

1

2 **Chromosome-scale genome assembly of the diploid oat *Avena***  
3 ***longiglumis* reveals the landscape of repetitive sequences, genes**  
4 **and chromosome evolution in grasses**

5

6 Qing Liu<sup>1,2\*</sup>, Hongyu Yuan<sup>1,3</sup>, Mingzhi Li<sup>4</sup>, Ziwei Wang<sup>1,3</sup>, Dongli Cui<sup>1,3</sup>, Yushi Ye<sup>1</sup>, Zongyi  
7 Sun<sup>5</sup>, Xukai Tan<sup>5</sup>, Trude Schwarzacher<sup>1,6</sup> and John Seymour Heslop-Harrison<sup>1,6\*</sup>

8 <sup>1</sup> Key Laboratory of Plant Resources Conservation and Sustainable Utilization / Guangdong Provincial Key  
9 Laboratory of Applied Botany, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou,  
10 510650, China. <sup>2</sup> Center for Conservation Biology, Core Botanical Gardens, Chinese Academy of Sciences,  
11 Guangzhou, 510650, China. <sup>3</sup> University of Chinese Academy of Sciences, Beijing, 100049, China. <sup>4</sup> Bio&Data  
12 Biotechnologies Co. Ltd., Guangzhou, 510700, China. <sup>5</sup> Grandomics Biosciences, Beijing 102200, China. <sup>6</sup>  
13 Department of Genetics and Genome Biology, University of Leicester, Leicester LE1 7RH, UK.

14 \* Correspondence: [liuqing@scib.ac.cn](mailto:liuqing@scib.ac.cn) (QL); [phh4@le.ac.uk](mailto:phh4@le.ac.uk), [phh@molcyt.com](mailto:phh@molcyt.com) (JSHH)

15

16 Running head: Genome assembly and evolutionary genomics of oats

17

18 **Abstract**

19 **Background:** Oat (*Avena sativa*,  $2n=6x=42$ ) is an important crop, and with its wild relatives  
20 including *A. longiglumis* (ALO,  $2n=6x=14$ ), has advantageous agronomic and nutritional traits.  
21 A *de-novo* chromosome-level ALO genome assembly was made to investigate diversity and  
22 structural genome variation between *Avena* species and other Poaceae in an evolutionary  
23 context, and develop genomic resources to identify the pangenome and economic traits within  
24 Pooideae.

25 **Results:** The 3.85 gigabase ALO genome (seven pseudo-chromosomes), contained 40,845  
26 protein-coding genes and 87% repetitive sequences (84.21% transposable elements). An LTR  
27 retrotransposon family was abundant at all chromosome centromeres, and genes were  
28 distributed without major terminal clusters. Comparisons of synteny with *A. eriantha* and *A.*  
29 *strigosa* showed evolutionary translocations of terminal segments including many genes.  
30 Comparison with rice ( $x=12$ ) and the ancestral grass karyotype showed synteny and features of  
31 chromosome evolution including fusions, translocations and insertions of syntenic blocks  
32 across Pooideae species. With a genome size 10 times larger than rice, ALO showed relatively  
33 uniform expansion along the chromosome arms, with few gene-poor regions along arms, and  
34 no major duplications nor deletions. Linked gene networks were identified (mixed-linkage  
35 glucans and cellulose synthase genes), and *CYP450* genes may be related to salt-tolerance.

36 **Conclusions:** The high-continuity genome assembly shows gene, chromosomal structural  
37 and copy number variation, providing a reference for the *Avena* pangenome, defining the full  
38 spectrum of diversity. Chromosomal rearrangements and genome expansion demonstrate  
39 features of evolution across the genus and grass BOP-clade, contributing to exploitation of  
40 gene and genome diversity through precision breeding.

41 **Key words:** *Avena longiglumis*, Ancestral karyotype, Chromosome rearrangement, Copy  
42 number variation (CNV), Genome assembly, Genome expansion, Retrotransposons, Structural  
43 variation, Translocations, Transposable elements

## 44 **Background**

45 In the Poaceae, cultivated common oat (*Avena sativa* L.;  $2n = 6x = 42$ , AACCCDD) and its wild  
46 relatives ( $2x$ ,  $4x$  and  $6x$ ) belong to the Aveneae tribe, which diverged from the Triticeae (syn.  
47 Hordeae) tribe (including wheat, barley and rye) around 27.8 million years ago (Mya) [1], and  
48 are part of the BOP- or BEP- clade [Bambusoideae, Oryzoideae (syn. Ehrhartoideae), and  
49 Pooideae] including half of all grasses, separating 70 Mya from the other grass lineages [2, 3,  
50 4, 5]. Global production of oat reached 22 million tons in 2021 (<http://www.fao.org/faostat/>).  
51 Human studies have demonstrated the beneficial effects of consuming oats for the reduction of  
52 serum cholesterol and cardiovascular disease, associated with the soluble  $\beta$ -glucan component  
53 [6], and a favourable glycaemic index with a low value and slow carbohydrate breakdown. Oat  
54 has substantial concentrations of phytochemicals (e.g., avenine, avenacin and phenolic  
55 compounds) and tolerance to the harsh environments such as sandy loam soil, short growing  
56 seasons and desert climate, making the crop resilient [7]. The oil content of oat grain (6%) is  
57 high among cereals, suggesting a possible important future use, like maize, for food oils.

58 Due not least to the large genome size of *A. sativa* [1C genome size  $\sim$ 12.5 gigabases (Gbp)  
59 [8], oat genomics has lagged behind those of other crops such as rice (*Oryza sativa*) [9],  
60 sorghum (*Sorghum bicolor*) [10] or foxtail millet (*Setaria italica*) [11] although there are  
61 increasing amounts of oat genome sequence available in databases [7, 12]. Exploitation and  
62 utilization of germplasm resources preserved in wild oat species are a pressing need for oat and  
63 related crop breeding.

64 The genus *Avena* contains about 25 species, including several edible species and invasive  
65 weeds, with characteristic erect culms and solitary spikelets on a panicle, and distribution  
66 throughout temperate regions of the Mediterranean Basin, Africa, Europe, Asia, Australia and  
67 the Americas [13]. Extensive chromosomal rearrangements following recurrent polyploidy  
68 events [14, 15] may have increased oat genomic variation and provided a selective advantage

69 in the adaptation to changing growth environments. A-genome diploid *A. longiglumis* (ALO)  
70 has important traits including the high content of linoleic content in the grains [16], drought-  
71 adapted phenotypes [17], and resistance to crown rust disease [18]. The known genetic  
72 resources from wild diploid *Avena* species are limited, which impedes progress on  
73 understanding the genomic variation related to responses to biotic and abiotic stressors as well  
74 as quality traits. Beyond variation in genes and regulatory sequences, structural and copy  
75 number variation (CNV) has proved difficult to assess with short-read sequencing but its  
76 importance in control of complex traits in both farm animal [19] and plant [20] breeding is  
77 increasingly recognized and must be characterized as part of the pangenome [21].

78 The diploid *Avena* genomes, similar in size to the wheat group but ten times larger than  
79 rice [9] (Ouyang et al., 2007), are characterized by a high proportion of repetitive DNAs, i.e.,  
80 interspersed repeats including transposable elements (TEs) and tandem repeats [22]. Along  
81 with the frequent chromosome translocations [22], the repetitive DNA makes complete genome  
82 assembly difficult. Growing evidence suggests that repetitive stretches of DNAs may cause  
83 sequencing breakage or genome assembly collapse [23]. Now, long-read sequencing  
84 technologies (eg. Oxford Nanopore Technologies, ONT) combined with genome scaffolding  
85 methods (eg. high-throughput chromatin conformation capture, Hi-C), together with Illumina  
86 short-reads used for sequence correction and optical mapping, have improved the genome  
87 contiguity and repeat annotation integrity [24]. For example, hybrid ONT/Illumina study  
88 revealed the genomic landscape among diploid *Brassica* species in unprecedented detail [25]  
89 and allowed *de novo* genome assembly of *Ensete glaucum*, identification of repeats and  
90 chromosomal rearrangements to the related genus *Musa* [26]. High-continuity assembly of  
91 large genomes with a high proportion of repeats is becoming possible [27, 28] and enables  
92 comparison between crops and their wild relatives as well as definition of the pangenome.

93 Here, we generate a high-quality *Avena longiglumis* ALO genome assembly by  
94 integration of Illumina, ONT and Hi-C data, aiming to uncover not only the range of genes and  
95 regulatory elements but also chromosomal rearrangements of wild oat relatives including  
96 features of genome expansion and structural variation. Besides showing structural variation  
97 within *Avena* species, we aimed to identify any intra- and inter-chromosomal rearrangements  
98 compared with the distantly related grasses rice and *Brachypodium distachyon* in the Pooideae.  
99 We aimed to identify gene families that expanded and contracted in ALO and nine grass species  
100 genomes, as well as the potential biosynthetic gene clusters that may play a role in salt stress  
101 response and  $\beta$ -glucan biosynthesis of wild oat species.

## 102 **Results**

### 103 **Genome assembly and annotation**

104 The *Avena longiglumis* (ALO,  $2n = 14$ ) genome size was estimated to be  $4.60 \pm 0.11$  Gbp/1C  
105 by flow cytometry (FCM, Additional file 1: Fig. S1), this value is similar to 4.7 Gbp reported  
106 by Yan et al. [8] and 3.97 Gbp we calculate by  $k$ -mer analysis ( $k = 17$  using raw reads of  
107 Illumina data (Additional file 1: Fig. S2). The ALO genome assembly size we report here is  
108 3.96 Gbp (Table 1), and 3.85 Gbp (97.14%) were anchored into seven pseudo-chromosomes  
109 (Fig. 1; Table 2).

110 Our assembly strategy is shown in Additional file 1: Fig. S3 with  $67.55 \times$  Illumina reads  
111 (150 bp paired-end),  $63.82 \times$  Nanopore and  $99.55 \times$  Hi-C sequencing data (Table 1 [summary  
112 of assembly], Additional file 1: Figs. S2c [sequence depth], d [N50], S3 [assembly strategy]  
113 and S4 [HiC contact map], Additional file 2: Tables S1 [libraries], S2 [programs] and S3  
114 [details of all assemblies]). A total of 252.78 Gbp qualified ONT read sequences (mean\_qscore  
115  $\geq 7$ ) were gained from 12 libraries (Additional file 2: Table S1). After self-correction, the 128  
116 Gbp of read sequences generated 2,379 contigs with a first-pass assembly of 3.71 Gbp with

117 N50 read length 11.92 megabases (Mbp). The ONT raw data and Illumina whole-genome  
118 shotgun sequencing data were used for interactive error correction three or four times, yielding  
119 a 3.96 Gbp polished assembly with a contig N50 of 12.68 Mbp and the longest contig of  
120 99,445,397 bp (summary in Table 1, details in Table S3). We further used 394.30 Gbp Hi-C  
121 data to improve the assembly and anchor 1,974 of 2,379 scaffolds into 414 super-scaffolds  
122 (Additional file 2: Table S3). Super-scaffolds were clustered and ordered into a chromosome-  
123 scale assembly with seven chromosomes, ranging from 453.69 to 594.55 Mbp (Fig. 1, circle a,  
124 Table 2). The chromosome-scale assembly was 3.85 Gbp (97.14% of the with a super-scaffold  
125 N50 of 583.93 Mbp, 97.14% of genomic sequences (Table 2 and supplementary Table S3)  
126 were assigned to discrete chromosome locations using Hi-C assembly (Additional file 1: Fig.  
127 S4, Additional file 2: Table S3). ALO sequencing data were uploaded to National Genomic  
128 Data Center (<https://bigd.big.ac.cn/bioproject/>) under the accession no.  
129 PRJCA004488/CRR275304-CRR275326, CRR285670-285674. The ALO genome assembly  
130 was uploaded to <https://figshare.com/s/34d0c099e42eb39a05e2>.

131 Genome completeness was evaluated by several approaches. Of the Illumina reads,  
132 99.94% (1,348,307,165 of 1,349,168,075) could be mapped onto the assembly after filtering  
133 chloroplast/mitochondrial/bacterial/fungal/human reads. Benchmarking Universal Single-  
134 Copy Orthologs (BUSCO) [29] identified the complete BUSCOs (C, 95.35%), complete and  
135 single-copy BUSCOs (S, 81.02%), complete and duplicated BUSCOs (D, 14.33%), fragmented  
136 BUSCOs (F, 1.16%) and missing BUSCOs (M, 3.49%), respectively (Additional file 2: Table  
137 S5a). With Conserved Core Eukaryotic Gene Mapping (CEGMA) [30], our assembly captured  
138 243 of 248 (98.0%) conserved core eukaryotic genes from CEGMA [30], and 241 (97.18%) of  
139 these were complete [Additional file 2: Tables S4 (assembly consistency statistics), S5a  
140 (BUSCO analysis), b (CEGMA analysis)]. Assembly base accuracy was also assessed based  
141 on Illumina short read mapping. In total, 90.53% of RNA-seq reads were uniquely mapped to

142 the genome assembly (Additional file 2: Table S5c). All of these evaluations indicate the high  
143 completeness, high continuity and high base accuracy of the genome assembly.

144 The long terminal repeat (LTR) Assembly Index (LAI) [31] evaluates the contiguity of  
145 intergenic and repetitive regions of genome assemblies based on the intactness of LTR  
146 retrotransposons (LTR-RTs, Additional file 2: Table S6a, b). The LAI value of the ALO  
147 genome assembly was 10.54, which was higher than that of *Aegilops tauschii* (ATA) [32], and  
148 lower than those of *A. strigosa* (AST, 11.51) [7], *Brachypodium distachyon* (BDI, 11.08) [33],  
149 *Oryza sativa* (OSA, 21.95) [9], *Sorghum bicolor* (SBI, 13.91) [10], *Setaria italica* (SIT, 17.44)  
150 [11] and *Zea mays* (ZMA, 25.57) [34] genomes (Additional file 1: Fig. S5). The LAI indicates  
151 the 3.85 Gbp genome assembly is high quality.

152 A heterozygosity rate of 0.48% was estimated from the frequency peaks of 17-mers from  
153 the Illumina reads following [35], with a main peak depth of 52 [Additional file 1: Fig. S2c  
154 (sequence depth), Additional file 2: Table S4a (Illumina read statistics)]. From mapping  
155 Illumina reads to the assembly, we found 3,914,721 heterozygous SNPs and 185,546  
156 heterozygous indels (4,007,839 polymorphisms / 3,960,768,570 bp) giving a heterozygosity in  
157 single-copy regions of 0.10%, or one polymorphism per kbp (Additional file 2: Tables S4b).  
158 The extremely low value (99.9% homozygous) suggests the species (accession PI 657387) is  
159 strongly inbreeding and this is consistent with reports in other diploid *Avena* species (0.07%  
160 heterozygosity in *A. atlantica* and 0.12% in *A. eriantha*) [12].

