nanopore Technologies, UK) at the Genome Center of Grandomics (Wuhan, China).

Base calling was completed with the ONT basecaller Guppy (v3.2.2) with the following parameters: -c dna_r9.4.1_450bps_fast.cfg. Raw Nanopore reads were filtered, and only reads with a mean_qscore_template ≥7 were retained for downstream analyses. A total of 71, 8, and 7 libraries were sequenced for SFS, *A. longiglumis*, and *A. insularis*, generating 1,282.7 Gb, 268.74 Gb, and 481.39 Gb of raw data, respectively. The post-filtered Nanopore reads produced a total of 1,027.83 Gb, 218.67 Gb, and 374.77 Gb of sequencing data, providing approximately 100-, 60- and 60-fold coverage of the genomes, respectively. A summary of ONT read sizes, including the average read length and read N50 value is summarized in **Supplementary Table 2**.

1.4. Hi-C library preparation and sequencing

The Hi-C libraries were prepared as described previously [1] with some modifications. In brief, oat plants *(A. sativa* ssp. *nuda* cv. Sanfensan and *A. insularis*) were grown in a growth chamber for two weeks. Samples of 2-4 g of tender leaves were harvested, cut into pieces of ca. 2 cm^2, and transferred to a 50 ml tubes containing 15 ml of ice-cold nuclear isolation buffer (NBE) with 2% formaldehyde, followed by vacuum infiltration (400 mbar), and incubation with a supplemented crosslinking agent for 1 h. Crosslinking was quenched by adding 2 M glycine to a final concentration of 0.125 M with incubation for 5 min under vacuum, followed by fixation on ice. Then, the fixed leaf pieces were washed three times with sterile Milli-Q water, ground in liquid nitrogen and used for nucleus isolation. The isolated nuclei were purified, checked for quality and quantity and digested with 100 units of *DpnII*. The next steps were Hi-C-specific, including marking the DNA ends with biotin-14-dATP and performing the blunt-end ligation of crosslinked fragments. After ligation, crosslinking was reversed by overnight incubation with proteinase K at 65 °C. Biotin-14-dATP was further removed from nonligated DNA ends using the exonuclease activity of T4 DNA polymerase. DNA was purified by phenol: chloroform (1:1) extraction, precipitated and washed as described. The purified DNA was physically sheared to a size of 300-600 bp by sonication and was size-fractionated using standard 2% agarose gel electrophoresis to obtain fragments in the range of 300-600 bp. The fragmented ends were blunt-end repaired, A-tailed, and subjected to Illumina PE sequencing adapter addition, followed by purification through biotin-streptavidin-mediated pulldown. PCR Amplification was conducted through 12-15 cycles of PCR to enrich the ligation products. After the quality check, the Hi-C libraries were sequenced using the Illumina HiSeq X-Ten instruments with 2 × 150 bp reads. A total of 1312.83 Gb and 816.93 Gb of Hi-C raw data were generated for SFS and *A. insularis*, respectively (**Supplementary Table 1**).

1.5. PacBio Iso-Seq

The three ONT-sequenced oat species were grown in the greenhouse or the field to different growth stages, and the following seven types of samples were collected for RNA isolation: two-week-old seedlings, flag leaves at the booting (Zodoks 45) and heading (Zodoks 58) stages, and panicles at the booting (Zodoks 45), heading (Zodoks 50 and 58) and grain dough (Zodoks 83) stages. The above seven types of RNA samples were mixed in equal amounts and subjected to quality checks using 0.75% agarose gel electrophoresis, a Qubit fluorometer (Thermo Fisher) and an Agilent 2100 BioAnalyzer. Full-length cDNA Iso-Seq template libraries were prepared by following the protocol provided by Pacific Biosciences with some modifications. For each sample, 500 ng of total RNA was employed for reverse transcription using a SMARTer PCR cDNA Synthesis Kit (Clontech). Then large-scale PCR was performed to amplify the cDNAs using KAPA HiFi PCR Kits. To minimize artefacts during large-scale amplification, the number of cycles was optimized and determined to be 14. After large-scale PCR, the resulting PCR products were purified using 1× AMPure PB Beads, followed by additional purification with 0.4× AMPure PB Beads. The purified amplicons were fractionated, and fractions with sizes between 0.5-6 k were harvested using the BluePippin™ Size Selection System to generate SMRTbell™ libraries using the PacBio Template Prep Kit. The SMRTbell templates were then sequenced on a PacBio Sequel II machine at the Genome Centre of Grandomics (Wuhan, China). A total of 46,759,952, 26,389,556 and 13,550,480 reads covering 81.14 Gb, 49.94 Gb, and 25.74 Gb were generated for the hexaploid, tetraploid and diploid species, respectively (**Supplementary Table 1**).

1.6. Contigs assembly, polish and evaluation

To provide guidance regarding genome assembly, the genome sizes of the three *Avena* species were estimated by counting the 17-mer frequency among the clean short reads with Jellyfish (v2.0) [2] software, which resulted in estimated genome sizes of 10.98 Gb, 7.96 Gb and 4.04 Gb for SFS, *A. insularis* and *A. longiglumis*, respectively (**Supplementary Table 3**).

De novo assembly was performed based on Nanopore long reads using the NextDenovo (v2.0-beta.1) pipeline (https://github.com/Nextomics/NextDenovo). Cleaned Nanopore reads were first self-corrected using the NextCorrect module with the default settings, and the corrected reads were then assembled into contigs to obtain the draft assembly using NextDenovo (parameters: reads_cutoff: 1k and seed_cutoff: 54 k for SFS, 25 k for *A. insularis* and *A. longiglumis*). The sizes, contig numbers and contig N50 values of the draft assembled genomes are summarized in **Supplementary Table 3.**

To obtain a high-quality genome assembly, the draft assemblies were further improved by using short reads and corrected Nanopore long reads. For this purpose, raw Illumina or MGI reads were processed with Trimmomatic (v.0.40) [3] to remove adapter sequences, low-quality reads, and short reads (reads with lengths of less than 70 bp). This produced 649.7 Gb, 451.9 Gb, and 204.7 Gb clean reads for SFS, *A.*

insularis and *A. longiglumis*, respectively, achieving ~50-fold coverage of their genomes. Two steps were included to improve the draft genome assemblies: first, using mininmap2 (v2.18) [4] (parameters: -x map-ont) and Racon (v1.4.21) [5] (default settings), the corrected Nanopore reads were aligned to the draft assembly for correction; second, the filtered short reads were employed to polish the draft assemblies using NextPolish. After three rounds of Racon polishing and four rounds of NextPolish polishing, the corrected genomes of SFS, *A. insularis* and *A. longiglumis* had sizes of 10,759,349,041 bp, 7,520,994,703 bp and 3,738,867,912 bp, respectively, which accounted for 97.98%, 94.49% and 92.54% of the genome sizes estimated from K-mer analysis (**Supplementary Table 3**).

Supplementary Table 3 | Statistics of each of the three de novo assembled genomes

	Stat type	Preliminary assembly		Polished genome	
		Contig length (bp)	Contig number	Contig length (bp)	Contig number
A. longiglumis	N50	6,994,235	160	7,297,603	160
	N60	5,694,560	218	5,940,850	218
	N70	4,402,736	290	4,594,691	290
	N80	3,215,601	386	3,358,481	386
	N90	2,004,128	524	2,088,414	524
	Longest	27,786,782	1	29,014,927	1
	Total	3,586,284,815	960	3,738,867,912	960
	Length≥1 kb	3,586,284,815	960	3,738,867,912	960
	Length≥2 kb	3,586,284,815	960	3,738,867,912	960
	Length≥5 kb	3,586,284,815	960	3,738,867,912	960
	Estimated genome size			4,040,471,759	
A. insularis	N50	7,506,894	297	7,836,599	297
	N60	5,836,206	406	6,085,207	406
	N70	4,506,187	548	4,689,328	548
	N80	3,306,342	734	3,435,674	734
	N90	2,039,556	1,004	2,124,822	1,004
	Longest	35,041,226	1	36,557,065	1
	Total	7,213,697,221	1,932	7,520,994,703	1,932
	Length≥1 kb	7,213,697,221	1,932	7,520,994,703	1,932
	Length≥2 kb	7,213,697,221	1,932	7,520,994,703	1,932
	Length≥5 kb	7,213,697,221	1,932	7,520,994,703	1,932
	Estimated genome size			7,959,398,247	
SFS	N50	91,712,002	34	93,262,735	34
	N60	74,100,035	47	75,353,051	47
	N70	59,130,319	63	60,156,522	63
	N80	43,002,173	84	43,730,326	84
	N90	20,584,744	119	20,933,943	119
	Longest	398,393,187	1	405,550,188	1
	Total	10,575,387,261	329	10,759,349,041	329
	Length≥1 kb	10,575,387,261	329	10,759,349,041	329
	Length≥2 kb	10,575,387,261	329	10,759,349,041	329
	Length≥5 kb	10,575,387,261	329	10,759,349,041	329
	Estimated genome size			10,981,026,862	

1.7. Chromosome construction and validation

The genome assembly of the diploid species *A. longiglumis* was anchored and arranged into seven pseudomolecules with RaGOO [6] using the previously published reference genome of the *Avena* A genome diploid *A. atlantica* [7] as the reference. For the tetraploid and hexaploid assemblies, contig anchoring and orientation were performed with the aid of Hi-C data (**Extended Data Fig. 1**). For this purpose, the raw reads from the Hi-C libraries were filtered using fastp [8] with the default settings, resulting in a total of 803,368,743,610 bp and 1,296,125,167,024 bp of clean data. Then the clean Hi-C reads were aligned to the assemblies by using Bowtie2 (v.2.3.2) [9] with the end-to-end model (parameters: -very-sensitive -L 30), which resulted in 45.43% and 48.37% uniquely mapped paired-end reads out of the total ~2,691 million and ~4,221 million read pairs of clean reads for *A. insularis* and SFS, respectively. After considering the map position and orientation of these unique reads, ~870 and ~1,372million read pairs were retained as valid interaction pairs for *A. insularis* and SFS, which represented 71.16% and 67.24% of the uniquely mapped reads and 32.33% and 32.52% of the clean reads, respectively. Second, LACHESIS [10] software was used to cluster, order and orient the contigs into chromosome-length pseudomolecules on the basis of the validated Hi-C dataset with the following parameters: CLUSTER MIN RE SITES=100; CLUSTER MAX LINK DENSITY=2.5; CLUSTER NONINFORMATIVE RATIO=1.4; ORDER MIN N RES IN TRUNK=60; ORDER MIN N RES IN SHREDS=60. After LACHESIS scaffolding, the final SFS assemblies contained 21 pseudomolecules with a total length of 10,438,597,837 bp, accounting for 97.02% of total assembly length, leaving 320,751,204 bp unanchored, whereas the *A. insularis* assemblies contained 14 pseudomolecules with a total length of 7,154,017,286 bp, accounting for 95.12% of the total assembly length. To evaluate the consistency of the Hi-C maps and the consensus genetic maps generated by Bekele *et al.* [11], we aligned the marker sequences from the consensus genetic maps against chromosomes in our SFS assemblies using BLASTN and then summarized the number of best hits (**Extended Data Fig. 2a**). The completeness of the assembly was evaluated using BUSCO (v3.1.0) program [12]. The results showed that 1344 (97.75%), 1349 (98.11%) and 1341 (97.53%) BUSCO genes were identified in the SFS, *A. insularis* and *A. longiglumis* assemblies, respectively (**Extended Data Fig. 2b**).