## 161 Identification and chromosomal distribution of transposable elements

162 A total of 3.45 Gbp, representing 87% of the ALO genome assembly, could be classified as  
163 repetitive DNAs using EDTA v.1.7.0 [36]. It included 3.34 Gbp (84%) as 2,205,936 complete  
164 or fragmented TEs (Additional file 2: Table S6a) and corresponds to the genome proportion in  
165 other *Avena* species [22]. The overall chromosomal distribution of TEs was relatively uniform  
166 along chromosomes (Fig. 1, circles b, c, d, except for centromeric domains), with no notable

167 depletion in terminal or gene-rich regions (contrasting with other species with both large and  
168 small genomes such as for example wheat [32] or *Ensete* [26]) where there are abundant  
169 transposons and few genes in broad centromeric regions. In ALO, retrotransposons (Class I  
170 TEs) and their fragments were the dominant and accounted for 94.71% of the TE content  
171 (79.76% of the ALO assembly), with LTR retroelements (*Gypsy* and *Copia*) occupying 51.65%  
172 and 26.15% of the ALO assembly, respectively (Fig. 2a, Additional file 2: Table S6a). Among  
173 Class II TEs (DNA transposons), *Helitron* was the most abundant class, constituting 1.11% of  
174 the ALO assembly (Fig. 1, circle e, Additional file 2: Table S6a. There was a notable small but  
175 sharp increase in density of LTR elements at the centromere of all seven chromosomes (Fig. 1,  
176 circle c), with a uniform distribution along the rest of chromosomes. Comparison with Liu et  
177 al. [22] (their Fig. 5d) shows that a tandemly repeated sequence Ab\_T105, shared widely  
178 among *Avena* species, was localized at the centromeres of all chromosomes.

179 Cross-genome comparisons with ATA, BDI, OSA, SIT, SBI and ZMA showed that the  
180 ALO TE content was similar to other published Poaceae genomes analysed with parallel data  
181 and software (Fig. 2a, Additional file 2: Table S6): ATA (81.71% TEs 3.5 Gbp genome size)  
182 and ZMA (81.34%, 1.9 Gbp), and a higher proportion than those of grasses with smaller  
183 genomes [ $< 1$  Gbp [37]; BDI (35.75%, 93.0 Mbp), OSA (47.56%, 185.0 Mbp), SBI (67.83%,  
184 495.2 Mbp), SIT (43.48%, 183.9 Mbp), but in all cases LTR-*Gypsy* elements were more  
185 frequent than LTR-*Copia* elements]. Comparison of the absolute and relative repetitive  
186 sequence composition (Fig. 2a) shows total retrotransposons (*Gypsy*, *Copia*, LINE and  
187 unclassified) were more abundant than DNA transposons in the species analysed, but, relative  
188 to other species, ALO had a lower proportion of DNA transposons (Fig. 2a, red). The ten most  
189 abundant TE families (five *Gypsy*, two *Copia*, LINE, and two DNA TEs) together represented  
190 about 59.66% of the ALO assembly, with the most abundant elements being from *Angela*, a  
191 family of autonomous *Copia* retrotransposons comprising 17.39% of the ALO assembly (Fig.

192 2b). Of the elements identified, 65% more *Copia* were intact (11,021; fragments and intact  
193 elements representing 33% of LTR-elements) compared to *Gypsy* (13,315; 65%). Two LTR-  
194 RT (intact) families, *Tekay-Gypsy* and *Athila-Copia* exhibited elevated abundance in ALO  
195 relative to other grasses (Fig. 2b, Additional file 2: Table S6c).

196 A unimodal distribution was found for the insertion times of all intact LTR-RTs in  
197 analysed grasses (Fig. 2c). The non-domesticated, temperate species with large genomes, ALO  
198 and ATA had the peak of amplification at around 1 Mya and a moderate proportion of older  
199 LTR-RT insertions; a younger peak, occurring approximately 0.1–0.4 Mya, was seen in OSA,  
200 SBI and ZMA (Fig. 2c). The insertion peaks overall have occurred long after the 12 Mya for  
201 ALO speciation and 20 Mya for *Avena* species separation [13, 38], but long before strong  
202 domestication pressures in 9000 to 2000 years ago [39]. Estep et al. [40] emphasized the  
203 contrasting behaviour of individual retroelement families, and indeed in ALO, bursts of both  
204 *Gypsy* and *Copia* show peaks at 1 Mya, but there is a notable additional recent burst of *Copia*  
205 elements only (Fig. 2d). Amplification of retrotransposons (*Copia*, notably *Angela*, and *Gypsy*,  
206 notably *Tekay*, Fig. 2b) contributed directly to the ALO genome expansion. Consistent with  
207 our analysis, previous studies showed that ancestral TE families followed independent  
208 evolutionary trajectories among related species, highlighting the evolution of TE populations  
209 as a key factor of genome expansion [40], and the differential dynamics of TE families within  
210 and between species.

211 Centromere locations were identified by multiple genomic features, some discussed in  
212 detail below. The location of the centromeric retrotransposon *Cereba* [41], SynVisio [42]  
213 results of gaps and conserved regions between *Avena* and other species assemblies,  
214 discontinuities in the Hi-C contact map (Additional file 1: Fig. S4), regions with a relatively  
215 high abundance of TEs (Fig. 1, circles b, c, d), regions of low gene density (Fig. 1, circle i,  
216 Additional file 1: Fig. S6), and examination of chromosome morphology with metacentric, sub-

217 metacentric, and unequal-armed chromosomes, and the secondary constriction at the NOR, all  
218 gave consistent centromere core localizations on chromosomes (Additional file 1: Fig. S7,  
219 Additional file 2: Table S7a, b, c).

220 **Identification and chromosomal distribution of genes**

221 Through a combination of *ab initio* prediction, homology searches, and RNA-seq-aided  
222 prediction, 40,845 protein-coding genes were identified in the ALO genome. Compared with  
223 other published Poaceae genomes, the number of genes in *A. longiglumis* is similar but slightly  
224 greater than that in BDI, OSA, SBI, SIT and ZMA. The mean gene and exon lengths of ALO  
225 genes were 3,272 bp and 268 bp (4.59 exons per gene), respectively (Table 2, Additional file  
226 2: Table S8). A total of 39,558 (96.85%) protein-coding genes were assigned functions, and  
227 86.82% of these genes exhibited homology protein domains in COG (Clusters of Orthologous  
228 Groups of proteins [43]), further 78.59% of these genes exhibited homology protein domains  
229 in Swiss-Prot database. Most of the genes were annotated with the non-redundant protein (NR)  
230 sequence database (96.44% in NCBI NR), and 93.23% of the genes were annotated in NOG  
231 (Non-supervised Orthologous Groups). The 87.09% of genes were annotated with Pfam [44,  
232 45] annotation, 51.81% being classified according to GO terms [46], 42.32% being mapped to  
233 known plant biological pathways based on the KEGG pathway database [47], 5.88% being  
234 annotated in PlantTFDB v.5.0 [48] and 2.28% in CAZy database [49] (Additional file 2: Table  
235 S9). In addition, we predicted at least 17,712 noncoding RNAs consisting of transfer RNAs  
236 (0.0027%), microRNAs (0.0532%), and small nuclear RNAs (0.0016%) (Additional file 2:  
237 Table S10). A total of 33,271 (81.46%) high-confidence (HC) and 7,574 (18.25%) low-  
238 confidence (LC) protein-coding genes were annotated based on *de novo* prediction, homology  
239 annotation, and RNA sequence data (Table 1, Additional file 2: Table S11a, b, c).

240 Ribosomal DNAs (rDNAs) were collapsed in the assembly. The 45S rDNA monomer was  
241 10,215 bp long, representing 0.30% of the genome (1,160 copies) and located on ALO07 (*A.*

242 *longiglumis* chromosome 7) around 429,873,253 bp with a minor site on chromosome ALO01  
243 at 476,793,000 bp (Fig. 1, circle a). The 5S rDNA monomer was 314 bp long, representing  
244 0.005% of the genome (584 copies), with a single locus located on chromosome ALO07 around  
245 173,926,000 bp.

246 Gene density along chromosomes varied, with broad regions both sides of the centromeres  
247 being depleted of genes (Fig. 1, circle i) in six of the seven chromosomes. Notably, ALO07  
248 with the major NOR region (45S rDNA site) had a different pattern: on the long arm, there was  
249 a similar density of genes along most of the arm. Very few genes were identified between the  
250 centromere and NOR locus, while the satellite had a high gene density.

## 251 Genome evolutionary and whole genome duplication (WGD) analysis

### 252 *Orthologous genes in Avena and related grasses*

253 We clustered the annotated genes into gene families among ALO and nine grass species (AAT,  
254 *Avena eriantha* (AER) [12], AST, ATA, BDI, OSA, SBI, SIT and ZMA) with *Arabidopsis*  
255 *thaliana* (ATH) [50] as the outgroup using Orthofinder v.2.3.14 [51]. A total of 1,880 single-  
256 copy genes (Additional file 2: Tables S12 and S13) were identified among 11 species, which  
257 were used for phylogenetic reconstruction (Fig. 3a, b). ALO is sister to the lineage of AAT and  
258 AST, and in turn clustered with AER, ATA, BDI and OSA, subfamily Pooideae (Fig. 3a, b,  
259 Additional file 2: Table S14). We found that ALO diverged phylogenetically from the lineage  
260 of AAT and AST at 2.34 (1.17–3.69) Mya after the divergence of *Avena* at 22.46 (15.12–29.08)  
261 Mya (Fig. 3a). The divergence times are consistent with phylogenies based on morphology or  
262 chloroplast and single-copy nuclear genes, showing when *Avena* diverged from ATA (20.04  
263 Mya) [13] or wheat (19.90 Mya) [38]. The C-genome diploid *A. eriantha* diverged from A-  
264 genome diploid species at 9.34 (5.29–13.48) Mya.

265 We identified homologous gene pairs in ALO, AER, AST, OSA and ZMA genomes and  
266 estimated species divergence times by analysis of synonymous nucleotide substitution rates

267 (Fig. 3c). The results indicated that some gene pairs within *Avena* showed a peak at 0.05–0.13,  
268 probably reflecting the whole genome duplication (WGD) event in *Avena*, with an additional  
269 shallow peak at 0.8 reflecting the ancient rho ( $\rho$ ) WGD event [52] (Fig. 3c). ALO comparison  
270 with OSA (0.54) and ZMA (0.61) showed a single Ks peak. Based on sequence homology  
271 among the analysed species, we assigned gene number ranging from 27,416 of ATH to 49,542  
272 of ATA and gene number in families ranging from 24,421 of ATH to 44,363 of AAT,  
273 respectively (Additional file 2: Tables S12 and S13). A total of 2,277 gene families (845 in  
274 AAT, 355 in AST, 259 in ALO, and 818 in AER) were unique to *Avena* species (Additional  
275 file 2: Table S13). Five species (ALO, AER, AST, OSA and ZMA) were selected to identify  
276 unique and shared gene families in the Poaceae. We found a total of 14,548 shared orthologous  
277 gene families, notably with more than a third ( $5,423 = 2,421 + 1,337 + 1,061 + 604$ ; Fig. 3d)  
278 gene families unique in all *Avena* species, emphasizing the novel gene pool in *Avena*.

### 279 *Comparative genomics of gene families*

280 Based on sequence homology, from the OrthoFinder [51] analysis, the 35,039 genes in families  
281 identified across ALO and 10 plant species genomes, we identified gene families showing  
282 expansion (1,440) and contraction (962) after the *Avena* divergence from ATA (Fig. 3a,  
283 Additional file 2: Table S14a). GO enrichment analysis revealed that the expanded genes of  
284 ALO were notably enriched ( $p < 0.05$ ) in molecular functions associated with terpenoid  
285 biosynthesis and polysaccharide binding, cellular components such as DNA packaging and  
286 signal recognition particle, as well as biological processes associated with hormone, water,  
287 biotic and abiotic stimulus (Additional file 2: Table S15). KEGG analysis showed that the  
288 expanded genes of ALO were involved in biosynthesis of other secondary metabolites (09110),  
289 metabolism of other amino acids (09106), metabolism of terpenoids and polyketides (09109),  
290 environmental adaptation (09159), signal transduction (09132), lipid metabolism (09103) and  
291 membrane transport (09131) at hierarchy B level, while the expanded genes of AAT were

292 involved in translation (09122), energy metabolism (09102), and environmental adaptation  
293 (09159), the expanded genes of AST were involved in replication and repair (09124), energy  
294 metabolism (09102) and folding, sorting and degradation (09123), while the expanded genes  
295 of AER were involved in energy metabolism (09102) and translation (09122) (Additional file  
296 1: Fig. S8a, b, c and d, Additional file 2: Table S16).

297 We also examined the unique gene families in *Avena* species: 845 gene families were  
298 unique in AAT, 355 in ALO, 259 in AST and 818 in AER (consistent with 1,100 genes unique  
299 to AER, 420 in ALO to 539 in AST; Fig. 3d, Additional file 2: Table S13). At hierarchy B  
300 level, genes associated with energy metabolism (09102), carbohydrate metabolism (09101) and  
301 membrane transport (09131) were uniquely enriched in ALO, and carbohydrate metabolism  
302 (09101), membrane transport (09131) and amino acid metabolism (09105) were uniquely  
303 enriched in AAT, while transcription (09121) was uniquely enriched in AST (Additional file  
304 1: Fig. S9a, b, c). Compared with the A-genome species, genes associated with folding, sorting  
305 and degradation (09123), glycan biosynthesis and metabolism (09107), transport and  
306 catabolism (09141), energy metabolism (09102) and biosynthesis of other secondary  
307 metabolites (09110) were uniquely in AER (Additional file 1: Fig. S9d). The expansion of gene  
308 families occurs during a long-term evolution and drives the evolutionary difference between  
309 wild oat species.

310 **Ancestral linkage group evolution**

311 *Intraspecific syntetic blocks in Avena longiglumis*

312 To determine the chromosome structure in ALO, we performed an intragenomic synteny  
313 analysis. About 15 major syntetic blocks exist between pairs of ALO chromosomes based on  
314 paralogous genes (so gene-poor regions around centromeres are not represented). Examples of  
315 shared major blocks of paralogous genes between ALO01 (*A. longiglumis* chromosome 1) and  
316 ALO02; ALO03 and ALO05; ALO04 and ALO06 (Fig. 1, centre m, Additional file 1: Fig.

317 S10a, d). The pairs of synteny blocks are likely to identify the signature of the ancient ρ WGD  
318 event in the grasses. The block covered most of the gene-rich parts of the genome with no  
319 indication of major deletions following WGD. There were no regions with three or more  
320 copies, suggesting no major segmental duplications. In contrast to *Musa* (D'Hont et al. [53]  
321 their supplementary Fig. S12) and *Ensete* [26] that do not share the grass ρ WGD event, there  
322 is clear evidence for two rounds of WGD in *Avena*.

### 323 *Avena* *Intergeneric chromosome rearrangements*

324 To investigate the relationship between ALO, AST (both A-genome, Additional file 1: Fig.  
325 S10b) and AER (C-genome, Additional file 1: Fig. S10c), we conducted a synteny analysis  
326 between *Avena* genomes, all sharing the same WGD events. We found a total of 29,030, 27,116  
327 and 21,536 pairs of collinear genes between ALO-AST, ALO-AER and AER-AST species  
328 pairs, respectively (Fig. 4a, Additional file 1: Fig. S11a, b, c, Additional file 2: Table S17).  
329 Visualization of regions of synteny between ALO, AST and AER, with SynVisio [42] shows  
330 large blocks of conservation between ALO and AST, with much more rearrangement with the  
331 more distant AER (Fig. 4a). Between ALO and AER, chromosome ALO01 was largely  
332 collinear with AER03, and ALO06 with AER02 (Fig. 4a; Additional file: Fig. S11a, b). Other  
333 AER chromosomes had multiple synteny regions, each involving most ALO chromosomes.  
334 Whether the higher level of structural variation reflects synteny gene clusters and the  
335 adaptation to a sandy-soil and arid environment of AER needs further investigation. Most  
336 notably, numerous evolutionary inter-chromosomal translocations represented 10% to 25% of  
337 the length of nearly all chromosomes with many in large distal domains (Fig. 4a, Additional  
338 file 1: Fig. S11a, b, c). Interestingly, these distal (sub-terminal) intragenomic evolutionary  
339 rearrangements, identified here for the first time in diploid species, are entirely consistent with  
340 distal nature and size of translocations identified using genome-specific repetitive DNA

341 sequence probes in polyploids [22] where translocations between genomes have occurred since  
342 the polyploidy event.

343 In order to understand the structural chromosomal variation including duplications and  
344 deletions, we examined the extent and nature of chromosomal rearrangements across Pooideae  
345 species. Two species with small genome sizes, OSA ( $x = 12$ ) and BDI ( $x = 5$ ) have been used  
346 extensively as reference genomes. Previous studies have suggested that Poaceae genomes  
347 evolved from a pre- $\rho$  WGD ancestral grass karyotype (AGK), with 7 protochromosomes, to a  
348 post- $\rho$  AGK with 12 protochromosomes [14]. Conserved genes for each AGK chromosome  
349 have been identified (with the proposed 12 protochromosomes showing extensive similarities  
350 with rice). Here, the AGK genes were mapped to the chromosomes of ALO and six grass  
351 species (AAT, AST, AER, ATA, BDI and OSA), and the corresponding regions of  
352 chromosomes were designed by a colour (Fig. 4b). The signature of the ancient  $\rho$  duplication  
353 (see also in Fig. 1, centre m) is shown by pairs of chromosomes with shades of similar colours.  
354 Because of divergence in gene sequences, there is some ambiguity in assignment of extant  
355 chromosome blocks to the duplicated chromosomes in the ancestor (eg. the shades of orange  
356 from AGK01 and AGK05 in ATA03) and so cannot be interpreted as clearly. Dotplots of ALO  
357 against OSA and BDI (Additional file 1: Fig. S11d, e and f) were also made as the reference  
358 for a SynVisio [42] plot of ALO with OSA and BDI (Fig. 4c).

359 Large segments of the ancestral chromosomes are conserved across the analysed grasses  
360 with distinct rearrangements involving translocations and fusions of syntenic blocks between  
361 the species. Some rearrangement events are shared between all  $x = 5$  and  $x = 7$  species (e.g.,  
362 the fusion of AGK09 and AGK11; or AGK02 and AGK03; both are seen in BDI, ATA and the  
363 *Avena* species) or between the  $x = 7$  species (AGK12 and AGK06 giving ALO07; Fig. 4b,  
364 Additional file 2: Table S17).

365 While some evolutionary events from the ancestral 12 AGK chromosomes involve fusion  
366 and rearrangement of syntenic blocks, it is notable that three events are largely characterized  
367 by insertion of one chromosome (group of syntenic genes) into another chromosome. Thus,  
368 ALO04 has AGK07 inserted into AGK04; ALO06 has much of AGK06 inserted into AGK02;  
369 while ALO05 has insertion of AGK08 into ALO06. The inheritance of fusion events is  
370 consistent with the phylogeny (Fig. 3a) and time since separation from the most recent common  
371 ancestor with the proposed ancestral grass karyotype. Given the higher number of  
372 rearrangements from the AGK, Fig. 4b suggests AER is the most derived karyotype in *Avena*  
373 and the A-genome (including ALO) is more primitive. Overall, during evolution, chromosome  
374 rearrangement has been restricted to a small number of events, presumably avoiding alteration  
375 of interactions of gene groups and promoters along chromosomes without fusion and fissions.

### 376 **Genome-wide expression analysis**

377 As well as use for gene annotation, we analysed the relative transcription of genes in roots,  
378 salt-treated roots, leaves and salt-treated leaves of ALO to suggest key features of gene  
379 expression level variation. Differentially expressed genes (DEGs; fold change  $\geq 2$  and  
380  $FDR \leq 0.005$ ), comprised 17.48% of genes (3,076 up-regulated DEGs and 3,963 down-  
381 regulated DEGs; Additional file 2: Table S18a). KEGG analysis indicated that the 1,569 up-  
382 regulated genes in salt-treated roots enriched in environmental adaptation (09159), metabolism  
383 of terpenoids and polyketides (09109), biosynthesis of other secondary metabolites (09110),  
384 signal transduction (09132) and membrane transport (09131), while the 1,627 down-regulated  
385 DEGs were enriched in salt-treated roots in biosynthesis of other secondary metabolites  
386 (09110), replication and repair (09124), metabolism of other amino acids (09106) and  
387 carbohydrate metabolism pathways (09101) (Additional file 1: Fig. S12a and b). In salt-treated  
388 leaves, there were 1,507 up-regulated DEGs enriched in terpenoid backbone biosynthesis  
389 (00900) and transcription (09121) pathway, while 2,336 down-regulated DEGs enriched in

390 energy metabolism (09102), carbohydrate metabolism (09101), amino acid metabolism  
391 (09105), metabolism of terpenoids and polyketides (09109), membrane transport (09131),  
392 signal transduction (09132), metabolism of cofactors and vitamins (09108) and environmental  
393 adaptation (09159) pathways (Additional file 1: Fig. S12c and d).

394 KEGG enrichment pathways of up- and down-regulated DEGs were very similar in ALO  
395 expanded gene families, and most genes involved with salt adaptation mainly belonged to the  
396 expanded gene families. The environmental adaptation, metabolism terpenoids and polyketides,  
397 membrane transport pathways were enriched in up-regulated DEGs of salt-treated roots, while  
398 these pathways were enriched in down-regulated DEGs of salt-treated leaves, which may be  
399 related to different responses to salt stress between aboveground and underground parts of  
400 plants. Additionally, the pathways related to terpenoid synthesis were enriched in both roots  
401 and leaves, in which they may be extensively participated salt-tolerance of plants [54].

402 Further investigate the resilience effect of DEGs on the environmental adversity of ALO,  
403 we conducted transcriptional analyses of 4,329 expanded gene families (7,236 expanded genes,  
404 Additional file 2: Table S14). Expanded DEGs comprised 34% of total DEGs (1,352 expanded  
405 DEGs in salt-treated roots and 1,093 expanded DEGs in salt-treated leaves, Additional file 2:  
406 Table S18b, c). The number of DEGs in roots was slightly higher than in leaves.

407 We analysed the gene function of DEGs among expanded gene families in roots and  
408 leaves of ALO. The 599 up-regulated, expanded genes in salt-treated roots included protein  
409 kinase, cytochrome P450 (CYP450), cupin, and pathogenesis-related protein genes, and 753  
410 down-regulated expanded genes in salt-treated roots included nucleosome histone, protein  
411 kinase, transferase and other *CYP450* families (Additional file 1: Fig. S13a). Expanded DEGs  
412 showed relatively fewer in salt-treated leaves, the 451 up-regulated genes including zinc finger  
413 protein, protein kinase, peptidase and other genes, and the 642 down-regulated DEGs in salt-  
414 treated leaves including protein kinase, CYP450 and receptor kinase (Additional file 1: Fig.

415 S13b). For *CYP450* genes in salt-treated samples, 27 up- and 22 down-regulated DEGs were  
416 detected in salt-treated roots, and 17 down-regulated genes were detected in salt-treated leaves  
417 of ALO.

418 *Analysis of cytochrome P450 (CYP450) gene clusters*

419 In total we identified 109 biosynthetic gene clusters (BGCs) in the ALO genome, including  
420 alkaloid, lignin, polyketide, saccharide and terpene biosynthesis genes (Fig. 5a). The result  
421 (Additional file 2: Table S19) allows assessment of each locus for its likelihood to encode  
422 genes working together in one pathway [55]. To evaluate the potential of ALO for the genetic  
423 dissection of agriculturally important traits, we focus on the evolution of triterpene synthesis,  
424 including clusters of the *CYP450* gene families, which encode proteins involved in multiple  
425 metabolic pathways with complex functions and playing important roles in defence responses  
426 to abiotic stresses. The number of *CYP450* genes in ALO (557; 1.36% of total genes) was  
427 significantly higher than in nine analysed grasses (251–454) and ATH (246; Additional file 2:  
428 Table S20a). Overall, the *CYP450* genes were relatively equally distributed on all  
429 chromosomes (average 80, between 73 and 100 observed) (Additional file 2: Table S20b). The  
430 transcript level of one *CYP450* gene (AL2G04509) was over 913-fold in salt-treated roots than  
431 roots; and AL7G05074 was 409-fold in salt-treated leaves than leaves (Additional file 1: Fig.  
432 S14, Additional file 2: Table S21).

433 Using the presence of at least one orthogene in the identified gene clusters as the selection  
434 criteria, we assigned 46 putative *CYP450* gene clusters (Additional file 1: Fig. S15, Additional  
435 file 2: Table S22). Five key gene clusters identified in the ALO genome were CL10 and CL95,  
436 CL37, CL98 and CL106, which included functionally characterized UDP-glycosyl transferase  
437 (AS01G000200), serine carboxy peptidase-like acyltransferase (AS01G000190), subtilisin  
438 homologue (AS01G000130), O-methyltransferase (AS01G000040) together with enzymes  
439 annotated as *CYP450* (Li et al. [7] their supplementary table 7), dehydromodolichyl diphosphate

440 synthase, aldehyde oxidase and hydrolase proteins (Additional file 1: Fig. S16a, b, c, d, e, f and  
441 g).

442 We examined the gene number and conserved synteny around *CYP450* gene clusters in  
443 the four *Avena* species (sharing 24 *CYP450* genes) and 10 grass species (sharing six *CYP450*  
444 genes; Additional file 2: Table S23a, b). In ten clusters (Additional file 2: Table S22), *CYP450*  
445 genes present across four *Avena* species with loss of copies and without syntenic relationship  
446 among six grass species. We identified a further 11 *CYP450* gene clusters containing terpene  
447 synthase genes (Additional file 2: Table S22) with *CYP450* genes together with terpene  
448 synthase genes, showing conservation across all species including *Avena* suggesting these gene  
449 clusters related to functional expansion of specialized terpene metabolism [55]. Tandem  
450 duplications within 18 *CYP450* gene clusters (Additional file 1: Figs. S16a, b, c, d, e, f and g)  
451 were revealed in ALO. These gene clusters, with known functionally characterized genes  
452 involved in *CYP450* biosynthesis and extensive copy number variation (CNV) between  
453 species, can be taken to present the pangenome for *CYP450* biosynthesis [56] (Additional file  
454 1: Fig. S16a and b).

455 Overall, the analysis provides strong support for the non-random organization of *CYP450*  
456 biosynthetic genes and presence of *CYP450* gene clusters [7, 57] in *Avena* and other grasses.  
457 Indeed, the high-continuity ALO genome assembly shows the avenacin cluster (including  
458 antimicrobial terpene biosynthetic genes), in terms of both gene number and the diversity of  
459 gene families that it contains, was even stronger than the triterpene BGCs identified previously  
460 [7].

461 **Phylogenetic analysis of the *CesA* and *Cs* gene families**

462 *Identification of cellulose synthase (CesA) and cellulose synthase-like (Csl) gene*  
463 *families in ALO*

464 To gain insight into whether the physical location plays a role in expansion of  $\beta$ -glucan  
465 biosynthesis genes, we evaluate the genomic organization of *cellulose synthase A (CesA)* and  
466 *cellulose synthase-like (Csl)* gene families (Additional file 1: Figs. S17 and S18). They encode  
467 1,4- $\beta$ -glucan synthase superfamily serving as the predominant structural polymer in primary  
468 and secondary cell walls of caryopses [58]. Dataset searches using conserved Pfam motifs  
469 PF000535 and PF03552 [44], which are specific to the glycosyltransferase GT2 superfamily  
470 [59], resulted in the identification of 11 *CesA* and 55 *Csl* genes (Additional file 2: Table S24).  
471 The maximum likelihood (ML) tree (Fig. 5b) for CesA (a single branch with 11 proteins in  
472 ALO) and Csl proteins from ALO, *Arabidopsis thaliana*, rice, wheat and maize (Additional  
473 file 1: Fig. S19) shows the ALO Csl proteins group in seven subfamilies: CslA (10 proteins),  
474 CslC (6 proteins), CslD (8 proteins), CslE (11 proteins), CslF (13 proteins) and CslH (1 protein)  
475 and CslJ (1 protein), with five proteins unclassified. The closely related CslA and CslC  
476 subfamilies were conserved across the species, as were the sister sub-families of CslD and the  
477 grass-specific CslF.