2. Genome annotation

2.1. Protein-coding gene annotation

Protein-coding genes were predicted using an evidence-based annotation workflow by integrating different sources of evidence.

Transcriptome-based evidence was generated with the following methods. First, full-length transcripts from Iso-Seq were used to produce high-quality opening reading frame (ORF) predictions. For this purpose, raw Iso-Seq sequencing data were first processed with the IsoSeq3 pipeline in SMRT Link (v8.0). Briefly, the "ccs" command (--min-passes 1 --min-rq 0.8) was used to generate consensus sequences (CCSs), which resulted in 1,163,006 (2,476,793,041 bp), 726,902 (1,613,862,251 bp), and 374,567 CCSs (825,983,822 bp) for SFS, *A. insularis* and *A. longiglumis*, respectively. Then, LIMA and REFINE were used to identify the full-length, nonchimeric CCSs with the subsequent step of primer and poly-A tail sequence removal. These sequences were then clustered using an iterative clustering and error correction (ICE) algorithm to obtain unpolished consensus isoforms, which were subsequently polished by using the non-full-length reads and raw bam files with quiver parameters, resulting in a total of 1,150,752, 708,107 and 371,269 high-quality CCSs for SFS, *A. insularis* and *A. longiglumis*. The resulting high-quality CCSs were mapped to the reference genome using minimap2 [4] software with the default settings; then, "fusion_finder.py" and "collapse_isoforms_by_sam.py" implemented in cDNA_Cupcake (v24.3.0) software (https://github.com/Magdoll/cDNA_Cupcake) were sequentially used to filter out fusion genes and redundant sequences, which resulted in the retention of a total of 53,812, 36,397, and 17,961 nonredundant isoforms for SFS, *A. insularis* and *A. longiglumis*, respectively. Finally, the nonredundant full-length transcripts were mapped to the reference genome assemblies using GMAP [13] with the default settings and the resulting BAM files were used as the input for GeneMarkS-T [14] to determine the locations of potential intron-exon boundaries.

Second, a set of homologous proteins from other closely related species was employed as homology evidence using GeMoMa (v1.6.1) [15]. These species include *Avena atlantica*, *Avena eriantha*, *Brachypodium distachyon*, *Hordeum vulgare*, *Oryza* *sativa* and *Triticum aestivum*.

De novo gene predictions were generated using AUGUSTUS (v2.4) [16]. For this purpose, an oat-specific AUGUSTUS gene model was trained using GeneMark-ET (v4.0) [17] with the following parameters: -max_intron max_intron -soft_mask soft_length -pbs -sequence=genome -ET=introns.gff. GeneMark-ET uses Iso-Seq evidence as training data and performs two rounds of iterative gene predictions to train model parameters. The 2000 gene models with the highest scores were used as training data for AUGUSTUS. The resulting gene models were employed to predict the coding genes using AUGUSTUS (-gff3=on-hintsfile=hints.gff-extrinsicCfgFile=extrinsic.cfg-allow_hinted_splicesites= gcag, atac-min_intron_len=30-softmasking=1).

Finally, all gene predictions were integrated into a final gene annotation set using EVidenceModeler (v1.1.1) [18] (parameters: -segmentSize 1000000 -overlapSize 100000) after removing transposable element-related genes, pseudogenes and noncoding genes by using TransposonPSI [19] with the default settings. The results of the annotation process are summarized in **Supplementary Table 4**.

Supplementary Table 4 | Gene models predicted from different types of evidence

Genome	Gene set	$^{\#}$Genes	Average gene length(bp)	Average CDS length(bp)	Average exons number per gene	Average exon length (bp)	Average intron length (bp)
A. longiglumis	De novo	44656	3261.7	1126.51	4.35	258.77	636.74
	Homology	114562	2623.65	827.48	2.6	318.23	1122.4
	RNA-seq	21150	3644.39	1373.62	5.35	256.54	521.48
	Final set	43477	3514.61	1163.98	4.34	268.32	704.19
A. insularis	De novo	106462	3164.93	1188.89	4.29	276.95	600.12
	Homology	170476	6893.24	882.38	3.01	292.86	2985.99
	RNA-seq	33669	4140.16	1619.69	7.2	224.85	406.3
	Final set	89995	3218.01	1195.01	4.48	266.5	580.64
SFS	De novo	130178	2787.59	1106.14	4.06	272.57	549.83
	Homology	92429	3526.35	1239.09	4.36	284.1	680.44
	RNA-seq	35769	3680.68	1499.6	5.97	251.25	438.98
	Final set	120769	2940.27	1136.65	4.28	265.53	549.78

2.2. Functional annotation of gene models

Functional assignments for the predicted protein-coding genes was performed 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 InterProScan (https://github.com/ebi-pf-team/interproscan) with the default settings. A total of 103,773, 81,027 and 40,216 genes were functionally annotated for SFS, *A. insularis* and *A. longiglumis*, respectively, comprising 88.41%, 90.04% and 92.50% of the predicted gene models of each genome assembly (**Supplementary Table 5**).

Supplementary Table 5 | Annotated genes in each of the assembled genomes

Sources	Genome assemblies					
	SFS		*A. insularis*		*A. longiglumis*	
	Number	Percent (%)	Number	Percent (%)	Number	Percent (%)
SwissProt	78,653	65.13	52,614	58.46	28,811	66.27
KEGG	34,790	28.81	23,422	26.03	12,222	28.11
KOG	52,307	43.31	35,987	39.99	19,612	45.11
GO	63,458	52.54	42,028	46.7	23,268	53.52
NR	106,050	87.81	80,568	89.52	39,980	91.96
Annotated genes	106,773	88.41	81,027	90.04	40,216	92.5
Total gene models	120,769		89,995		43,477	

2.3. Noncoding RNA prediction

Noncoding RNAs (ncRNAs), including microRNAs, small nuclear RNAs, rRNAs and regulatory elements, were identified using the Infernal (v1.1.2) [20] program to search against the Rfam database [21]. RNAmmer (v1.2) [22] (parameters: -S euk -m lsu,ssu,tsu –gff) was additionally used to predict rRNAs in more detailed subclasses. Transfer RNAs (tRNAs) were predicted using tRNAscan-SE (v2.0) [23] with eukaryotic parameters. miRNAs were predicted using miRanda (v3.0) (http://www.microrna.org). A total of 59,916, 40,282 and 15,706 ncRNAs were identified in SFS, *A. insularis* and *A. longiglumis*, respectively. The details of these ncRNAs are given in **supplementary Table 6.**

2.4. Repetitive element annotation

Tandem repeats (TRs) in the genome assemblies were identified using GMATA v2.2 [24] and Tandem Repeats Finder (v4.07b) [25] with the following parameters: 2 7 7 80 10 50 500 -f -d -h -r.

Species-specific de novo repeat libraries were constructed with the following steps. First, MITE-Hunter [26] software (parameters: -n 20 -P 0.2 -c 3) was used to identify miniature inverted TEs (MITEs). Then, LTR_FINDER (v1.05) [27] and LTR_harvest (v1.5.10) [28] were used for long terminal repeat (LTR) identification, and the results were processed with LTR_retriever (v2.8) [29] to generate an LTR library. Third, the TR soft-masked reference genome assemblies were hard-masked with both MITE and LTR libraries by using RepeatMasker (v1.331) [30] with the following parameters: nolow -no_is -gff -norna -engine abblast -lib lib, and other de novo repetitive elements were identified with RepeatModeler (v1.0.11) (parameters: -engine wublast) (https://github.com/Dfam-consortium/RepeatModeler) and classified using TEclass [31] (default parameters). Finally, the libraries obtained from MITE, LTR and RepeatModeler were merged to generate the species-specific de novo repeat library, which was used along with the repetitive elements in Repbase (v19.06) [32] to annotate the genomes with RepeatMasker. The results of repetitive element annotation are summarized in **Supplementary Table 7.** The distribution of TEs along each chromosome is visualized in **Fig. 1.**

2.5. Pseudogene annotation

The pseudogenes in each species were identified using Pseudopipe [33]. Each of these pseudogenes was then aligned to the parent gene using MACSE (v2) [34] and only genes with a frameshift or nonsense mutation were considered as the candidate pseudogenes. The total number of pseudogenes in each assembled genome is given in **Table 1**, and their distributions on the chromosomes are visualized in **Extended Data Fig. 9d.**

3. Subgenome assignment, validation and nomenclature

A reference-guided strategy based on subgenome homeology was used to distinguish the subgenomes of *A. insularis* and SFS. For the subgenome assignments of SFS, we first divided the sequenced *A. longiglumis* genome into 100 bp chunks (referred to as markers), which were subsequently aligned to the SFS reference genome using BWA [35] with default settings. Uniquely mapped markers were retained (**Extended Data Fig. 3a**). A syntenic block was generated when more than five markers were consecutively distributed in a syntenic manner (distance between every two adjacent markers of less than 200 kb). This successfully split the 21 chromosomes of SFS into three homoeologous groups **(Extended Data Fig. 3c**). The group showing the highest synteny to *A. longiglumis* was assigned as the A subgenome, the group with moderate synteny to *A. longiglumis* was assigned as the D subgenome, and the remaining group was assigned as the C subgenome according to previous studies which have reporting high homology between the A and D subgenomes but a relatively low homology between the A and C subgenomes [36,37]. Similarly, the genome sequences of *A. insularis* were divided into 100 bp markers and then aligned to the SFS reference genome. The 14 chromosomes were split into two groups, which showed high synteny with the C or D chromosomes of SFS and were hence assigned as the C and D subgenomes, respectively (**Extended Data Fig. 3b, d**).