478 *CesA* and *Csl* genes were relatively equally distributed over all chromosomes (average 9,  
479 between 6 and 13 observed) (Additional file 2: Table S25). There were large differences in  
480 expression of *CesA* and *Csl* genes between roots, salt-treated roots, leaves and salt-treated  
481 leaves (Fig. 5c) with the fragments per kilobase of exon model per million mapped reads  
482 (FPKM) showed ratios up to 150 that are suggestive of their functional role (Additional file 2:  
483 Table S26). Some genes presenting higher levels of expression in roots than leaves, may be  
484 involved in  $\beta$ -glucan synthesis [60]. Among 109 gene clusters, we identified 10 metabolic gene  
485 clusters representing 2 *CesA* and 10 *Csl* gene models across six of seven chromosomes  
486 (Additional file 1: Fig. S17, Additional file 2: Table S27). Synteny analysis showed the

487 conservation among *CesA* and *Csl* genes of gene clusters: four *Avena* species shared 1 *CesA* and  
488 3 *Csl* genes, and 10 grass species shared two *Csl* genes (Additional file 2: Table S28a, b). In  
489 four clusters of *Csl* genes present across four *Avena* species, the loss of copies, and lack of  
490 syntenic relationships to six other grass species, suggests the opportunity to exploit the (1,4)-  
491  $\beta$ -glucan biosynthesis pathway outside oats may be limited (Additional file 2: Table S24). The  
492 two gene clusters (CL32 and CL58) contained alkaloid and saccharide biosynthetic genes  
493 showed the conserved synteny relationship with other grasses. Tandem duplications within  
494 *CesA* and *Csl* gene cluster were observed in ALO (Additional file 1: Fig. S18a). It is thus  
495 important to study the role of individual *CesA* and *Csl* in primary and secondary cell wall  
496 biosynthesis to attempt effective modification of biomass composition.

497 Comparing the cellulose-like synthesis clusters with homologous genomic loci in AAT  
498 genome can give important information on its evolutionary conservation or diversification  
499 (Additional file 2: Tables S28, S29 and S30). Whereas strong conservation of clusteredness  
500 across larger periods of evolutionary time may point to a selective advantage of clusteredness  
501 for these genes, diversification of *Csl* genes by co-option of glyoxalase genes may give clues  
502 to find novel variants of natural products that have been generated through directional pathway  
503 evolution (Additional file 1: Fig. S18a, b, c, d, e and f). A better understanding of the cell wall  
504 gene expression under abiotic stress is important to design strategies to produce crops in  
505 marginal lands with less  $\beta$ -glucan accumulation. Gene families play an important role in  
506 enhancing salt-tolerance and adaptation of ALO, which was also found in desert plants [61].

### 507 *Identification of callose synthase (CalS) enzyme families in ALO*

508 Callose (1,3- $\beta$ -glucan), encoded by the *callose synthase (CalS)* or *Glucan synthase-like (GSL)*  
509 gene families, plays an important role in plants grown both normal and unfavourable  
510 environments [62]. Dataset searches for ALO using conserved Pfam motifs PF02364 and  
511 PF14288, identified 13 CalS or GSL proteins. With the *CalS* genes, distributed along five

512 chromosomes (ALO01/03/05/06/07, with no notable chromosomal clusters), and five *CalS*  
513 genes were identified in the expanded gene families of ALO (Additional file 2: Table S29).  
514 The transcript level of *CalS* gene (AL6G02277) was 11-fold higher in salt-treated roots than  
515 roots, while AL3G02326 was one-fold in salt-treated leaves than leaves (Additional file 1: Fig.  
516 S20, Additional file 2: Table S30) supporting involvement in plants' resilience to salt stress.

517 *Comparative phylogenetic analysis of the CalS gene family*

518 The 13 CalS or GSL proteins were placed in a ML tree (Additional file 1: Fig. S21) along with  
519 23 GSL proteins from *Arabidopsis thaliana* and rice (Additional file 1: Fig. S22). The analysis  
520 grouped the proteins into eight clades, with Clades VII and VIII, and Clade II (*Arabidopsis*)  
521 and III (rice and *Avena*), being sisters. Five clades (IV to VIII) included all plant species  
522 suggesting diversification present in the common ancestors, with further duplications reflecting  
523 ancient whole genome duplication events ( $\alpha$ ,  $\beta$ ,  $\gamma$  in eudicots, and  $\tau$ ,  $\sigma$ ,  $\rho$  in monocots) or more  
524 recent segmental duplications.

525 **Discussion**

526 The diploid wild oat, *Avena longiglumis* (ALO), with distribution around the Mediterranean  
527 Basin, is an important genetic resource for oat breeding, and a valuable reference for genomic  
528 organization and evolution in the grasses. We used a combination of Nanopore (436 Gbp),  
529 Illumina (269 Gbp) and Hi-C (331 Gbp) sequencing technologies to assemble the  
530 3,847,578,604 bp long ALO genome into the seven pseudo-chromosomes (Fig. 1, Tables 1 and  
531 2). The high-quality ALO genome assembly gives not only insight into the gene diversity but  
532 also into the variation in repetitive DNA content and structural variation (SV) in the genome  
533 including chromosome duplication and arrangements.

534 Assembled plant genome sizes range from 61 Mbp (*Genlisea tuberosa*, bladderwort [63])  
535 to 26,454 Gbp (*Sequoia sempervirens*, coast redwood [64]). The ALO assembly falls into this

536 range, and a combination of ONT, Illumina and Hi-C sequence approaches was essential for  
537 the high continuity chromosome level assemblies with high presence of core genes, as is for  
538 many important crop genomes, particularly cereals, with genomes larger than 2 Gbp (barley  
539 [65]; wheat [66, 67]). *Avena*, as all large genomes, includes abundant copies of TEs [22] and  
540 the long-read sequencing technology allowed examination of their organization (Figs 1 and 2).

541 Most transposable element classes are distributed widely and rather uniformly along the  
542 ALO chromosomes (Fig. 1). In many species with smaller genomes (<3.96 Gbp of ALO), broad  
543 pericentromeric regions are reservoirs for the accumulation of a medley of (often lineage-  
544 specific) TEs [26, 68, 69], but in ALO, the overall TE density is relatively similar along the  
545 chromosomes. However, there is a high and localized abundance of an LTR retrotransposon  
546 (Fig. 1 circle c) at the centromeres of all seven ALO chromosomes. Such contrasting  
547 distributions of LTR retrotransposon clades has been found in several species [40, 70, 71, 73].  
548 Centromeres of some species harbor arrays of tandemly repeated satellite sequences (e.g.,  
549 *Arabidopsis thaliana* [74, 75], *Beta vulgaris* [76] and *Ensete glaucum* [26]), but we found no  
550 equivalent tandem repeat in ALO. However, often in genomes with centromeric satellite  
551 sequences, abundant families of retroelements are also found at the centromeres, such as the  
552 *Nanica* LINE of *Musa acuminata* [77] and *E. glaucum* [26], *Arabidopsis* retroelement domains  
553 [67] or the wheat *Quinta* and other elements [32, 78].

554 In a short evolutionary timescale, young TEs (< 2 Mya [79]) were frequent (Fig. 2) in  
555 ALO. LTR retrotransposons may be beneficial to their hosts by providing regulatory genetic  
556 elements [80, 81] or by disruption of genes and their promoters. While TEs are unlikely to be  
557 the only causal factor responsible for subgenome expression dominance in polyploids,  
558 methylated TEs can reduce the expression of nearby genes [82]. Further studies are needed to  
559 address whether oat A-genome dominance is determined by methylation pattern differences of  
560 retrotransposons [83], and their contribution to genetic variation in different *Avena* species.

561 The  $\rho$  WGD events occurred 50–70 Myr ago, after Poales separated from other monocot  
562 orders [4, 84, 85]. Based on detailed paleogenomics, using inference from  $x = 5$ –12 grasses in  
563 terms of gene order and content, Murat et al. [14] proposed an ancestral grass karyotype (AGK)  
564 with similarity to the extant *Oryza sativa* ( $2n = 2x = 24$ ) genome including 14,241 conserved  
565 genes (Fig. 3). We delineated genome sequences between OSA, BDI, ATA and four *Avena*  
566 species, representing four tribes and different polyploidization events, confirming that the  $\rho$   
567 Poaceae event is shared by the ancestral BOP clade and Poales (Fig. 4). Most notably, the  
568 conservation of large syntetic blocks and the orthologous relationships of the seven extant  
569 ALO chromosomes to the 12 chromosomes of OSA and AGK was evident, with defined fusion  
570 and translocation events but limited major duplications or deletions. Apart from the long-term  
571 evolutionary conservation, such regions harbour conserved sequence regions that might be  
572 synthesized as oligonucleotides for *in situ* hybridization to label linkage group 1 across all  
573 Poales grasses [86], and to use as baits (cf. <https://treeoflife.kew.org/methods> and Johnson et  
574 al. [87]) to identify the variation of all AGK01 genes across the group. Other chromosomes  
575 have well-defined range of fusions from the AGK or OSA reference, reducing the chromosome  
576 number from  $x = 12$  to  $x = 5$  or  $x = 7$ , but notably some chromosomes have evolutionary  
577 insertion of one ancestral chromosome into another.

578 The conservation of many syntetic blocks and the chromosome structure occurs despite  
579 of the huge expansion in genome size, with ALO being ten times larger than OSA ( $15.2 \times$   
580 BDI). Most notably, there is expansion in genome size throughout the chromosomes, largely  
581 involving the amplification of retroelements that are dispersed uniformly along all chromosome  
582 arms (Fig. 1) and is evidenced by the lines of synteny between ALO and the corresponding  
583 BDI and OSA syntetic blocks spreading out relatively uniformly over a much-expanded region  
584 of ALO. There are a few gene rich regions (on chromosomes ALO01, ALO04, ALO05 and  
585 ALO07) shared with BDI but not OSA that are worth further investigation. Overall, the genome

586 structure revealed in the syntenic comparison reveals the evolutionary history of the Poales at  
587 the chromosome level, and encourages exploitation of the whole gene pool in both biodiversity  
588 studies and for plant breeding.

589 *In situ* hybridization using repetitive DNA probes has shown that many chromosomes in  
590 the hexaploid *A. sativa* show intergenomic translocations (i.e., between chromosomes of the  
591 diploid ancestral genomes [16, 88]), involving the terminal 10% to 25% of many chromosome  
592 arms. Such translocations have not been seen in the tribe Triticeae (syn. Hordeae, sister tribe  
593 to Aveneae). Remarkably, the three *Avena* species (ALO, AST and AER) have multiple  
594 terminal translocations between seven chromosomes of three species (Fig. 4), occurring only  
595 on one arm. Six chromosomes are involved in clear non-reciprocal translocations between ALO  
596 and AST, but no terminal regions have been lost during the translocation events. With respect  
597 to the evolutionarily more distant AER chromosomes, every ALO chromosome has a terminal  
598 translocation as well as a greater number of other rearrangements. The terminal rearrangements  
599 do not only involve repetitive DNA sequences, as is likely to be the case in maize (e.g., The  
600 P53 knob [89]) or rye (pSc250 tandem repeat sequence [73]), but also involve many genes  
601 within syntenic groups [90].

602 Poales species occupy differentiated environmental and ecological niches, with  
603 contrasting selective pressures so we looked at groups of syntenically conserved genes where  
604 phenotype and selection may be affected. ALO is restricted to sandy loam soils and mesic  
605 habitats in the Mediterranean desert, while AER populations thrive on shallow calcareous hills  
606 or terra rossa soil steppes around the Mediterranean Basin [91]. Notably, while there are few  
607 insertions or deletions between ALO and AST, there are many gaps, but not break in synteny  
608 along chromosome arms between ALO and AER (Fig. 4). It will be interesting to see if these  
609 regions are related to functional or selective changes in copy number and the unique paralogues  
610 of AER (Fig. 3).

611 The physical clustering of multiple genes from a single metabolic pathway is now  
612 established in plants [7, 92]. Clustering should favour co-inheritance of beneficial combination  
613 of alleles that confer a selective advantage together [93]. Our results show clustering of genes  
614 and regulators including terpene (Fig. 5), cellulose or phytohormone pathway enzymes. In  
615 terpene and cellulose clusters, *CYP450s* exhibit down-regulation among different tissues, while  
616 most are considered as highly tissue-specific genes. The common expression trends of  
617 homologous genes also exist in wheat and maize, implying a unique highly conserved function  
618 for each clustered gene [93]. Consistent with the model, our survey-expression data indicates  
619 some *CYP450s* are up-regulated, and others are down-regulated under salt stress, suggesting  
620 the need for detailed investigation of *CYP450* functions under salt stress [93].

621 The *CesA/Csl* gene families play a critical role in the biosynthesis of cellulose and  
622 hemicellulose. We identified 66 *CesA/Csl* genes which could be divided into four lineages in  
623 ALO. Orthologous genes (in different species) can be more similar than paralogous genes (of  
624 the same species), eg, the *Arabidopsis* (dicot)-specific CslB lineage was closer to grass-specific  
625 CslH lineage than CslF lineage, suggesting that CesAs and Csls diverged before the split of  
626 monocots and eudicots, c. 150 Mya [94]. This indicates that the *CesA/Csl* genes established  
627 their roles early in higher plant evolution, and could be a reason why there are so many  
628 *CesA/Csl* gene families in Poaceae. Our findings indicate that the larger *CesA/Csl* superfamily  
629 is the consequence of recent duplications (Fig. 3), and particularly chromosomes ALO02 and  
630 ALO06 have more *Csl* genes than other grass species.

631 Some expanded *CesA/Csl* genes may be retained simply owing to sub-functionalization  
632 where the functions of the ancestral genes were partitioned among the duplication. In the case  
633 of *CslD1* subfamily members, which are involved in root hair-deficient phenotypes of maize  
634 [95], some *CslD1* copies can be lost without any phenotypic consequences. Intragenic  
635 complementation has not been observed among alleles with mutations in different *CSLD1*

636 domains of *Lotus japonicas* [96], suggesting that the *CSLD1* has not yet undergone the  
637 complete functional differentiation. *CslF6* and *CslH* have a functional role in the synthesis of  
638 mixed-linkage (1,3;1,4)- $\beta$ -glucan (MLG [97]), and the sequence divergence of the *Csl* genes  
639 we found is likely a reflection of their functional divergence although MLG synthesis is tightly  
640 regulated and thus maximizing the yield of end-product cellulose might be difficult. For  
641 example, isoforms may utilize the same donor but a different acceptor molecule in the synthesis  
642 of the same polysaccharide, and thus, having multiple genes may be a requirement for synthesis  
643 of some types of plant polysaccharides [58].

## 644 **Conclusions**

645 The 3.85 gigabase sequence assembly of the wild oat species *Avena longiglumis* has enabled  
646 chromosome evolution to be defined within *Avena* and diverse Poaceae species. The diversity  
647 revealed in gene and gene network will accelerate the analysis of trait genes and their control.  
648 Beyond diversity in genes and regulatory sequences, the spectrum of chromosomal structure  
649 variation and sequence copy number variation (both of genes and repetitive DNAs), can be  
650 shown by comparison with our high-continuity genome assembly, and will enable  
651 characterization of the *Avena* and broader grass pangenome. There is increasing recognition of  
652 the role of structural and copy number variation in diversity, going beyond the well-studied  
653 differences in gene alleles, networks, and transcription factors, both in plants and animals [19,  
654 20].

655 Between rice (*Oryza sativa*) and *A. longiglumis* (a genome 10.18 times larger than rice),  
656 the amplification of non-coding sequences lying between genes has occurred throughout the  
657 chromosome arms: syntenic regions show relatively uniformly expansion with a few  
658 substantial gaps. The repetitive sequence component of the ALO genome—the repeatome—is  
659 characterized by both ancient and more recently amplified transposable elements (TEs) and  
660 tandem repeats occurring both along chromosome arms and at centromeres. It remains

661 unclearly why genome size should be so different in two successful crop genera, *Avena* and  
662 *Oryza*, and whether selective pressures (dynamics of repeat replication and transposition)  
663 enhance options for evolvability. Given the high synteny observed, with presence of well-  
664 defined inter-chromosomal translocations and fusions between the species (including insertion  
665 of ancestral syntenic blocks within another), conserved nucleotide sequences and domains can  
666 be identified by major linkage blocks in grasses. We suggest that these can form the basis for  
667 synthetic pan chromosome oligonucleotide pools for *in situ* hybridization to identify major  
668 chromosomal and karyotypic rearrangements across the Poales.

669 The *Avena* genome assembly and analysis here, along with those of Triticeae and *Oryza*  
670 species in the BOP (Bambusoideae-Oryzoideae-Pooideae)-clade, provide insight into the  
671 extent and nature of chromosomal rearrangements and genome expansion in the pangenome,  
672 contributing to exploitation of the diversity present in the common gene pool across grasses  
673 through precision breeding using a range of approaches.

## 674 Materials and methods

675 Plant germplasm, genome sequencing and assembly

### 676 *Plant material*

677 The *Avena longiglumis* (ALO) (PI 657387; US Department of Agriculture at Beltsville,  
678 <https://www.ars-grin.gov/>, originally collected in Spain) was used for genome sequencing.  
679 After sowing, seedlings were grown in South China Botanical Garden Greenhouse at 25°C, 16  
680 h light/8 h dark with 70% relative humidity. Four weeks later, the plants were moved outside  
681 and further grown for 4 weeks under natural day-light condition (dry season in Guangzhou).

### 682 *Genome survey sequencing and assembly*

683 Genomic DNA for Illumina mate-pair sequencing was extracted using the DNeasy Plant Mini  
684 Kit (Qiagen) from 8-week-old leaves of ALO seedlings. An amplification-free approach was

685 used to prepare sequencing libraries with insert sizes of 350 bp, following the manufacture's  
686 protocol [98]. The paired-end reads were loaded into two lanes of an Illumina HiSeq2500  
687 platform and raw data generated reads with  $2 \times 150$  bp length (Table 1; Additional file 2: Tables  
688 S1 and S2). ALO genome size, heterozygosity and repeat content were determined by  $k$ -mer  
689 (17-mer) analysis by Jellyfish v.2.2.6 [99] with the parameter “-c -m 51 -s 10G -t 50”. The  
690 output file was used as the input for GenomeScope [100] to estimate the genome size. Project  
691 data have been deposited at Genome Sequence Archive  
692 (<https://ngdc.cncb.ac.cn/gsa/browse/CRA003996>; Additional file 2: Tables S1 and S2a).

### 693 *Oxford Nanopore Technology (ONT) sequencing and assembly*

694 For ONT PromethION library construction and sequencing, genomic DNA was extracted from  
695 3-week-old leaves of ALO seedling using the QIAGEN® Genomic DNA Extraction Kit (Cat.  
696 13323, Qiagen) according to the manufacturer protocol. DNA quantification was carried out  
697 using Qubit® 3.0 Fluorometer (Invitrogen, USA). DNA purification was confirmed (OD  
698 260/280, 1.8–2.0; OD 260/230, 2.0–2.2) and fragments in the range of 10–50 kbp recovered  
699 using a BluePippin automatic nucleic acid recovery instrument (Sage Science, USA). The 3'  
700 and 5' overhangs were converted into blunt ends with NEBNext® FFPE DNA Repair Mix  
701 (NEB, Cat. M6630) and then 'A' base was added to 3' blunt ends using the A-Tailing reaction  
702 (NEBNext® Ultra™ II End Repair/dA-Tailing Module, NEB, Cat. E7546). The purified A-  
703 tailed DNA was ligated with adaptors from the Ligation Sequencing Kit (SQK-LSK109,  
704 Oxford Nanopore Technologies) and the NEBNext® Quick Ligation Module (NEB, Cat.  
705 E6056). The purified ligation products were used as the constructed sequencing library. The  
706 DNA libraries were accurately quantified using a Qubit® 3.0 Fluorometer (Cat. E33216,  
707 Invitrogen, USA) and loaded into 12 lanes of a PromethION, R9.4.1 flow cell (Oxford  
708 Nanopore Technologies, UK) for SMRT (single molecular real-time) sequencing. Sequencing  
709 results (fast5 files) were processed using the Guppy v.3.2.2 [101] (Additional file 2: Table S2b).

710 A total of 31.2 million passed reads (Q score  $\geq 7$ ; 252.8 Gbp) were generated with read length  
711 N50 12,682,464 bp (Additional file 2: Table S3).

712 NextDenovo v.1.0 [102], wtdbg2.huge [103] and SMARTdenovo v.1.0.0 [104] have been  
713 used for self-correction of ONT reads. The pass reads were sent into NextDenovo v.1.0 for  
714 read correction. We tested parameters and found that using corrected reads to SMARTdenovo  
715 v.1.0.0 with the assembler parameters ‘-c 3’ and ‘-k 11’ gave good results, yielding a  
716 preliminary assembly consisting of 2,379 contigs (contig N50 11.92 Mbp). Contigs were  
717 polished three times with ONT raw data by NextPolish v.1.01 [102] and four times by the  
718 filtered Illumina whole-genome shotgun data by Fastp v.0.20.1 [105]. This procedure increased  
719 the contig N50 size to 12.68 Mbp (Additional file 2: Table S3).

720 *Hi-C library preparation and sequencing*

721 For Hi-C sequencing, 3-week-old leaves of ALO seedlings were fixed in 2% formaldehyde  
722 solution. The nuclei/chromatin was extracted from the fixed tissue and digested with DpnII  
723 (NEB, Cat. E0543L). Hi-C libraries were constructed and sequenced on the Illumina Novaseq  
724 6000 platform to obtain 150 bp paired-end reads (Additional file 2: Table S3). Raw data were  
725 processed by trimming adaptor and removing low-quality reads (Phred quality scores  $< 15$ ) by  
726 Fastp v.0.20.1 [105] with default parameters. A total of 1,453 million clean reads were kept for  
727 the mapping process. The quantity of informative Hi-C reads was estimated by Hi-C\_Pro  
728 v.2.10.0 [106].

729 The 585 million paired-end reads (40.79% of the clean reads) were uniquely mapped to  
730 the draft assembly sequence using Bowtie2 v.2.3.2 [107] (-end-to-end --very-sensitive -L 30).  
731 The de-duplicated list of alignments of Hi-C reads to the draft ALO assembly was generated  
732 using Juicer v.1.5.7 [108]. Nine base pair-delimited resolutions (2.5, 1 Mbp, 500, 250, 100, 50,  
733 25, 10, 5 kbp) were used to bin the reads and describe the interaction intensity of chromosome  
734 conformation. The 431 million (73.68% of unique mapped reads) valid paired-end reads were

735 used to assemble the draft assembly into chromosome-length scaffolds with the linking  
736 information by LACHESIS [109]. Only these scaffolds >15 kbp were taken into the processes  
737 of cluster, order and orientation. The iterative round for mis-correction was set as zero time.  
738 The pseudomolecules were generated by concatenating the adjacent contigs with 100 'N's  
739 [110]. Hi-C contact maps were processed by Pheatmap package for R v.3.6.3 [111] and  
740 reviewed in Juicer v.1.5.7 [108] (Additional file 1: Fig. S4, Additional file 2: Tables S4 and  
741 S5).

742 *Estimation of genome size*

743 Nuclear DNA content was estimated by flow cytometry [112]. The 20 mg of *A. brevis* (PI  
744 657352) leaves (2C = 8.98 pg [8] served as an internal reference standard, were chopped with  
745 blades in 500 µl Otto I buffer solution (0.1 M citric acid, 0.5% v/v Tween 20 [113]). The  
746 homogenate was filtered through a 40 µm nylon mesh (BD Falcon<sup>TM</sup>, Cat. 352340). Nuclei  
747 were pelleted by centrifugation and resuspended in 400 µl of Otto I buffer. After 30 min  
748 incubation at room temperature, 800 µl of Otto II solution (0.4 M Na<sub>2</sub>HPO<sub>4</sub>) supplemented  
749 with 50 µg/ml RNase and 50 µg/ml propidium iodide was added. Samples were analysed by a  
750 CyFlow Space flow cytometer (Sysmex Partec GmbH, Görlitz, Germany) equipped with 533  
751 nm laser. At least 5,000 nuclei were analysed per sample. Five plants were measured, and each  
752 plant was analysed three times on three different days. The 2C DNA content of ALO was  
753 calculated as  $9.23 \pm 0.20$  pg (mean  $\pm$  SD) by the ratio of G1 peak mean and standard value,  
754 then 1C genome size was calculated as  $4,513 \pm 0.099$  Mbp (1 pg = 978 Mbp [112]).

755 Total 268.60 Gbp clean data were used for *k*-mer analysis by Kmerfreq\_AR v.2.0.4 [114]  
756 (Luo et al., 2012) from SOAPec v.2.01 package (<http://soap.genomics.org.cn/about.html>) and  
757 Jellyfish v.2.2.6 [99] at 17-mer (Additional file 1: Fig. S2b). The genome size of *A. longiglumis*  
758 was estimated by the formula  $G = k\text{-mer number} / k\text{-mer depth}$ , where the *k*-mer number is the  
759 total numbers of *k*-mers, and *k*-mer depth refers to the depth of main peak. The genome size is

760 expected to be  $206,214,840,000/52 = 3.97$  Gbp, which was close to the flow cytometry result.  
761 The  $k$ -mer ( $k=17$ ) result indicated the heterozygosity of the ALO genome was approximately  
762 0.48%.

### 763 *Quality of genome assembly*

764 The Illumina paired-end data were mapped to assembled scaffolds with Bowtie2 v.2.3.2 [107]  
765 (Langmead and Salzberg, 2012). The overall alignment rate was 99.94% with 96.90% properly  
766 paired alignments. We identified 3,830,731 heterozygous SNPs and 177,108 indels (depth  $\geq$   
767  $10 \times$ ) in the ALO genome (Additional file 1: Table S4). The nanopore long reads were mapped  
768 to the assembled scaffolds using Minimap2 v.2.17 [115], and the depth of long reads was  
769 calculated using SAMtools [116] with default parameters.

770 The gene completeness of ALO assembly (Fig. 1) was evaluated by BUSCO v.4.0.5 [29]  
771 and CEGMA v.2.5 [30]. In BUSCO, a set of 1,375 plant-specific orthologous genes  
772 (Embryophyta\_odb10) was used to search against genome assembly with parameters ‘-  
773 lineage\_path embryophyta\_odb10 –mode geno’ (Additional file 1: Table S5a). In CEGMA, a  
774 collection of 241 most conserved eukaryotic genes was searched against genome assembly with  
775 default parameters (Additional file 1: Table S5b). The gene completeness was defined by the  
776 proportion of completely matched proteins out of 1,375 embryophyta genes or 241 conserved  
777 eukaryotic genes. Finally, the LTR Assembly Index (LAI = 10.54) was calculated using the  
778 LTR\_retriever [117].

### 779 *RNA preparation and sequencing*

780 Total RNA of four tissues [roots, salt-treated (the salt water of 4 mM NaCl for 48 h) roots,  
781 leaves and salt-treated leaves] were extracted using Column Plant RNAout 2.0 (Tiandz Inc.,  
782 Beijing, China) according to the manufacturer’s protocol. Extracted RNA was treated with  
783 DNase (Tiandz Inc., Beijing, China) to remove genomic DNA. The RNA quality was validated  
784 using agarose gel electrophoresis, Nanodrop 2000 (Nanodrop Technologies Inc., NanoDrop

785 2000, Wilmington, USA), and Agilent 2100 (Agilent Technologies Inc., Pleasanton, USA) to  
786 confirm the purity, concentration and integrity, respectively. Library construction and  
787 sequencing were performed using Illumina Novaseq 6000 platform (Illumina Inc., San Diego,  
788 USA).

789 The clean data was generated by removing adaptor sequences, ambiguous reads  
790 ('N' > 10%) and low-quality reads (greater than 50% of bases in reads with a quality value  
791 Q ≤ 20) using Fastp v.0.20.1 [105]. The quality control of clean reads was filtered by FastQC  
792 v.0.11.3 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc>) for further genome-  
793 wide expression dominance analysis (<https://ngdc.cncb.ac.cn/gsa/browse/> CRA004247;  
794 Additional file 2: Tables S1 and S2).

## 795 Genome sequence annotation

### 796 *Repeat analysis*

797 *De novo* repeat prediction of the ALO genome was carried out by EDTA v.1.7.0 (Extensive  
798 *de-novo* TE Annotator [36]) being composed of eight softwares (Fig. 2). LTRharvest [118],  
799 LTR\_FINDER\_parallel [36], LTR\_retriever [117] were incorporated to identify LTR  
800 retrotransposons; Generic Repeat Finder [119] and TIR-Learner [120] were included to identify  
801 TIR transposons; HelitronScanner v.1.0 [121] identified *Helitron* transposons; RepeatModeler  
802 v.2.0.2a [122] was used to identify TEs (such as *LINEs*); Finally, RepeatMasker v.4.1.1 [123]  
803 was used to annotate fragmented TEs based on homology to structurally annotated TEs. In  
804 addition, TEsorter v.1.1.4 [124] was used to identify TE-related genes (Additional file 2: Table.  
805 S6a).