To validate the correction of the subgenome assignments, two independent approaches were used. First, trimmed short reads from *A. longiglumis* and *A. insularis* were individually mapped to the SFS reference genome using the default settings of BWA [35]. The median depth coverage of the sliding windows (window size: 1 Mb, step size: 0.5 Mb) for *A. longiglumis* or *A. insularis* was calculated using the Mosdepth (v0.3.0) [38] program. The results showed that a much higher mapping depth was achieved for the hexaploid A subgenome chromosomes than for the chromosomes of the other two subgenomes after mapping the *A. longiglumis* reads to the SFS genome, while the chromosomes assigned to the C and D subgenomes showed higher mapping depths than the A subgenome chromosomes after mapping the *A. insularis* reads to the reference genome (**Extended Data Fig. 3e, f**). All of these analyses resulted in consistent subgenome assignments for *A. insularis* and SFS. Second, the abundances and distributions of two types of satellite repeats, As120a and Am1, in all three assembled genomes were investigated by BLASTN analyses. As120a and Am1 are DNA repeats that selectively hybridize to the hexaploid A and C subgenome chromosomes, respectively. The results showed that these two types of repeats were overrepresented on seven pseudochromosomes assigned to the A and C subgenomes in *A. insularis* and SFS, whereas the abundance of these repeats on the D subgenome chromosomes was much lower, providing additional strong evidence of the correct subgenome assignments (**Fig. 1**).

The nomenclature system for wheat chromosomes was adopted for naming the homologous groups (1-7) of SFS. For this purpose, whole-genome protein sequences and gene positions from bread wheat (IWGSC RefSeq v2.1) were retrieved from the GrainGenes database (https://urgi.versailles.inrae.fr/download/iwgsc/IWGSC_RefSeq_Assemblies/v1.1/). If a gene had more than one transcript, only the longest transcript was retained as the representative sequence. The synteny between the bread wheat and SFS was analysed using the MCScanX program with the default settings. The numbers of conserved genes on every pair of chromosomes between SFS and bread wheat are given in **Extended Data Fig. 4a**. The degree of synteny between the wheat genome and the reference SFS genome is displayed in **Extended Data Fig. 4b**.

4. Phylogenomics and comparative genomics analyses of cereal crops

4.1. Phylogenetic tree construction and divergence time estimation

Protein sequences of 43 plant species were downloaded from NCBI, JGI and the official website (**Supplementary Table 8**). Only the longest transcript was selected for each gene locus with alternative splicing variants. Additionally, genes encoding proteins with fewer than 50 amino acids were removed.

Each proteome was subjected to BLAST searches against *Amborella trichopoda* sequences according to an E-value ≤ $1e^{-5}$. Reciprocal best hits (RBHs) in each pair were obtained and the gene families conserved in all the 43 species (52 subgenomes) were retained. The protein sequences from each family were aligned using MUSCLE (v3.8.31) [39] with the default parameters, and the corresponding CDS alignments were back translated from the corresponding protein alignments. The conserved CDS alignments were extracted by Gblocks (v0.9b) [40], and the retained CDS alignments of each family were used for further phylogenomic analyses.

For phylogenetic tree construction, the CDS alignments of each single-copy family were concatenated to generate a supermatrix of 652,068 unambiguously aligned nucleotide positions. Then, 99,3154 DTV sites were extracted from these supergenes and subject to RAxML (v8.2.7) analysis [41] to generate a maximum likelihood tree with the GTR+I+Γ model.

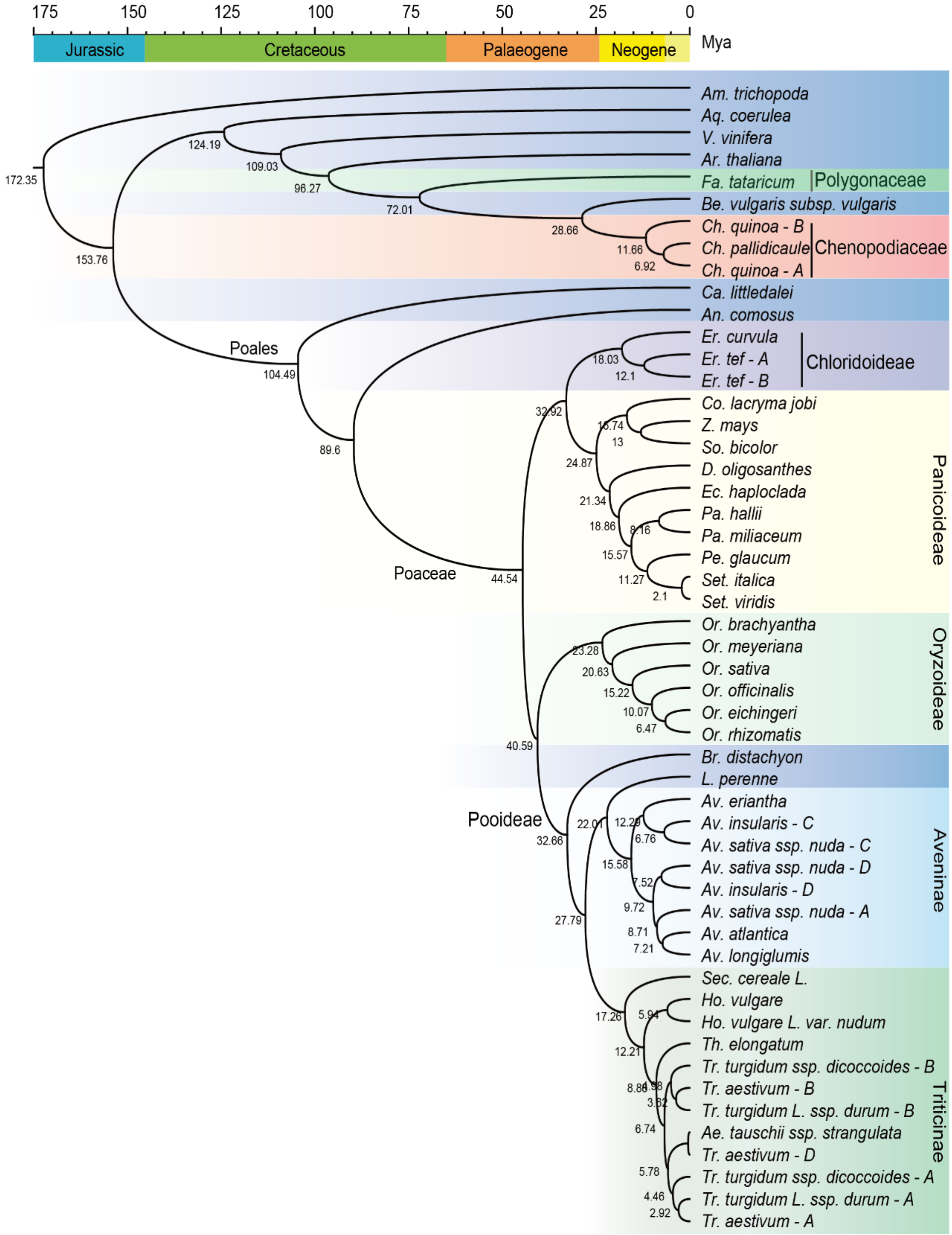
422 **Supplementary Table 8 | List of 43 species with high-quality reference genomes**

Species	Abbreviation*	Accession	Level	ID	Database
Aegilops tauschii ssp. *strangulata*	Atau	AL8/78	Chromosome	ATGSP	official
Amborella trichopoda	Atri	-	scaffold	-	NCBI
Ananas comosus	Acom	F153	Chromosome	GCF_001540865.1	NCBI
Aquilegia coerulea	Acoe	-	scaffold	-	JGI
Arabidopsis thaliana	Atha	-	Chromosome	TAIR10	official
Avena atlantica	Aatl	CC7277	Chromosome	CoGe_53337_v1.0	official
Avena eriantha	Aeri	CN19238	Chromosome	CoGe_53381_v1.0	official
Avena insularis	Ains				This
Avena longiglumis	Alon				This
Avena sativa ssp. *nuda*	Asat				This
Beta vulgaris ssp. *vulgaris*	Bvul	-	Chromosome	-	NCBI
Brachypodium distachyon	Bdis	Bd21	Chromosome	GCF_000005505.3	NCBI
Carex littledalei	Clit	-	Chromosome	GCA_011114355.1	NCBI
Chenopodium pallidicaule	Cpal	-	scaffold	-	official
Chenopodium quinoa	Cqui	-	Chromosome	Cq_PI614886_genome_V1	official
Coix lacryma jobi	Clac	-	Chromosome	Adlay_V1	official
Dichanthelium oligosanthes	Doli	Kellogg 1175	Scaffold	GCA_001633215.2	NCBI
Echinochloa haploclada	Ehap	-	Chromosome	-	official
Eragrostis curvula	Ecur	Victoria	Chromosome	GCA_007726485.1	NCBI
Eragrostis tef	Etef	-	Chromosome	-	official
Fagopyrum tataricum	Ftat	-	Chromosome	-	official
Hordeum vulgare	Hvul	Morex	Chromosome	V2	official
Hordeum vulgare var. *nudum*	Hnud	Lasa Goumang	Scaffold	-	official
Lolium perenne	Lper	-	scaffold	-	official
Oryza brachyantha	Obra	-	Chromosome	GCF_000231095.1	NCBI
Oryza eichingeri	Oeic	-	scaffold	-	official
Oryza meyeriana var. *granulata*	Omey	Menghai	Chromosome	GCA_005223365.2	NCBI

Oryza officinalis	Ooff	-	Chromosome	-	official
Oryza rhizomatis	Orhi	-	scaffold	-	official
Oryza sativa Indica Shuhui498	Osat	-	Chromosome	-	official
Panicum hallii	Phal	FIL2	Chromosome	GCF_002211085.1	NCBI
Panicum miliaceum	Pmil	-	Chromosome	-	official
Pennisetum glaucum	Pgla	-	Chromosome	-	official
Secale cereale	Scer	Lo7	scaffold	Secale_cereale_Lo7_v2	official
Setaria italica	Sita	Yugu1	Chromosome	GCF_000263155.2	NCBI
Setaria viridis	Svir	A10	Chromosome	GCF_005286985.1	NCBI
Sorghum bicolor	Sbic	BTx623	Chromosome	GCF_000003195.3	NCBI
Thinopyrum elongatum	Telo	-	Chromosome	-	official
Triticum aestivum	Taes	Chinese_Spring	Chromosome	IWGSC_WGA_v1.0	official
Triticum turgidum ssp. *durum*	Tdur	Svevo	Chromosome	v1	official
Triticum turgidum ssp. *dicoccoides*	Tdic	Zavitan	Chromosome	151210_zavitan_v2	official
Vitis vinifera	Vvin	-	Chromosome	IGGP_12x	official
Zea mays	Zmay	B73	Chromosome	GCF_000005005.2	NCBI

* The first character of the genus name and the first three characters of the species name or the subspecies/variety name were concatenated to represent the species.

Considering that evolutionary rates are varied at different codon positions, the three codon positions of a concatenated supergene were treated as three different partitions. Divergence times were estimated under a relaxed clock model using the MCMCTree program in the PAML4.7 package [42]. The “Independent rates model (clock=2)” and “JC69” model in MCMCTree program were used. The MCMC process was run for 6,000,000 iterations after a burn-in of 2,000,000 iterations. We ran the program twice for each data type to confirm that the results were similar between runs. The chronogram was produced using FigTree (v1.4.0) (http://tree.bio.ed.ac.uk/) with the first run (**Fig. 2a**, **Supplementary Fig. 1**).



Supplementary Figure 1 |Phylogeny and time scale of 43 plant species, including 33 assembled cereal crops. The number on each branch represents the divergence time.

4.2. Gene family analysis

The pairwise sequence similarities between all input protein sequences were calculated using BLASTP [43] according to an E-value cut-off of 1e-05 followed by the removal of low-quality hits (identity <30% and coverage <30%). Orthologous groups were constructed by OrthoFinder2 (v2.2.7) [44] using the default settings based on the filtered BLASTP results. The results showed that 2,202 clusters contained sequences from all 43 species (52 subgenomes). An overview of the cluster structure is shown in Fig. 2b. Expanded and contracted gene families for each subgenome were identified by comparing the cluster size differences between the ancestor and each species by using CAFÉ (v5) [45]. A random birth-and-death model was employed to evaluate changes the changes in gene families along each lineage of the phylogenetic tree. A probabilistic graphical model (PGM) was used to calculate the probability of transitions in each gene family from parent to child nodes in the phylogeny. Using conditional likelihoods as the test statistics, we calculated the corresponding P-values of each lineage, and a P-value＜0.05 was used as the cutoff to determining the significance of family size change (**Supplementary Table 9**).

The genes that were exclusively found in *Avena* species (>60%) were defined as *Avena* specific. Significantly overrepresented GO terms in each group were identified using the topGO package in the R programming language (https://www.r-project.org/). The significantly overrepresented GO terms were identified with an adjusted P-value of 0.05 or below. (**Supplementary Table 10**).

Supplementary Table 9 | The number of expanded and contracted gene families for each subgenome identified by CAFÉ.

Species*	Expanded	Contracted	Species	Expanded	Contracted
Atri	862	2,544	Osat	767	834
Acoe	2,023	1,861	Ooff	505	919
Vvin	1,693	1,834	Oeic	424	378
Atha	2,606	1,226	Orhi	699	375
Ftat	3,319	911	Bdis	618	1,090
Bvul	676	596	Lper	643	2,920
CquiB	799	1,367	Aeri	1,124	473
Cpal	321	637	AinsC	1,065	632
CquiA	805	854	AsatC	444	2,405
Clit	1,667	2,938	AsatD	648	1,029
Acom	1,536	1,717	AinsD	860	1,178
Ecur	4,048	1,006	AsatA	1,096	1,599
EtefA	449	666	Aatl	823	596
EtefB	405	729	Alon	721	590
Clacr	1,847	1,293	Scer	561	4,632
Zmay	3,381	667	Hvul	550	437
Sbic	451	960	Hnud	925	1,642
Doli	554	2,858	Telo	1,825	346
Ehap	1,326	1,403	TdicB	334	4,108
Phal	92	1,199	TaesB	748	479
Pmil	8,639	166	TdurB	371	1,187
Pgla	992	1,395	Atau	756	809
Sita	283	362	TaesD	565	662
Svir	277	255	TdicA	298	4,173
Obra	350	1,423	TdurA	326	1,270
Omey	971	1,114	TaesA	746	449

* The uppercase letter after the abbreviation for a polyploid species indicates the subgenome.

463 **Supplementary Table 10 | GO term enrichment of *Avena* specific gene families**

GO	Class	#total annotated	#group specific	P value	Term
GO: 0004842	MF	628	33	1.50E-18	ubiquitin-protein transferase activity
GO: 0008270	MF	3,957	72	5.80E-13	zinc ion binding
GO: 0004657	MF	9	6	1.10E-11	proline dehydrogenase activity
GO: 0004869	MF	116	9	1.90E-07	cysteine-type endopeptidase inhibitor activity
GO: 0042393	MF	117	9	2.10E-07	histone binding
GO: 0004222	MF	202	11	3.20E-07	metalloendopeptidase activity
GO: 0003984	MF	25	5	9.10E-07	acetolactate synthase activity
GO: 0008970	MF	34	5	4.50E-06	phospholipase A1 activity
GO: 0050664	MF	38	5	8.00E-06	oxidoreductase activity, acting on NAD(P)H, oxygen as acceptor
GO: 0030410	MF	29	4	5.60E-05	nicotianamine synthase activity
GO: 0005515	MF	17,916	169	6.30E-05	protein binding
GO: 0003700	MF	2,229	30	0.0014	DNA-binding transcription factor activity
GO: 0004713	MF	75	4	0.0022	protein tyrosine kinase activity
GO: 0004601	MF	702	12	0.0057	peroxidase activity
GO: 0016747	MF	1,298	18	0.0072	transferase activity, transferring acyl groups other than amino-acyl groups
GO: 0008233	MF	2,660	20	0.0135	peptidase activity
GO: 0017025	MF	42	2	0.0372	TBP-class protein binding
GO: 0016567	BP	614	33	9.50E-20	protein ubiquitination
GO: 0006562	BP	9	6	8.40E-12	proline catabolic process
GO: 0006511	BP	536	17	2.40E-07	ubiquitin-dependent protein catabolic process
GO: 0007275	BP	423	15	3.10E-07	multicellular organism development
GO: 0009082	BP	52	5	3.00E-05	branched-chain amino acid biosynthetic process
GO: 0030418	BP	29	4	4.60E-05	nicotianamine biosynthetic process
GO: 0006886	BP	548	12	0.00048	intracellular protein transport
GO: 0006633	BP	479	10	0.00197	fatty acid biosynthetic process
GO: 0006367	BP	52	3	0.0056	transcription initiation from RNA polymerase II promoter
GO: 0016192	BP	601	10	0.00943	vesicle-mediated transport
GO: 0005992	BP	74	3	0.0147	trehalose biosynthetic process

GO: 0008152	BP	29661	223	0.02089	metabolic process
GO: 0000160	BP	236	5	0.02443	phosphorelay signal transduction system
GO: 0006352	BP	172	6	0.04946	DNA-templated transcription, initiation
GO: 0030117	CC	154	10	2.40E-09	membrane coat
GO: 0005672	CC	23	3	0.00017	transcription factor TFIIA complex
GO: 0005741	CC	52	3	0.00189	mitochondrial outer membrane
GO: 0005634	CC	2722	22	0.00274	nucleus
GO: 0005852	CC	61	3	0.00299	eukaryotic translation initiation factor 3 complex
GO: 0005840	CC	906	10	0.00964	ribosome

4.3. Karyotype evolution

The AGK (Ancestral Grass Karyotype) genome, which includes 7 protochromosomes and 7,010 ordered protogenes, was downloaded [46], and the protein sequences of rice, bread wheat, and four *Avena* species (*A. eriantha*, *A. longiglumis*, *A. insularis* and SFS) were aligned with the AGK protogenes. Syntenic blocks that were defined based on the presence of at least five syntenic gene pairs were identified using the MCScanX [47] package with the default settings. These syntenic blocks were then used to deduce the homologous relationships between the AGK marker genes and the protein sequences of *Avena* and the related cereal crop species (**Supplementary Table 11**).

Supplementary Table 11 | Number of protogenes in rice, bread wheat and the three assembled *Avena* genomes

Species	AGK genes	Orthologues	# Syntenic blocks
Aeri	5,463	6,563	234
Ains	5,651	12,577	546
Alon	5,269	6,410	297
Asat	5,669	19,112	732
Osat	5,849	7,386	199
Taes	5,473	17,894	814

5. Origin of tetraploid and hexaploid species

5.1. Whole-genome sequencing-based analyses

Plant material

To clarify the evolutionary history of hexaploid oat, 14 *Avena* accessions, representing all extant diploid and tetraploid genomes were chosen for whole-genome sequencing. These included As, Al, Ad, Ac, Cv and Cp genome diploids, AB and CD genome tetraploids and ACD genome hexaploid species. Detailed information on these species, including their genome constitutions, accession numbers, and geographical origins, is listed in **Supplementary Table 1**.

Whole-genome sequencing

For the sequencing of the selected accessions, DNA was isolated from the young leaf tissue of a single plant using the Qiagen DNeasy Plant Mini Kit and 400-bp paired-end (PE) libraries were prepared. Sequencing was conducted on an Illumina HiSeq X-Ten sequencer at the Genome Centre of Grandomics (Wuhan, China) (**Supplementary Table 1)**. Raw data were processed through the Trimmomatic pipeline as described above. Summary statistics for the whole-genome sequencing accessions are shown in **Supplementary Table 1**.

Identity plots

For each accession that was subjected to whole-genome sequencing, approximately 1X clean short paired-end reads were randomly extracted from the resequencing data. Then, these reads were mapped to the repeat hard-masked SFS reference genome using BWA with the default parameters. Uniquely mapped reads were extracted using SAMtools [48] (samtools view -bS -f 3 -q 10). The best hit for each read was retained when the BLASTN score was 15 greater than that of the suboptimal hit and the query coverage was over 60 bp. The average identity over a sliding window of 20 Mb was calculated and plotted against the chromosomes of the SFS assemblies with a step size of 1 Mb.