806 For comparison, the same protocol was applied to analyse repeat of six grass genomes,  
807 including *Aegilops tauschii* (ATA [32] Luo et al., 2017), *Brachypodium distachyon* (BDI) [33],  
808 *Oryza sativa* (OSA) [9], *Sorghum bicolor* (SBI) [10], *Setaria italica* (SIT) [11] and *Zea mays*  
809 (ZMA) [34]. For LTR-RTs, the families were clustered based on their LTR sequences. The

810 final set of repetitive sequences in the ALO genome was obtained by integrating the *ab initio*-  
811 predicted TEs and those identified by homology through RepeatMasker (Additional file 2:  
812 Table. S6a). Intact LTR-RTs were identified and analyzed using LTR\_retriever [116]. A  
813 nucleotide substitution rate ( $r$ ) of  $1.3 \times 10^{-8}$  mutations per site per year [125] was used to  
814 estimate the insertion time ( $T$ ) of intact LTR-RTs with the formula of  $T = K/(2r)$  [126], where  
815  $K$  is the divergence rate of 5'-LTR and 3'-LTR estimated by the Jukes-Cantor model  
816 (Additional file 2: Table. S6b, c).

817 Locations of centromeres were identified by multiple genomic features: (1) High  
818 abundance repetitive areas of repeat sequences on chromosome dotplots (Fig. 2, Additional file  
819 2: Table S7a); (2) Discontinuities in the Hi-C contact map (Additional file 1: Fig. S4); (3)  
820 Location of barely (*Hordeum vulgare*) Gypsy LTR *Cereba* (KM948610 [41]) sequences are  
821 used to identify centromeres of wheat (*Triticum aestivum*, TAE; IWGSC, 2018), thus to  
822 identify the ALO centromeric regions, the *Cereba* sequence [42] was aligned to the ALO  
823 genome using Blastn to identify the centromere cores by Geneious Primer v.2021.1.1  
824 (<https://www.geneious.com/>; Additional file 2: Table. S7b); (4) SynVisio [42] visualization of  
825 gaps and conserved regions between ALO and OSA assemblies (Additional file 2: Table. S7c);  
826 (5) Regions of low gene density along each ALO chromosome. The centromere cores are  
827 identified by the overlap regions of the high abundance repetitive areas on ALO chromosome  
828 dotplots and the low gene density areas on ALO chromosomes.

829 *Gene prediction and functional annotation*

830 Gene structure prediction depended on the application of three methods, i.e., *ab initio*  
831 prediction, homology-based prediction and RNA-seq-assisted prediction [127]. Augustus  
832 v.3.3.2 [128] was used for *de novo*-based gene prediction with default parameters to predict  
833 genes of the ALO genome. Additionally, the filtered proteins in genomes of six species ATA  
834 [31], BDI [32], *Hordeum vulgare* [65], SBI [10], *Triticum aestivum* [65] and ZMA [34] were

835 used for homology-based prediction by GeMoMa v.1.6.1 [129] with default parameters  
836 (Additional file 2: Tables S8). Then, TransDecoder v.5.5.0 [130] were used for RNA-seq-based  
837 gene prediction. All predicted gene structures from three approaches were integrated into  
838 consensus gene models using EVidenceModeler v.1.1.1 [131]. These gene models were filtered  
839 sequentially to obtain a precise gene set, some genes whose sequences included transposable  
840 elements were removed with TransposonPSI v.2 (<http://transposonpsi.sourceforge.net>).

841 Gene functional annotation were carried out by performing BLASTP (E-value  $\leq$  1E-5)  
842 searches against NCBI non-redundant protein (NR) and Swiss-Prot (<http://www.uniprot.org/>)  
843 protein databases using BLASTP under the best match parameter [132]. NOG (Non-supervised  
844 Orthologous Groups), COG (Clusters of Orthologous Groups of proteins) [43], KEGG (Kyoto  
845 Encyclopedia of Genes and Genomes) [47], CAZy (Carbohydrate-Active enZYmes) [49],  
846 Pfam [44] annotations were performed with eggNOG v.5.0 [45]. The gene ontology (GO) IDs  
847 [46] for each gene were determined using the BLAST2GO v.1.44 [133]. Then transcription  
848 factors annotation was performed with PlantTFDB v.5.0 [48] (Additional file 2: Tables S9).

#### 849 *Identification of non-coding RNA genes*

850 Genome-wide prediction of non-coding RNA gene set (ncRNA) was performed (Additional  
851 file 2: Tables S10). Firstly, the data set was aligned to the Rfam library v.11.0 [134] noncoding  
852 database to annotate genes encoding ribosomal RNA (rRNA), small nuclei RNA (snRNA) and  
853 microRNA (miRNA). Then the transfer RNA (tRNA) sequences were identified using  
854 tRNAscan-SE v.2.0 [135]. Meanwhile, miRNAs were predicted by miRanda v.3.0 [136], while  
855 rRNA and its subunits were predicted by RNAmmer v.1.2 [137].

#### 856 *Identification of high- and low-confidence genes*

857 The 40,845 gene set was filtered to identify high-confidence (HC) protein-coding genes by two  
858 methods. Transcriptome raw reads were preprocessed by Fastp v.0.20.1 [105] with default  
859 parameters in order to trim adaptors and remove the low-quality RNA-seq reads (Phred quality

860 scores < 20). The clean reads were aligned to the ALO genome by STAR aligner [138]. The  
861 initial SAM-to-BAM conversion was performed by SAMtools [116]. The mapped RNA-seq  
862 reads (in BAM file) were assembled to transcript by Stringtie v.2.0.6 [139], which was used to  
863 call the fragments per kilobase of exon model per million mapped reads (FPKM) values.  
864 Subsequently, the genes with FPKM value larger than zero were classified as HC (Additional  
865 file 2: Table S11a). For the genes without transcriptome transcript abundance support, the  
866 alignment was performed with *A. atlantica* (identity > 95%, coverage > 95%), *A. eriantha*  
867 (identity > 90%, coverage > 90%), *Hordeum vulgare* and *Triticum aestivum* (identity > 80%,  
868 coverage > 80%) by BLASTP (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>; *E* value=1e-5),  
869 respectively. Those supported by alignment results of two or more species alignments were  
870 defined as HC genes (Additional file 2: Table S11b). Finally, the HC genes were supported by  
871 transcriptome data or homology, while the low-confidence (LC) genes were not supported by  
872 either one method (Additional file 2: Table S11c).

## 873 Evolutionary analysis

### 874 *Gene family identification and phylogenetic tree reconstruction*

875 To examine evolution and divergence of the ALO genome, protein-coding gene sequences  
876 from 10 species, *Avena atlantica* (AAT) [12], *A. eriantha* (AER) [12], *A. strigosa* (AST) [7],  
877 ATA [32], *Arabidopsis thaliana* (ATH) [50], BDI [33], OSA [9], SBI [10] (McCormic et al.,  
878 2018), SIT [11] (Yang et al., 2020) and ZMA [34], were downloaded from Phytozome v.13  
879 [140] and NCBI website (<https://www.ncbi.nlm.nih.gov/>) for comparative analyses (Additional  
880 file 2: Tables S12 and S13). When one gene had multiple transcripts, only the longest transcript  
881 in the coding region was kept for further analysis. Paralogs and orthologs were clustered with  
882 OrthoFinder v.2.3.14 [51] through standard mode parameters with Diamond v.0.9.24 [141].  
883 The numbers of shared and species-specific gene families among five species (ALO, AER,  
884 AST, OSA and ZMA) were visualized by UpSetR v.1.4.0 [142] for R v.3.6.3 [111].

885 Single-copy of orthologous genes were extracted from the OrthoFinder [51] clustering  
886 results and MAFFT v.7.48 [143] was used to align the concatenated protein sequences to give  
887 a super-gene matrix. RAxML v.8.1.17 [144] was used to reconstruct a phylogenetic tree with  
888 the GTR+G+I model and a bootstrap value of 1000 (Fig. 3). The phylogenetic tree was  
889 visualized by FigTree v.1.4.4 (<http://tree.bio.ed.ac.uk/software/figtree/>). Species divergence  
890 time and the 95% confidence intervals were inferred by MCMCTree in the PAML v.4.9i [145]  
891 via the Markov Chain Monte Carlo method. A correlated rate model (clock = 3) was established,  
892 and MCMC was performed (burnin = 2,000, sample number = 20,000, sample-frequency = 2).  
893 The crown ages for Stipeae (25.5–30.1 Mya), Oryzeae (28.4–33.4 Mya) and BEP/PACMAD  
894 split (45.0–57.4 Mya) [4], obtained from the TimeTree database (<http://www.timetree.org/>),  
895 were applied to calibrate the divergence times.

#### 896 *Expansion and contraction of gene families*

897 To analyse the expanded and contracted gene families, the Computational Analysis of Gene  
898 Family Evolution software (CAFE v.4.2) [146] was run to compute changes in gene families  
899 along each lineage of the phylogenetic tree under a random birth-and-death model (Additional  
900 file 2: Table S14). The expanded and contracted gene families were localized on the  
901 chromosomes (Fig. 1, circles f and g). Although these regions may be under selective pressure  
902 and reflect major duplications or deletions at the whole-genome level, the distribution of  
903 changed gene families largely reflects overall chromosomal gene density. The clustering results  
904 and the information from the estimated divergence times were used. Using conditional  
905 likelihood as the test statistics, the corresponding *P*-value of each lineage were calculated (*P*-  
906 value  $\leq 0.01$ ). Additionally, the GO enrichment of expanded genes and KEGG enrichment of  
907 unique genes were analysed to determine their functions (Additional file 2: Tables S15 and  
908 S16). Genomic landscape of repeats, genes and expanded genes were plotted across the ALO  
909 chromosomes by TBtools v.1.092 [147].

910 *Genome synteny and whole genome duplication analysis*

911 Protein sequences within and between genomes were searched against one another to detect  
912 putative homologous genes ( $E$  value  $< 1\text{e-}5$ ) by BLASTP. With information about homologous  
913 genes as input, MCscanX [148] were implemented to infer homologous blocks involving  
914 collinear genes within and between genomes. The maximum gap length between collinear  
915 genes along a chromosome region was set to 50 genes [149]. Then, homology dotplots were  
916 constructed by a perl script to reveal genomic correspondence. Then, homology dotplots were  
917 constructed by SynVisio [42] to reveal genomic correspondence within ALO, between three  
918 *Avena* species, between ancestral grass karyotype (AGK) and seven grass species, and between  
919 ALO, OSA and BDI (Fig. 4, Additional file 2: Table S17). Subsequently, we applied the WGDI  
920 v.0.4.7 [150] to identify the whole genome duplication events based on the high synonymous  
921 (Ks) peak of ALO versus AST, ALO versus AER, AST versus AER collinear gene pairs. Non-  
922 synonymous (Ka) and Ks substitution rates for gene pairs were calculated with  
923 KaKs\_Calculator v.2.0 [151] under the YN model. The synonymous substitutions rate per site  
924 per year ( $r$ ) equaling  $1.3 \times 10^{-8}$  was applied to the recent WGD estimation [152].

925 **Anti-salinity and secondary metabolism gene cluster analysis**

926 *Genome-wide expression analysis*

927 To investigate the expression dominance of salt-responsive gene, the FPKM values were  
928 calculated for genes in roots, leaves, salt-treated roots and salt-treated leaves. Differentially  
929 expressed genes (DEGs) were identified by DEseq2 v.3.11 [153]. We filtered the DEGs with a  
930 minimum of two-fold differential expression ( $|\text{fold change}| \geq 2$ ; false discovery rate  
931 (FDR)  $\leq 0.005$ ). The DEGs were performed KEGG enrichment by TBtools [147] (Additional  
932 file 2: Table S18).

933 *Metabolic gene cluster prediction and CYP450 gene identification*

934 The plantiSMASH v.1.3 [154] was used to identify the potential metabolic gene clusters in the  
935 ALO genome with parameter setting “run\_antismash.py -c 16 --taxon plants tanxaing.gb --  
936 outputfolder tanxiang” (Fig. 5, Additional file 2: Table S19). To identify the *CYP450* gene  
937 numbers in genomes of ALO, nine grass species (AAT, AER, AST, ATA, BDI, OSA, SBI, SIT  
938 and ZMA) and *Arabidopsis thaliana*, all proteins of each species was searched against hidden  
939 Markov model (HMM) profile of the Pfam domain (PF00067) by hmmsearch  
940 (<http://hmmer.org/>). Putative *CYP450* genes were further verified in the Pfam database  
941 (PF00067) to confirm the *CYP450* proteins of ALO. The *CYP450* gene copy number and  
942 syntenic relationships between ALO and the nine grass species and *Arabidopsis thaliana* were  
943 visualized by TBtools v.1.092[147] (Additional file 2: Tables S20–S23).

944 *CesA, Csl and CalS gene identification*

945 To identify the *CesA*, *Csl* and *CalS* (or *GSL*) gene family members in ALO, all proteins of  
946 ALO was searched against hidden Markov model (HMM) profile of the Pfam domain  
947 [(PF00535 or PF03552 for *CesA* and *Csl*) and (PF02364 and PF14288 for *CalS*)] by hmmsearch  
948 (<http://hmmer.org/>). Putative *CesA*, *Csl* and *CalS* genes were further verified in the Pfam  
949 database [155] (<http://pfam.xfam.org/>), screened for Pfam domains [(PF00535 or PF03552 for  
950 *CesA* and *Csl*) and (PF02364 and PF14288 for *CalS*)] to confirm as the *CesA*, *Csl* and *CalS*  
951 proteins of ALO (Additional file 2: Table S24). The *CYP450* gene copy number and syntenic  
952 relationships between ALO and nine grass species [AAT, AER, AST, ATA, BDI, OSA, SBI,  
953 SIT and ZMA] and *Arabidopsis thaliana* were visualized by TBtools v.1.092 (Additional file  
954 2: Tables S24–S30). Previously known *CesA* and *Csl* protein sequences were downloaded for  
955 *Arabidopsis thaliana*, rice, wheat and maize (Kaur et al. [156] their Fig. S1). *GSL* protein  
956 sequences of *Arabidopsis thaliana* and rice were downloaded from RGAP  
957 <http://rice.uga.edu/index.shtml>) and TAIR (<https://www.arabidopsis.org/>). Multiple sequence

958 alignments of CesA and Csl proteins and GSL proteins were performed by MAFFT v.7.48 [142]  
959 with default parameters (Additional file 1: Fig. S34, Additional file 2: Tables S31 and S32). A  
960 maximum likelihood (ML) phylogenetic tree was constructed using FastTree v.2.1.10 with GTR  
961 model [157] and 1000 bootstrap replicates. The phylogenetic tree was visualized by FigTree  
962 v.1.4.4 (Fig. 5b).