Variant calling

For all sequenced accessions, we used the BWA [35] program to map the paired-end clean reads to the reference SFS genome. The resulting BAM files were sorted by

SAMtools, PCR duplicates were removed using Picard and deduped BAM files were merged using SAMtools. The mapping rate of each sample was calculated (**Supplementary Table 12**). The mpileup and call functions of BCFtools [48] were used for variant calling. The resulting variants were further filtered using BCFtools with the following parameters: -Ob -g 7 -G 10 -e 'QUAL < 20 || DP < 5'. The numbers of variants identified in each subgenome are listed in **Supplementary Table 12**.

Supplementary Table 12 | Mapping rate and number of SNPs identified based on short paired-end reads using each of the SFS subgenome as the reference sequences.

Sample	# snps in A	# snps in C	# snps in D	Mapping rate (%)
AclaCN21388	5,198,496	75,881,279	7,521,744	97.6642
AvenCN21405	4,829,563	71,615,650	7,008,480	96.4649
AlonCN58138	42,291,586	2,521,110	35,673,134	95.1203
AlonCN58139	36,316,204	1,361,101	26,845,283	98.0327
AstrCN88610	36,926,996	1,511,234	30,064,718	98.8379
AnudCN58062	36,540,711	1,478,917	29,630,618	98.8892
AcanCN23017	38,286,964	2,626,209	38,951,262	96.4071
AdamCN19457	38,947,899	2,197,424	38,280,348	95.1850
AbarCN65538	59,228,472	4,039,965	57,904,127	98.2189
AagaCN25869	62,467,219	3,749,960	60,797,312	99.1863
AsatC_Ogle	10,300,438	15,021,607	12,987,665	99.3413
AdamCN19457	40,502,159	2,364,012	39,902,336	98.8000
AwieCN90217	36,823,766	1,509,793	29,911,128	98.4024
AinsCN108634	11,700,657	36,216,761	27,094,898	97.0268
AinsINS-4	10,538,413	32,828,326	24,788,646	97.9428
AmarCN57945	17,841,454	42,696,686	32,909,786	99.4978
AmurCN21989	18,606,875	46,893,801	33,759,736	99.1080

Phylogenetic tree construction using SNPs

Phylogenetic analysis based on the SNPs identified across the whole genome was carried out using RAxML (v1.0.1, parameters: --all --model GTR+ASC_LEWIS --tree pars{10} --bs-trees 200) with the defaulting settings (**Fig. 3b**). To clarify which species showed the closest relationships to the different hexaploid subgenomes, these SNPs were extracted and compared to each subgenome to construct A-, C- and D-genome phylogenetic trees (**Extended Data Fig. 5**).

5.2. Transcriptome sequencing-based analyses

Plant growth and RNA isolation and sequencing

All diploid accessions that were subjected to whole-genome sequencing were included in the transcriptome analysis. Plants were grown in the greenhouse or the field to different growth stages. Seven sample types from each line, (as described in section 1.5 PacBio Iso-Seq), were collected for RNA extraction. RNA was extracted using a Qiagen RNA isolation kit and RNA quality was accessed by 0.75% agarose gel electrophoresis and on an Agilent 2100 Bioanalyzer. High-quality RNAs from the seven sample types from each accession were mixed in equal amounts. Sequencing libraries were prepared using the MGIEasy RNA Directional Library Prep Kit (BGI, China) according to the manufacturer's protocol and 400-bp paired-end (PE) sequencing was performed using an MGISEQ2000 instrument at the Genome Centre of Grandomics (Wuhan, China) (**Supplementary Table 1**).

Transcript assembly and CDS prediction

MGI raw reads were filtered via the following steps. Read pairs with adapter contamination, read pairs with N contents higher than 3% and read pairs with more than 20% low-quality bases (quality < 20) were first removed. Then, reads with potential low-quality regions were trimmed by applying Trimmomatic (v0.40) [3]. Reads with a quality score below 15 at both ends were also trimmed off, and reads containing 3' or 5' ends with an average quality score dropping below Q20 in a 4 bp sliding window were trimmed. Finally, all reads shorter than 32 bp were excluded to obtain clean data for further analyses. The clean reads were de novo assembled using Trinity (v2.0.3) [49] with the default parameters. The CDSs were predicted using TransDecoder (v5.5.0) (**Supplementary Table 13**).

Supplementary Table 13 | Transcripts de novo assembled by Trinity and the total number of genes identified

Sample	#genes	#transcripts
AlonCN58139	108,830	165,351
AlonCN58138	145,631	214,516
AstrCN88610	101,672	155,779
AstrCN3065	187,658	250,417
AnudCN58062	123,088	188,456
AnudCN79349	164,491	270,651
AnudCN79351	103,147	221,021
AcanCN23017	122,914	180,044
AdamCN19457	114,754	169,443
AclaCN21388	113,988	169,846
AwieCN90217	116,980	175,369

Phylogenetic tree construction and divergence time estimation

Each proteome from a diploid species was subjected to BLAST searches against *Hordeum vulgare* sequences according to an E-value ≤ 1e-5. The RBHs in each pair were obtained, and the gene families that were conserved in all the species were retained for further study. The protein sequences from each conserved gene family were aligned using MUSCLE (v3.8.31) [39] with the default parameters, and the corresponding CDS alignments were back-translated from the corresponding protein alignments. The same methods described in section 4.1 were used for phylogenetic tree construction and divergence time estimation.

5.3. Organelle-based analyses

The chloroplast genomes of *A. longiglumis*, *A. insularis*, SFS, and the other taxa subjected to whole-genome sequencing were assembled using high-quality short paired-ended reads (**Supplementary Table 1**) with NOVOPlasty (v3.7) (https://github.com/ndierckx/NOVOPlasty), in which chloroplasts from *A. murphyi* were employed as the reference (GenBank Accession: NC_044174.1) (**Supplementary Table 14**). We downloaded 26 additional *Avena* chloroplast genomes (**Supplementary Table 15**) to obtain a more comprehensive dataset.

Multiple sequence alignments were performed using MUSCLE, and the informative sites were used for phylogenetic tree construction, in which *Triticum aestivum* was used as the outgroup. All of these analyses were performed with RAxML (v8.2.7) with the following parameter settings: -m GTRGAMMAI -N 100 -f a -k -d -p 12345 -x 12345).

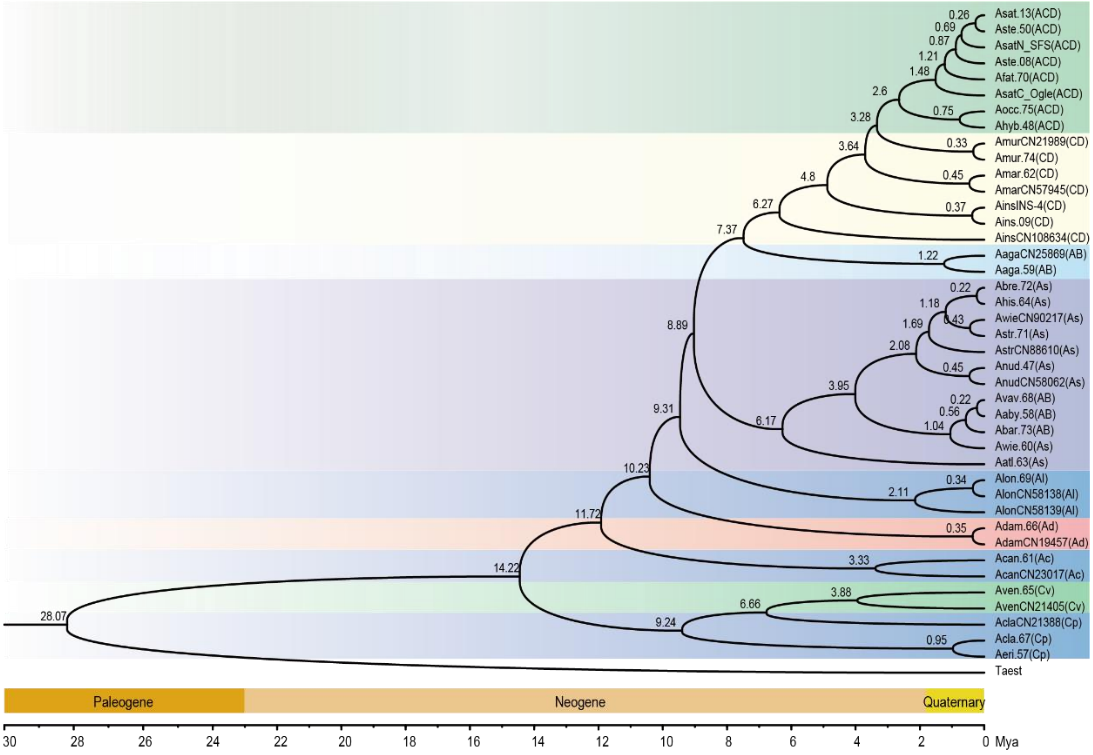
Supplementary Table 14 | Assembled chloroplast genomes and their features

Sample	**Species**	**Haplome**	**Length**	**Content of N**	**Number of Gaps**
AagaCN25869	*A. agadiriana*	AB	135,946	0	0
AbarCN65538	*A. barbata*	AB	135,940	0	0
AcanCN23017	*A. canariensis*	Ac	135,948	0	0
AclaCN21388	*A. clauda*	Cp	135,906	0	0
AdamCN19457	*A. damascena*	Ad	135,926	0	0
AinsCN108634	*A. insularis*	CD	135,944	0	0
AinsINS-4	*A. insularis*	CD	135,967	0	0
AlonCN58138	*A. longiglumis*	Al	135,727	0	0
AlonCN58139	*A. longiglumis*	Al	135,728	0	0
AmarCN57945	*A. maroccana*	CD	135,884	0	0
AmurCN21989	*A. murphyi*	CD	135,890	0	0
AnudCN58062	*A. nuda*	As	135,935	0	0
AsatN_SFS	*A. sativa* ssp. *nuda*	ACD	135,891	0	0
AsatC_Ogle	*A. sativa*	ACD	135,883	0	0
AstrCN88610	*A. strigosa*	As	135,930	0	0
AvenCN21405	*A. ventricosa*	Cv	135,761	0	0
AwieCN90217	*A. wiestii*	As	135,935	0	0

Supplementary Table 15 | Chloroplast genomes of *Avena* species from public databases

Sample	Species	Haplome	Accession
Aaby.58	*Avena_abyssinica*	AB	NC_044158.1
Aaga.59	*Avena_agadiriana*	AB	NC_044159.1
Aatla.63	*Avena_atlantica*	As	NC_044163.1
Abar.73	*Avena_barbata*	AB	NC_044173.1
Abre.72	*Avena_brevis*	As	NC_044172.1
Acan.61	*Avena_canariensis*	Ac	NC_044161.1
Acla.67	*Avena_clauda*	Cp	NC_044167.1
Adam.66	*Avena_damascena*	Ad	NC_044166.1
Aeri.57	*Avena_eriantha*	Cp	NC_044157.1
Afat.70	*Avena_fatua*	ACD	NC_044170.1
Ahis.64	*Avena_hispanica*	As	NC_044164.1
Ahyb.48	*Avena_hybrida*	ACD	NC_044148.1
Ains.09	*Avena_insularis*	CD	MG674209.1
Alon.69	*Avena_longiglumis*	Al	NC_044169.1
Alus.49	*Avena_lusitanica*	As	NC_044149.1
Amar.62	*Avena_maroccana*	CD	NC_044162.1
Amur.74	*Avena_murphyi*	CD	NC_044174.1
Anud.47	*Avena_nuda*	As	NC_044147.1
Aocc.75	*Avena_occidentalis*	ACD	NC_044175.1
Asat.13	*Avena_sativa*	ACD	MG687313.1
Aste.08	*Avena_sterilis*	ACD	MG687308.1
Aste.50	*Avena_sterilis*	ACD	NC_031650.1
Astr.71	*Avena_strigosa*	As	NC_044171.1
Avav.68	*Avena_vaviloviana*	AB	NC_044168.1
Aven.65	*Avena_ventricosa*	Cv	NC_044165.1
Awie.60	*Avena_wiestii*	As	NC_044160.1

Divergence times were estimated under a relaxed clock model using the MCMCTree program in the PAML4.7 package [42]. The “Independent rates model (clock=2)” and “JC69” models in the MCMCTree program were used. The MCMC process was run for 6,000,000 iterations after a burn-in of 2,000,000 iterations. We ran the program twice for each data type to confirm that the results were similar between runs. The chronogram was visualized using FigTree (v1.4.0) with the first run (**Supplementary Fig. 2**).



Supplementary Figure 2 | Phylogenetic relationship among *Avena* species based on chloroplast genome sequences.

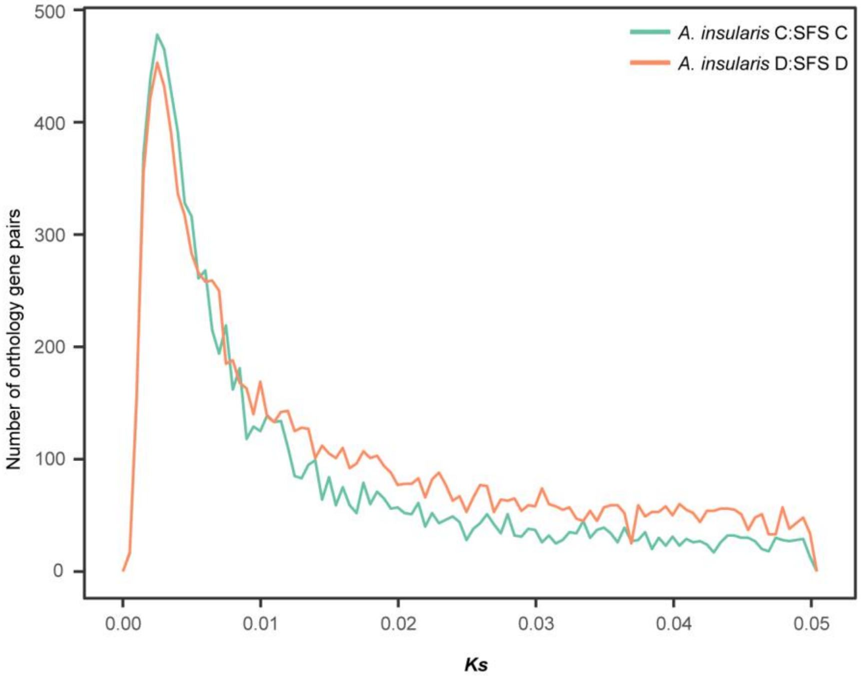
5.4. Timing of allo-hexaploidy formation

To dating the time of hexaploid origin, we obtained the orthologous gene pairs between the two C subgenomes and two D subgenomes of *A. insularis* and SFS, and calculated the synonymous substitution rate (*Ks*) values of the orthologous gene pairs using the yn00 module of the PAML4.7 package. Divergence time was estimated using the method described by Salse *et al.* [50]. The results suggested the hexaploid oat formed around 0.523~0.585 mya (**Supplementary Table 16, Supplementary Fig. 4**)). For pseudogenes, the nucleotide sequences before the frameshift or nonsense mutation sites were removed, and the remaining nucleotide sequences were aligned by MUSCLE. Divergence was calculated by distmat, and the time of pseudogenization was estimated using a mutation rate of 1.3 × 10^{-8} mutations per site per year [51] (**Supplementary Fig. 4**).

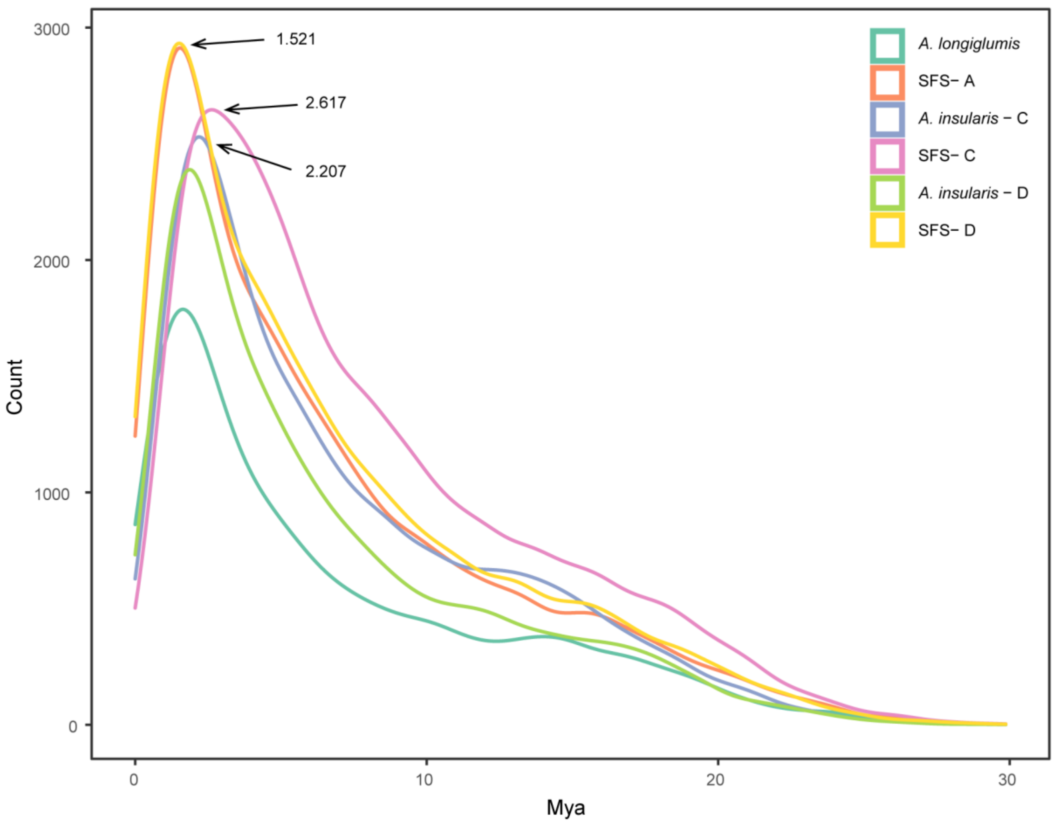
Supplementary Table 16 | Peaks of each *Ks* distribution of orthologues in the subgenomes of *A. insularis* and SFS.

Orthologs	*Ks* peak value	Divergence time (mya)
A. insularis D vs SFS D	0.0034	0.523
A. insularis C vs SFS C	0.0038	0.585

Note: The formula T=Ks/r was used to estimate the divergence time between the subgenomes as described by Salse *et al*. [50], where r is the average substitution rate for grass species which was determined to be 6.5 ×10^{-9} substitutions per synonymous site per year [52].



Supplementary Figure 3 | Dating the divergence of the tetraploid and hexaploid oats. The *Ks* distribution is shown for orthologous gene pairs between two C subgenomes and two D subgenomes of *A. insularis* and SFS. Data are grouped into *Ks* units of 0.001.



Supplementary Figure 4 | Time of pseudogenization in the Al genome (*A. longiglumis*) and the subgenomes of *A. insularis* and SFS.

6. Subgenome evolution

6.1. Chromosome rearrangement

Synteny analysis

Subgenome synteny among the subgenomes of *A. insularis* and SFS was individually analysed by plotting the positions of homoeologous pairs in the subgenome pairs within the context of 14 and 21 chromosomes using Circos [53] (**Extended Data Fig. 6**). The syntenic blocks between the SFS subgenomes and the tetraploid *A. insularis* and the diploid *A. longiglumis* were identified using MCScanX and were visualized using Circos (**Fig. 1**).

To explore broad-scale structural variations after polyploidization, we used SFS to perform in silico painting of the *A. insularis* and *A. longiglumis* genomes with the method described previously [54]. In brief, the SFS genome was divided into 100 bp markers, which were then aligned to concatenated repeat hard-masked genomes of *A. insularis* and *A. longiglumis* using BWA with the default settings. The uniquely mapped markers with alignment lengths over 50 bp in the target genome were retained. We then processed the markers on each chromosome by requiring at least five consecutive markers supporting homology to the same SFS chromosome. We consolidated each group of five consecutive potential markers as one confirmed block. These confirmed blocks with a distance of less than 2 Mb were further consolidated as superblocks (**Fig 4a, bottom**). A similar painting analysis was performed by painting 100 bp marker from *A. insularis* onto concatenated genomes of *A. longiglumis* and *A. eriantha* (**Fig. 4a, top**).

To further explore the genomic exchanges between *A. insularis* and SFS after polyploidization, clean short paired-end reads from the Cp genome diploid *A. eriantha* and the Al genome diploid *A. longiglumis* were individually mapped onto the reference *A. insularis* and SFS genomes using BWA. The signle-base depth coverage of properly paired reads from the *A. longiglumis* and *A. eriantha* mapping results was calculated using the Mosdepth program and plotted along each chromosome of the reference genome (**Fig. 4c, Extended Data Fig. 7a, c, d**). A similar analysis was performed by aligning reads from *A. insularis* to the SFS genome **(Extended Data Fig. 7b**).