## 963 **Supplementary information**

964 **Supplementary information** accompanies this paper at  
965 <https://doi.org/10.xxxx/syyyyy-yyy-yyyy-y>.

966 **Additional file 1: Fig. S1.** Flow cytometric estimation of the nuclear genome size of *Avena*  
967 *longiglumis*. Nuclei were isolated from *A. longiglumis* (PI 657387) and *A. brevis* (CN 1979;  
968 used as an internal reference standard), stained and analyzed simultaneously. **Fig. S2.** *Avena*  
969 *longiglumis* spikelets and genome assembly. **Fig. S3.** Strategy for sequencing, assembly and  
970 annotation of the *Avena longiglumis* genome. **Fig. S4.** Inter-chromosomal contact matrix. The  
971 intensity of pixels represents the normalized count of Hi-C links between 100 kbp windows on  
972 ALO chromosomes on a logarithmic scale. **Fig. S5.** Evaluation of genome assemblies by LTR  
973 Assembly Index (LAI). **Fig. S6.** Gene density of 1 Mbp-sized sliding windows on seven  
974 chromosomes of *Avena longiglumis*. **Fig. S7.** Centromeric retrotransposon *Cereba*  
975 (KM948610 [41]) sequence locations on seven chromosomes of *Avena longiglumis*.  
976 Centromere area denoted by red dots. **S8.** KEGG enrichment of expanded genes in four *Avena*  
977 species. **Fig. S9.** KEGG enrichment of unique genes of four *Avena* species. **Fig. S10.**  
978 Syntenic relationships based on three *Avena* species genomes. **Fig. S11.** Syntenic relationships  
979 based on AST-ALO-AER and OSA-ALO-BDI genome homologous genes. **Fig. S12.** KEGG  
980 enrichment of up- and down-regulated DEFs in roots versus salt-treated roots and leaves versus  
981 salt-treated leaves of ALO. **Fig. S13.** Statistics of expanded DEGs (number  $\geq 5$ ) of ALO. **Fig.**  
982 **S14.** Heat map showing hierarchical clustering of *cytochrome P450* gene families in roots,

983 salt-treated roots, leaves and salt-treated leaves of ALO. **Fig. S15.** Total 557 *cytochrome P450*  
984 genes located within 57 clusters identified in the ALO genome. **Fig. S16.** *Cytochrome P450*  
985 genes inserted within gene-clusters identified on ALO chromosomes. **Fig. S16a.** *Cytochrome*  
986 *P450* genes inserted within gene-clusters identified on ALO01. **Fig. S16b.** *Cytochrome P450*  
987 genes inserted within gene-clusters identified on ALO02. **Fig. S16c.** *Cytochrome P450* genes  
988 inserted within gene-clusters identified on ALO03. **Fig. S16d.** *Cytochrome P450* genes  
989 inserted within gene-clusters identified on ALO04. **Fig. S16e.** *Cytochrome P450* genes  
990 inserted within gene-clusters identified on ALO05. **Fig. S16f.** *Cytochrome P450* genes  
991 inserted within gene-clusters identified on ALO06. **Fig. S16g.** *Cytochrome P450* genes  
992 inserted within gene-clusters identified on ALO07. **Fig. S17.** Total 11 *CesA* and 55 *Csl* genes  
993 located within 10 clusters identified in the ALO genome. **Fig. S18.** Ten *CesA* and *Csl* gene  
994 clusters on six chromosomes of ALO. **Fig. S18a.** *CesA* and *Csl* genes inserted gene-clusters  
995 identified on ALO02. **Fig. S18b.** *Csl* genes inserted within gene-cluster identified on ALO03.  
996 **Fig. S18c.** *Csl* genes inserted within gene-cluster identified on ALO04. **Fig. S18d.** *Csl* genes  
997 inserted within gene-clusters identified on ALO05. **Fig. S18e.** *Csl* genes inserted within gene-  
998 cluster identified on ALO06. **Fig. S18f.** *Csl* genes inserted within gene-cluster identified on  
999 ALO07. **Fig. S19.** FASTA sequences of *Csl* proteins of ALO used for the phylogenetic  
1000 analysis. **Fig. S20.** Heat map showing hierarchical clustering of the ALO *callose* gene (*CalS*,  
1001 *GSL*) families in roots, salt-treated roots, leaves and salt-treated leaves. **Fig. S21.** The  
1002 maximum likelihood phylogenetic tree constructed with *CalS* (*GSL*) proteins of ALO,  
1003 *Arabidopsis thaliana* and *Oryza sativa*. **Fig. S22.** FASTA sequences of *CalS* (*GSL*) proteins  
1004 of ALO used for the phylogenetic analysis.

1005 **Additional file 2: Table S1.** Summary of sequencing libraries of *Avena* species included in  
1006 the study. **Table S2.** Deposited data of *Avena longiglumis* (ALO) genome and software used  
1007 in the study. **Table S3.** Summary of genome assembly and annotation of ALO. **Table S4.**

1008 Statistics of the ALO genome assembly consistency. **Table S5.** Evaluation of gene space  
1009 completeness in the ALO genome assembly. **Table S6.** Repetitive DNA composition  
1010 comparison among genomes of ALO and six grass species. **Table S7.** Size and centromere  
1011 localization of the ALO pseudomolecules. **Table S8.** Gene characterization comparison  
1012 among ALO and ten other plant species. **Table S9.** Statistics of gene function annotation of  
1013 the ALO genome. **Table S10.** Statistics of annotated non-coding RNAs of the ALO genome.  
1014 **Table S11.** Identification of high-confidence (HC) and low-confidence (LC) protein-coding  
1015 genes annotated in the ALO genome. **Table S12.** Gene family categories in genomes of ALO  
1016 and ten plant species. **Table S13.** Gene family statistics of ALO and ten plant species. **Table**  
1017 **S14.** Summary of orthologous gene clusters analyzed in analysed species. **Table S15.** GO  
1018 enrichment analysis of expanded genes in the ALO genome. **Table S16.** KEGG enrichment  
1019 analysis of unique genes in the ALO genome. **Table S17.** The gene pair statistics of SynVisio  
1020 results between post- $\rho$  ancestral grass karyotype (AGK) and grass species and between grass  
1021 species. **Table S18.** Statistics of up- and down-regulated DEGs in salt-treated roots and salt-  
1022 treated leaves and expanded gene families of ALO. **Table S19.** Characterization of  
1023 biosynthetic gene clusters (BGCs) in the ALO genome. **Table S20.** The *CYP450* gene copy  
1024 number in analysed species and distribution along the ALO chromosomes. **Table S21.** FPKM  
1025 value of *CYP450* genes in roots, leaves, salt-treated roots and salt-treated leaves of ALO.  
1026 **Table S22.** Description of 46 BGCs containing 117 *CYP450* genes in the expression profiling  
1027 heat map. **Table S23.** The conserved synteny among *CYP450* genes within gene clusters in  
1028 ALO and nine grass species. **Table S24.** Statistics of 11 *CesA* and 55 *Csl* genes corresponding  
1029 to the expanded gene families in the ALO genome. **Table S25.** Distribution of 11 *CesA* and  
1030 55 *Csl* genes in the ALO chromosomes. **Table S26.** FPKM value of *CesA* and *Csl* genes in  
1031 roots, salt-treated roots, leaves and salt-treated leaves of ALO. **Table S27.** Gene description  
1032 of 10 BGCs containing *CesA* and *Csl* genes in the expression profiling heat map. **Table S28.**

1033 The conserved synteny among *CesA* and *Csl* genes within gene clusters in ALO and nine grass  
1034 species. **Table S29.** Statistics of 13 *CalS* (*GSL*) genes corresponding the expanded gene  
1035 families in the ALO genome. **Table S30.** FPKM value of *CalS* genes in roots, salt-treated  
1036 roots, leaves and salt-treated leaves of ALO.

1037 **Additional file 3:** Review history.

## 1038 **Declarations**

### 1039 **Acknowledgements**

1040 We thank Jun Wen, Dallas Kessler and Bockelman Harold for the seed collection, Shuyu Zhou,  
1041 Haoyan Pan and Dongli Cui for the assistance with growing and maintaining the plants, and  
1042 Yubo Wang, Kai Ouyang, and Huitong Tan for the statistical advice. We thank Grandomics  
1043 Biosciences Co., Ltd. (Wuhan, China) for sequencing support, Huawei Elastic Cloud Server  
1044 (Jiangsu, China) for supplying the computational resources.

### 1045 **Review history**

1046 The review history is available as Additional file 3.

### 1047 **Peer review information**

1048 XXX was the primary editor of this article and managed its editorial process and peer review  
1049 in collaboration with the rest of the editorial team.

### 1050 **Authors' contributions**

1051 QL and JSHH designed the study. ZWW, YSY and XKT collected samples. ZYS and XKT  
1052 sequenced DNA. MZL, ZWW and DLC performed genome assembly, polishing, validation,  
1053 annotation and analysis. HYY, MZL, ZWW and ZYS performed repeat and transcriptome  
1054 sequence analysis. TS and JSHH supervised genome assembly and analysis. QL, HYY and  
1055 JSHH wrote the manuscript. TS and JSHH revised the manuscript. All authors read and  
1056 approved the final manuscript.

1057 **Funding**

1058 This work was supported by grants from National Science Foundation of China (32070359),  
1059 Guangdong Basic and Applied Basic Research Foundation (2021A1515012410), Overseas  
1060 Distinguished Scholar Project of SCBG (Y861041001) and Undergraduate Innovation  
1061 Training Program of Chinese Academy of Sciences (KCJH-80107-2020-004-97).

1062 **Availability of data and materials**

1063 The sequencing data used in this study have been deposited into the Genome Sequence Archive  
1064 (GSA) database in BIG Data center under Accession Number PRJCA004488/CRR275304-  
1065 CRR275326 and CRR285670-285674 (<https://ngdc.cncb.ac.cn/gsa/browse/CRA003996> for  
1066 raw data of the ALO genome; <https://ngdc.cncb.ac.cn/gsa/browse/CRA004247> for raw data of  
1067 ALO transcriptome). The previously reported Illumina data for were deposited into the NCBI  
1068 database under Accession Number SRA: SRR6058489-SRR6058492 and from NCBI under  
1069 BioProject PRJNA407595 (<https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA407595>).

1070 The ALO genome assembly was uploaded to <https://figshare.com/s/34d0c099e42eb39a05e2>.

1071

1073 **Ethics approval and consent to participate**

1074 Not applicable.

1075 **Consent for publication**

1076 Not applicable.

1077 **Competing interests**

1078 The authors declare that they have no competing interests.

1079 **Author details**

1080 <sup>1</sup> Key Laboratory of Plant Resources Conservation and Sustainable Utilization / Guangdong  
1081 Provincial Key Laboratory of Applied Botany, South China Botanical Garden, Chinese

1082 Academy of Sciences, Guangzhou, 510650, China. <sup>2</sup> Center for Conservation Biology, Core  
1083 Botanical Gardens, Chinese Academy of Sciences, Guangzhou, 510650, China. <sup>3</sup> University of  
1084 Chinese Academy of Sciences, Beijing, 100049, China. <sup>4</sup> Bio&Data Biotechnologies Co. Ltd.,  
1085 Guangzhou, 510700, China. <sup>5</sup> Grandomics Biosciences, Beijing 102200, China. <sup>6</sup> Department  
1086 of Genetics and Genome Biology, University of Leicester, Leicester LE1 7RH, UK.

1087 **References**

- 1088 1. Sancho R, Cantalapiedra CP, López-Alvarez D, Gordon SP, Vogel JP, Catalán P, et al.  
1089 Comparative plastome genomics and phylogenomics of *Brachypodium*: flowering time  
1090 signatures, introgression and recombination in recently diverged ecotypes. *New Phytol.*  
1091 2018; 218(4):1631–44.
- 1092 2. Liu Q, Peterson PM, Ge XJ. Phylogenetic signals in the realized climate niches of Chinese  
1093 grasses (Poaceae). *Plant Ecol.* 2011;212:1733–46.
- 1094 3. Christin PA, Spriggs E, Osborne CP, Strömberg CAE, Salamin N, Edwards EJ. Molecular  
1095 dating, evolutionray rates, and the age of the grasses. *Syst Biol.* 2014;63(2):153–65.
- 1096 4. Soreng RJ, Peterson PM, Romaschenko K, Davidse G, Teisher J, Clark LG, et al. A  
1097 worldwide phylogenetic classification of the Poaceae (Gramineae) II: an update and a  
1098 comparison of two 2015 classification. *J Syst Evol.* 2017;55(4):259–90.
- 1099 5. Bianconi ME, Hackel J, Vorontsova MS, Alberti A, Arthan W, Burke SV, et al. Continued  
1100 adaptation of C<sub>4</sub> photosynthesis after an initial burst of changes in the Andropogoneae  
1101 grasses. *Syst Biol.* 2020;69(3):445–61.
- 1102 6. Grundy MLM, Fardet A, Tosh SM, Rich GT, Wilde PJ. Processing of oat: the impact on  
1103 oat's cholesterol lowering effect. *Food Funct.* 2018;9(3):1328–43.
- 1104 7. Li Y, Leveau A, Zhao QA, Feng Q, Lu HY, Miao JS, et al. Subtelomeric assembly of a  
1105 multi-gene pathway for antimicrobial defense compounds in cereals. *Nat Commun.*  
1106 2021;12(1):2563.

1107 8. Yan HH, Martin SL, Bekele WA, Latta RG, Diederichsen A, Peng YY, et al. Genome  
1108 size variation in the genus *Avena*. *Genome*. 2016;59(3):209–20.

1109 9. Ouyang S, Zhu W, Hamilton J, Lin HN, Campbell M, Childs K, et al. The TIGR rice  
1110 genome annotation resource improvement and new features. *Nucleic Acids Res.*  
1111 2007;35(Database issue):D883–7.

1112 10. McCormic RF, Truong SK, Sreedasyam A, Jenkins J, Shu S, Sims D, et al. The *Sorghum*  
1113 *bicolor* reference genome: improved assembly, gene annotations, a transcriptome atlas,  
1114 and signatures of genome organization. *Plant J.* 2018;93(2):338–54.

1115 11. Yang ZR, Zhang HS, Li XK, Shen HM, Gao JH, Hou SY, et al. A mini foxtail millet with  
1116 an *Arabidopsis*-like life cycle as a C<sub>4</sub> model system. *Nat Plants*. 2020;6(9):1167–78.