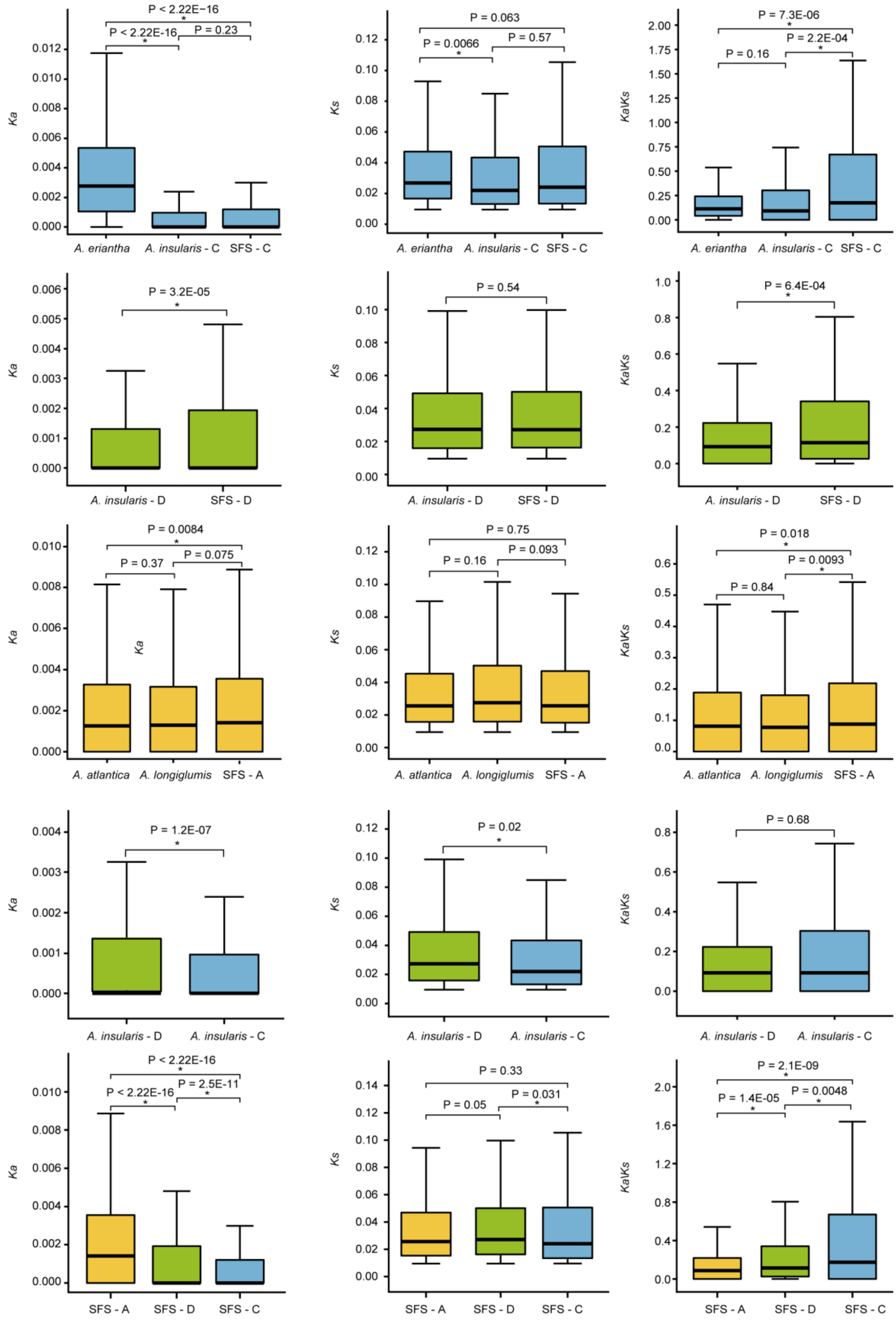
Fluorescence in situ hybridization (FISH)

To validate the observed C/D and C/A intergenomic exchanges in the *A. insularis* and SFS genomes, FISH analysis was performed using a C genome-specific repeat, Am1 as the probe. The FISH probe was prepared as described in Yan *et al.* [55]. The metaphase chromosome preparation method paralleled that employed in previous experiments [56] with some modifications. In brief, seeds of *A. insularis* and SFS were imbibed in distilled water for 18 h at 25°C in the dark and then placed in petri dishes with two layers of moist filter papers. The germinated seeds were transferred to a cabinet at a temperature of 4°C to synchronize cell division and allow to accumulation of metaphase plates. Root tips were harvested when they reached to 1.5-2.0 cm and were pre-treated in 1.0 MPa nitrous oxide gas for 3 h followed by fixation using glacial acetic acids for 20 min. The apical meristem was extruded from the fixed root tip and digested with 2% cellulase and 1% pectinase for 2 h. The digested apical meristem was squashed in a drop of 60% acetic acid, and the resulting suspension was dropped onto a clean glass side.

FISH analysis was performed as described by Fu *et al.* [57]. Briefly, air-dried slides were fixed for 10 min with 4% (w/v) paraformaldehyde and then immersed in 2× saline sodium citrate (SSC) for 10 min. After dehydration in an ice-cold ethanol series (75%, 95%, and 100%) for 5 min in each concentration, the slides were air dried. The air dried slides were then subjected to denaturing at 80°C for 2 min in deionized formamide (60 µl per slide), followed by dehydration in 75%, 95%, and 100% alcohol at -20°C for 5 min each before air drying again. A 10 µl aliquot of a hybridization mixture containing 0.5 µl of the FISH probe, 4.75 µl of 2× SSC, and 4.75 µl of 1× TE was applied to each slide, and the slides were then incubated for 2 h at 37°C. The slides were next counterstained with DAPI and Vectashield mounting medium (Vector Laboratories, Inc., Burlingame, CA, USA). Digital images were captured using an Olympus BX-51 epifluorescence microscope equipped with a Photometric SenSys Olympus DP80 CCD camera (Olympus, Tokyo) and processed using Photoshop V7.0 (Adobe Systems Incorporated, San Jose, CA) (**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 fetched from OrthoFinder2 results [44]. A total of 2,767 orthologous gene sets were obtained and then used for the nonsynonymous (*Ka*) and synonymous (*Ks*) rate calculations (Fig. 4e, Supplementary Fig. 5). For this purpose, the orthologous gene pair list was used as the input, and the protein sequences from each gene pair were aligned using MUSCLE [39]. PAL2NAL [58] 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/K*s values between genomes (subgenomes) was estimated using the Wilcoxon rank-sum test for nonnormal distributions in R.



Supplementary Figure 5 |Comparison of codon substitution rate distributions between the subgenomes of SFS and *A. insularis*, and the A (*A. longiglumis*, *A. atlantica*) and C (*A. eriantha*) genome diploid progenitors. Comparison of *Ka*, *Ks* and *Ka/Ks* distributions between subgenomes and the putative diploid progenitor genomes of *A. longiglumis* (Al genome), *A. atlantica* (As genome) [7] and *A. eriantha* (Cp genome) [7]. All estimates with *Ks*<0.01 were excluded from the analysis. The central line for each box plot indicates the median. The top and bottom edges of the box indicate the first and third quartiles and the whiskers extend 1.5 times the interquartile range beyond the edges of the box. The significance of the differences in the values between genomes (subgenomes) was estimated using the Wilcoxon rank-sum test (*, P < 0.05).

6.2. Subgenome contents

Kmer distribution

The 31-mer frequency in the sliding window (window size: 1 Mbp, step size: 0.5 Mbp) of the Al, CD, and ACD genome assemblies was counted using Jellyfish [2], and the highest frequencies in each window were plotted along the chromosomes (**Fig. 1**).

Full-length LTR analyses

Full-length LTRs (FL-LTRs) were identified using LTR_FINDER (**Extended Data Fig. 8b**). The average sequence length of FL-LTRs was calculated (**Extended Data Fig. 8c**). The retained FL-LTRs were classified into different families based on sequence similarity. For this purpose, these full-length LTRs were first searched against the Copia and Gypsy domains in Pfam using hmmsearch. Then, the un-classified full-length LTRs were subjected to BLAST searches against the TREP database (release 19). Finally, the remaining repeat elements were further classified using the RepeatClassifier module in RepeatModeler [30]. The results showed that the two superfamilies, Gypsy and Copia contributed largely to the LTRs in *Avena* genomes.

To estimate the insertion times for the full-length LTRs, the 5'- and 3'-LTR sequences were aligned and used to calculate K-value (the average number of substitutions per aligned site) using distmat [59]. The insertion times were estimated with the formula T=K/2r, where r represents the neutral mutation rate of 1.3×10^{-8} mutations per site per year [51] (**Extended Data Fig. 8d)**.

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 *A. eriantha* genome as the reference was employed to reveal the percentage of retained genes in the C and D subgenome of *A. insularis* (**Extended Data Fig. 9e**). The gene retention rates of the SFS subgenomes were calculated and plotted using the same methods (**Extended Fig. 9f**).

6.3. Subgenome dominance

Plant materials and transcriptome sequencing

RNAs were isolated from seven sample types of SFS, including seedlings, flag leaves, and panicles at different developmental stages (as described in section 1.5). Each type of RNA sample was sequenced with 3 biological repeats on an MGISEQ2000 instrument. To further understand the responses of genes in different subgenomes of SFS under abiotic stress, seedlings of SFS were exposed to heat, cold, drought, waterlogging, alkalinity and salt. For the abiotic treatments, oat plants were first grown in well-watered conditions in a growth chamber for 14 d at 20°C under 12 h of daily light, and plants were then either left in these growth conditions as controls or transferred to other growth chambers for stress treatments. For cold treatment, the plants were grown in a growth chamber at 4°C, while for heat treatment, the plants were grown in a growth chamber under a light cycle with 12 h of light at 37°C and 10 h of darkness at 32°C. For drought and waterlogging treatments, the plants were carefully transferred to other plots containing 10% PEG6000 or muddy soil. For the alkaline and salt treatments, water was replaced by a 6 mmol/L alkaline solutions (Na2CO3: NaHCO3=1:1) or a 40 mmol/L salt solution (NaCl: Na2SO4=1:1), respectively. One week after all treatments, the seedlings were harvested with 3 repeats from each treatment and used for RNA isolation. The same methods described in section 5.2 were adopted for RNA sequencing libraries construction and sequencing.

Quantification of gene expression levels

Paired-end MGI reads from the RNA samples described above were subjected to quality trimming using Trimmomatic (v0.40) with the default settings and aligned to the gene models with HISAT2 [60] software according to the default parameters. Gene expression levels were quantified using the HTseq (v0.9.1) [61] program with the SFS gene models as the reference. Expression levels were quantified as transcripts per million values.

Identification of differentially expressed genes in stress-treated samples

The differentially expressed genes (DEGs) between different stress-treated sample pairs were identified with the edgeR software package [62]. For each gene, an adjusted P-value (corrected for the false discovery rate (FDR)) was calculated using the one-sided Fisher exact test. Genes with an adjusted P-value below 0.05 and a $\log_2$ FC greater than 0.5 were considered differentially expressed (**Supplementary Table 17**).

Supplementary Table 17 | Distribution of the DEGs identified on each chromosome of SFS under different stresses.

Chromosome	Alkaline	Cold	Drought	Heat	Salt	Waterlogging
1A	21	165	1,480	476	322	10
2A	13	120	885	251	214	7
3A	18	113	856	244	213	7
4A	22	181	1,573	511	370	12
5A	22	156	1,336	351	291	13
6A	21	149	1,305	337	326	15
7A	25	126	1,129	346	224	8
1D	24	180	1,410	461	303	12
2D	15	175	1,478	419	331	8
3D	26	132	1,019	320	229	6
4D	27	171	1,511	473	353	7
5D	33	159	1,409	369	281	9
6D	15	76	706	226	186	9
7D	26	166	1,250	365	293	11
1C	6	66	614	201	126	5
2C	27	145	1,172	366	265	6
3C	17	146	1,090	304	256	6
4C	18	116	979	316	245	4
5C	15	138	1,239	378	292	9
6C	18	121	1,228	353	284	10
7C	9	70	657	188	135	2
Total	446	3,024	25,542	7,603	5,806	190

Note: The colour of each cell is proportional to the number of DEGs in each column.

Analysis of homoeologous gene expression

Differences in the expression patterns of homoeologous genes in SFS were analysed to test whether subgenome dominance, a striking whole-genome feature common to polyploids, was present. For this purpose, we used MCScanX [47] to detect syntenic blocks (regions with at least five collinear genes). Among these blocks, we identified 41,232 homoeologous genes that were present in 13,744 triads with a single gene copy per subgenome (an A:C:D configuration of 1:1:1). Then, the raw expression values (TPM values) of these triplets from seedlings, flat leaves, panicles at different developmental stages and seedlings under six abiotic stresses were transformed by adding 1 and taking the common logarithm, and the expression matrix was subjected to two-dimensional hierarchical clustering using the correlation distance and the average linkage method to form clusters (**Fig. 4f**). The differentially expressed orthologous genes (DEOGs) between different subgenome pairs were defined as gene triplets with a pairwise log2-fold change exceeding 0.5 and adjusted P-value below 0.05 (**Supplementary Table 18**). The expression patterns of these DEGOSs were visualized in the heatmap shown in **Fig. 4g** using the heatmap.2 command from the R package gplots.

Supplementary Table 18 | Dominant gene expression between the subgenomes in SFS

RNA Sample [*]	A vs C		A vs D		C vs D	
	Up in A	Up in C	Up in A	Up in D	Up in C	Up in D
AsatN_SFS_A_L	898	728	488	459	767	900
AsatN_SFS_CK_L	912	722	530	444	784	886
AsatN_SFS_C_L	916	759	472	451	810	901
AsatN_SFS_D_L	931	645	516	467	697	900
AsatN_SFS_H_L	996	838	566	533	871	1001
AsatN_SFS_S_L	1035	853	583	543	915	1044
AsatN_SFS_W_L	651	576	353	334	591	685
AsatN_SFS_L0	605	517	300	283	541	624
AsatN_SFS_L1	1030	864	565	539	887	1057
AsatN_SFS_L3	1035	876	590	534	892	1002
AsatN_SFS_S1	888	656	490	471	716	899
AsatN_SFS_S2	780	570	416	360	627	813
AsatN_SFS_S3	882	615	482	464	660	891
AsatN_SFS_S4	608	465	325	314	522	641
Total	12167	9684	6676	6196	10280	12244

[*] All RNA samples were isolated from different tissues of SFS (abbreviated as AsatN_SFS to distinguish it from another hexaploid taxon, “Ogle”) grown under normal conditions or abiotic stresses. A: alkaline, CK: control, C: cold, D: drought, H: heat, S: salt, W: waterlogging. L: seedling; L0-L3: two-week-old seedlings (L0), flag leaves at the booting (Zodoks 45, L1) and heading (Zodoks 58, L3) stages; S1-S4: panicles at the booting (Zodoks 45, S1), heading (Zodoks 50 and 58, S2 and S3) and grain dough (Zodoks 83, S4) stages.

Relationship between gene expression and TE-density

To test whether the density of nearby TEs was correlated with gene expression levels, as observed in previous studies [63,64], we calculated the TE densities of the 5 kb up- and downstream sequences, both separately and together, for each gene from the 13,744 triads. The results revealed that homoeologs from the C subgenome of SFS had a higher TE density than those from the A and D subgenomes (**Extended Data**

Fig. 10). We then divided the 13,744 triplets into 20 bins according to the TE density (both 5 kb up- and downstream sequences were included). The expression values of the genes in each bin were averaged. The results showed that the expression levels decreased with an increasing TE density, supporting a negative correlation between the expression level and the density of nearby TEs.

Supplementary References

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

2 Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770, doi:10.1093/bioinformatics/btr011 (2011).

3 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

4 Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100. doi:10.1093/bioinformatics/bty191 (2018).

5 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).

6 Alonge, M. et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol* **20**, 224. doi:10.1186/s13059-019-1829-6 (2019).

7 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).

8 Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890, doi:10.1093/bioinformatics/bty560 (2018).

9 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

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

doi:10.1038/nbt.2727 (2013).

11 Bekele, W. A., Wight, C. P., Chao, S., Howarth, C. J. & Tinker, N. A. Haplotype-based genotyping-by-sequencing in oat genome research. *Plant Biotechnol J* **16**, 1452-1463, doi:10.1111/pbi.12888 (2018).

12 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212, doi:10.1093/bioinformatics/btv351 (2015).

13 Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-1875, doi:10.1093/bioinformatics/bti310 (2005).

13 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).

15 Keilwagen, J. *et al.* Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res* **44**, e89-e89, doi:10.1093/nar/gkw092 (2016).

16 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).

17 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).

18 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).

19 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).

20 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).

21 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).

22 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).

23 Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964, doi:10.1093/nar/25.5.955 (1997).

24 Wang, X. & Wang, L. GMATA: An integrated software package for genome-scale SSR mining, marker development and viewing. *Front Plant Sci* **7**, doi:10.3389/fpls.2016.01350 (2016).

25 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580, doi:10.1093/nar/27.2.573 (1999).

26 Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res* **38**, e199-e199, doi:10.1093/nar/gkq862 (2010).

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

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

29 Ou, S. & Jiang, N. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol* **176**, 1410-1422, doi:10.1104/pp.17.01310 (2018).

30 Bedell, J. A., Korf, I. & Gish, W. MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics* **16**, 1040-1041, doi:10.1093/bioinformatics/16.11.1040 (2000).

31 Abrusán, G., Grundmann, N., DeMester, L. & Makalowski, W. TEclass-a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329-1330, doi:10.1093/bioinformatics/btp084 (2009).

32 Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements.

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

33 Zhang, Z. *et al.* PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* **22**, 1437-1439, doi:10.1093/bioinformatics/btl116 (2006).

34 Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N. & Delsuc, F. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol Biol Evol* **35**, 2582-2584, doi:10.1093/molbev/msy159 (2018).

35 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595, doi:10.1093/bioinformatics/btp698 (2010).

36 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).

37 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).

38 Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867-868, doi:10.1093/bioinformatics/btx699 (2018).

39 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).

40 Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**, 564-577, doi:10.1080/10635150701472164 (2007).

41 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).

42 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591, doi:10.1093/molbev/msm088 (2007).

43 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC*

Bioinformatics **10**, 421, doi:10.1186/1471-2105-10-421 (2009).

44 Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238, doi:10.1186/s13059-019-1832-y (2019).

45 De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFÉ: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269-1271, doi:10.1093/bioinformatics/btl097 (2006).

46 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).

47 Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* **40**, e49-e49, doi:10.1093/nar/gkr1293 (2012).

48 Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi: 10.1093/bioinformatics/btp352 (2009).

49 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652, doi:10.1038/nbt.1883 (2011).

50 Salse, J. *et al*. 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 Genomics* **9**, 555, doi:10.1186/1471-2164-9-555 (2008).

51 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).

52 Gaut, B. S., Morton, B. R., McCaig, B. C. & Clegg, M. T. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proc Natl Acad Sci USA* **93**, 10274, doi:10.1073/pnas.93.19.10274 (1996).

53 Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645, doi:10.1101/gr.092759.109 (2009).

54 Schield, D. R. *et al*. The origins and evolution of chromosomes, dosage

compensation, and mechanisms underlying venom regulation in snakes. *Genome Res* **29**, 590-601, doi: 10.1101/gr.240952.118 (2019).

55 Yan, H. *et al.* New evidence confirming the CD genomic constitutions of the tetraploid *Avena* species in the section *Pachycarpa* Baum. *PloS One* **16**, e0240703, doi:10.1371/journal.pone.0240703 (2021).

56 Fominaya, A., Loarce, Y., Montes, A. & Ferrer, E. Chromosomal distribution patterns of the $(AC)_{10}$ microsatellite and other repetitive sequences, and their use in chromosome rearrangement analysis of species of the genus *Avena*. *Genome* **60**, 216-227, doi:10.1139/gen-2016-0146 (2017).

57 Fu, S. et al. Oligonucleotide probes for ND-FISH analysis to identify rye and wheat chromosomes. Sci Rep 5, 10552, doi:10.1038/srep10552 (2015).

58 Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**, W609-612, doi:10.1093/nar/gkl315 (2006).

59 Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends Genet* **16**, 276-277, doi:10.1016/s0168-9525(00)02024-2 (2000).

60 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360, doi:10.1038/nmeth.3317 (2015).

61 Anders, S., Pyl, P. T. & Huber, W. HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169, doi:10.1093/bioinformatics/btu638 (2014).

62 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 (2010).

63 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).

64 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).