1117 12. Maughan PJ, Lee R, Walstead R, Vickerstaff RJ, Fogarty MC, Brouwer CR, et al.  
1118 Genomic insights from the first chromosome-scale assemblies of oat (*Avena* spp.) diploid  
1119 species. *BMC Biol.* 2019;17:92.

1120 13. Liu Q, Lin L, Zhou XY, Peterson PM, Wen J. Unraveling the evolutionary dynamics of  
1121 ancient and recent polyploidization events in *Avena* (Poaceae). *Sci Rep.* 2017;7:41944.

1122 14. Murat F, Xu JH, Tannier E, Abrouk M, Guilhot N, Pont C, et al. Ancestral grass karyotype  
1123 reconstruction unravels new mechanisms of genome shuffling as a source of plant  
1124 evolution. *Genome Res.* 2010;20(11):1545–57.

1125 15. Murat F, Armero A, Pont C, Klopp C, Salse J. Reconstructing the genome of the most  
1126 recent common ancestor of flowering plants. *Nat Genet.* 2017;49(4):490–6.

1127 16. Welch RW, Brown JCW, Leggett JM. Interspecific and intraspecific variation in grain  
1128 and great characteristics of wild oat (*Avena*) species: very high great (1→3),(1→4)-beta-  
1129 D-glucan in an *Avena atlantica* genotype. *J Cereal Sci.* 2000;31(3):273–9.

1130 17. Amosova A, Zoshchuk SA, Rodionov AV, Ghukasyan L, Samatadze TE, Punina EO, et  
1131 al. Molecular cytogenetics of valuable Arctic and sub-Arctic pasture grass species from  
1132 the *Aveneae/Poeae* tribe complex (Poaceae). *BMC Genet.* 2019;20(1):92.

1133 18. Saini Pa, Gani M, Saini Po, Bhat JA, Francies RM, Negi N, et al. Molecular breeding for  
1134 resistance to economically important diseases of fodder oat. In: Wani SH, editor. *Disease*  
1135 *resistance in crop plants*. Switzerland AG: Springer Nature; 2019. p. 199–239.

1136 19. Li R, Gong M, Zhang XM, Wang F, Liu ZY, Zhang L, et al. The first sheep graph pan-  
1137 genome reveals the spectrum of structural variations and their effects on different tail  
1138 phenotypes. *bioRxiv*. 2021;472709.

1139 20. Picart-Picolo A, Grob S, Picault N, Franek M, Llauro C, Halter T, et al. Large tandem  
1140 duplications affect gene expression, 3D organization, and plant-pathogen response.  
1141 *Genome Res.* 2020;30(11):1583–92.

1142 21. Della Coletta R, Qiu YJ, Ou SJ, Hufford MB, Hirsch CN. How the pan-genome is  
1143 changing crop genomics and improvement. *Genome Biol.* 2021;22(1):3

1144 22. Liu Q, Li XY, Zhou XY, Li MZ, Zhang FJ, Schwarzacher T, Heslop-Harrison JS. The  
1145 repetitive DNA landscape in *Avena* (Poaceae): chromosome and genome evolution  
1146 defined by major repeat classes in whole-genome sequence reads. *BMC Plant Biol.*  
1147 2019;19:226.

1148 23. Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, et al. Tandem  
1149 repeats lead to sequence assembly errors and impose multi-level challenges for genome  
1150 and protein databases. *Nuclei Acids Res.* 2019;47(21):10994–1006.

1151 24. Amarasinghe S, Su S, Dong XY, Zappia L, Ritchie ME, Gouil Q. Opportunities and  
1152 challenges in long-read sequencing data analysis. *Genome Biol.* 2020;21(1):30.

1153 25. Perumal S, Koh CS, Jin LL, Buchwaldt M, Higgins EE, Zheng CF, et al. A high-  
1154 contiguity *Brassica nigra* genome localizes active centromeres and defines the ancestral  
1155 *Brassica* genome. *Nature Plant*. 2020;6(8): 929–41.

1156 26. Wang ZW, Rouard M, Biswas M, Droc G, Cui DL, Roux N, et al. A chromosome-level  
1157 reference genome of *Ensete glaucum* gives insight into diversity, chromosomal and  
1158 repetitive sequence evolution in the Musaceae. *bioRxiv*. 2021;469474.

1159 27. Appels R, Eversole K, Stein N, Feuillet C, Keller B, Rogers J, et al. Shifting the limits in  
1160 wheat research and breeding using a fully annotated reference genome. *Science*. 2018;  
1161 361(6403):eaar7191.

1162 28. Belser C, Istance B, Denis E, Dubarry M, Baurens FC, Falentin C, et al. Chromosome-  
1163 scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat Plants*  
1164 2018;4(11):879–87.

1165 29. Simão FA, Waterhouse RM, Ioannidis P, Krivensseva EV, Zdobnov EM. BUSCO:  
1166 assessing genome assembly and annotation completeness with single-copy orthologs.  
1167 *Bioinformatics*. 2015;31(19):3210–2.

1168 30. Parra G, Bradnam K, Ning Z, et al. Accessing the gene space in draft genomes. *Nucleic  
1169 Acids Research*. 2009;37(1):289–97.

1170 31. Ou SJ, Chen J, Jiang N. Assessing genome assembly quality using the LTR Assembly  
1171 Index (LAI). *Nucleic Acids Research*. 2018;46(21):e126.

1172 32. Luo MC, Gu YQ, Puiu D, Wang H, Twardziok SO, Deal KR, et al. Genome sequence of  
1173 the progenitor of wheat D subgenome *Aegilops tauschii*. *Nature*. 2017;551(7681):498–  
1174 502.

1175 33. International *Brachypodium* Initiative. Genome sequencing and analysis of the model  
1176 grass *Brachypodium distachyon*. *Nature*. 2010;463(7282):763–8.

1177 34. Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, et al. The B73 maize  
1178 genome: complexity, diversity, and dynamics. *Science*. 2009;326(5956):1112–5.

1179 35. Zhang GB, Ge CX, Xu P, Wang SK, Cheng SN, Han YB, et al. The reference genome of  
1180 *Miscanthus floridulus* illuminates the evolution of Saccharinae. *Nat Plants*.  
1181 2021;7(5):608–18.

1182 36. Ou SJ, Jiang N. *LTR\_FINDER\_parallel*: parallelization of *LTR\_FINDER* enabling rapid  
1183 identification of long terminal repeat retrotransposons. *Mob DNA*. 2019; 10:48.

1184 37. Dodsworth S, Leitch AR, Leitch I. Genome size diversity in angiosperms and its influence  
1185 on gene space. *Curr Opin Genet Dev*. 2015;35:73–8.

1186 38. Fu YB. Oat evolution revealed in the maternal lineages of 25 *Avena* species. *Sci Rep*.  
1187 2018;8(1):4252.

1188 39. Zhou X, Jellen EN, Murphy JP. Progenitor germplasm of domesticated hexaploid oat.  
1189 *Crop Sci*. 1999;39(4):1208–14.

1190 40. Estep MC, DeBarry JD, Bennetzen JL. The dynamics of LTR retrotransposon  
1191 accumulation across 25 million years of panicoid grass evolution. *Heredity (Edinb)*.  
1192 2013;110(2):194–204.

1193 41. Tomás D, Rodrigues J, Varela A, Veloso MM, Viegas W, Silva M. Use of repetitive  
1194 sequences for molecular and cytogenetic characterization of *Avena* species from Portugal.  
1195 *Int J Mol Sci*. 2016;17(2):203.

1196 42. Bandi V, Gutwin C. Interactive exploration of genomic conservation. In: Proceedings of  
1197 the 46th graphics interface conference on proceedings of graphics interface 2020 (GI’20).  
1198 Waterloo, Canada: Canadian Human-Computer Communications Society; 2020.

1199 43. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The  
1200 COG database: an updated version includes eukaryotes. *BMC Bioinformatics*. 2003;4:41.

1201 44. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein  
1202 families database in 2019. *Nucleic Acids Res.* 2019;47(Database issue):D427–32.

1203 45. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund AK, Cook H, et  
1204 al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology  
1205 resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*  
1206 2019;47(Database issue):D309–14.

1207 46. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, et al. The Gene  
1208 Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 2004;32(Database  
1209 issue):D258–61.

1210 47. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives  
1211 on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2017;45(Database  
1212 issue):D353–61.

1213 48. Tian F, Yang DC, Meng YQ, Jin J, Gao G. PlantRegMap: charting functional regulatory  
1214 maps in plants. *Nucleic Acids Res.* 2020;48(Database issue):D1104–13.

1215 49. Levasseur A, Drula E, Lombard V, Coutinho PM, Henrissat B. Expansion of the  
1216 enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes.  
1217 *Biotechnol Biofuels.* 2013;6(1):41.

1218 50. The *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering  
1219 plant *Arabidopsis thaliana*. *Nature.* 2000;408(6814):796–815.

1220 51. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative  
1221 genomics. *Genome Biol.* 2019;20(1):238.

1222 52. Vanneste K, Baele G, Maere S, van de Peer Y. Analysis of 41 plant genomes supports a  
1223 wave of successful genome duplications in association with the Cretaceous–Paleogene  
1224 boundary. *Genome Res.* 2014;24:1334–7.

1225 53. D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, et al. The banana  
1226 (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature*.  
1227 2012;488(7410):213–7.

1228 54. Pandian BA, Sathishraj R, Djananguiraman M, Prasad PVV, Jugulam M. Role of  
1229 cytochrome P450 enzymes in plant stress response. *Antioxidants (Basel)*. 2020;9(5):454.

1230 55. Kemen AC, Honkanen S, Melton RE, Findlay KC, Mugford ST, Hayashi K, et al.  
1231 Investigation of triterpene synthesis and regulation in oats reveals a role for  $\beta$ -amyrin in  
1232 determining root epidermal cell patterning. *Proc Natl Acad Sci U S A*.  
1233 2014;111(23):8679–84.

1234 56. Varshney RK, Bohra A, Yu JM, Graner A, Zhang QF, Sorrells ME. Designing future crops:  
1235 genomics-assisted breeding comes of age. *Trends Plant Sci*. 2021;26(6):631–49.

1236 57. Polturak G, Liu ZH, Osbourn A. New and emerging concepts in the evolution and  
1237 function of plant biosynthetic gene clusters. *Curr Opin Green Sustain Chem*.  
1238 2022;33:100568.

1239 58. McFarlane HE, Döring A, Persson S. The cell biology of cellulose synthesis. *Annu Rev*  
1240 *Plant Biol*. 2014;65:69–94.

1241 59. Richmond TA, Somerville CR. The cellulose synthase superfamily. *Plant Physiol*.  
1242 2000;124(2):495–8.

1243 60. Yang J, Bak G, Burgin T, Barnes WJ, Mayes HB, Peña MJ, et al. Biochemical and genetic  
1244 analysis identify CSLD3 as a beta-1,4-glucan synthase that functions during plant cell  
1245 wall synthesis. *Plant Cell*. 2020;32(5):1749–67.

1246 61. Zeng L, Tu XL, Dai H, Han FM, Lu BS, Wang MS, et al. Whole genomes and  
1247 transcriptomes reveal adaptation and domestication of pistachio. *Genome Biol*.  
1248 2019;20:79.

1249 62. Vatén A, Dettmer J, Wu S, Stierhof YD, Miyashima S, Yadav SR, et al. Callose  
1250 biosynthesis regulates symplastic trafficking during root development. *Dev Cell*.  
1251 2011;21(6):1144–55.

1252 63. Fleischmann A, Michael TP, Rivadavia F, Sousa A, Wang WQ, Temsch EM, et al.  
1253 Evolution of genome size and chromosome number in the carnivorous plant genus  
1254 *Genlisea* (Lentibulariaceae), with a new estimate of the minimum genome size in  
1255 angiosperms. *Ann Bot*. 2014;114(8):1651–63.

1256 64. Neale DB, Zimin AV, Zaman S, Scott AD, Shrestha B, Workman RE, et al. Assembled  
1257 and annotated 26.5 Gbp coast redwood genome: a resource for estimating evolutionary  
1258 adaptive potential and investigating hexaploid origin. *G3 (Bethesda)*. 2022;12(1):jkab380.

1259 65. Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, et al. A  
1260 chromosome conformation capture ordered sequence of the barley genome. *Nature*.  
1261 2017;544(7651):427–33.

1262 66. International Wheat Genome Sequencing Consortium (IWGSC). A chromosomebased  
1263 draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science*.  
1264 2014;345(6194):1251788.

1265 67. International Wheat Genome Sequencing Consortium (IWGSC). Shifting the limits in  
1266 wheat research and breeding using a fully annotated reference genome. *Science*.  
1267 2018;361(6403): eaar7191.

1268 68. Heslop-Harrison JS, Brandes A, Taketa S, Schmidt T, Vershinin AV, Alkhimova EG, et  
1269 al. The chromosomal distributions of *Tyl-copia* group retrotransposable elements in  
1270 higher plants and their implications for genome evolution. *Genetica*. 1997;100(1-3):197–  
1271 204.

1272 69. Nishihara H. Transposable elements as genetic accelerators of evolution: contribution to  
1273 genome size, gene regulatory network rewiring and morphological innovation. *Genes*  
1274 *Genet Syst.* 2019;94(6):269–281.

1275 70. Santos FC, Guyot R, do Valle CB, Chiari L, Techio VH, Heslop-Harrison P, et al.  
1276 Chromosomal distribution and evolution of abundant retrotransposons in plants: *gypsy*  
1277 elements in diploid and polyploid *Brachiaria* forage grasses. *Chromosome Res.*  
1278 2015;23(3):571–82.

1279 71. Aragón-Alcaide L, Miller T, Schwarzacher T, Reader S, Moor G. A cereal centromeric  
1280 sequence. *Chromosoma.* 1996;105(5):261–8.

1281 72. Presting GG, Malysheva L, Fuchs J, Schubert I. A *TY3/GYPSY* retrotransposon-like  
1282 sequence localizes to the centromeric regions of cereal chromosomes. *Plant J.*  
1283 1998;16(6):721–8.

1284 73. Vershinin AV, Druka A, Alkhimova AG, Kleinhofs A, Heslop-Harrison JS. *LINEs* and  
1285 *gypsy*-like retrotransposons in *Hordeum* species. *Plant Mol Biol.* 2002;49(1):114.

1286 74. Maluszynska J, Heslop-Harrison JS. Localization of tandemly repeated DNA sequences  
1287 in *Arabidopsis thaliana*. *Plant J.* 1991;1(2):159–66.

1288 75. Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmücker A, et al. The  
1289 genetic and epigenetic landscape of the *Arabidopsis* centromeres. *Science.*  
1290 2021;374(6569):eabi7489.

1291 76. Menzel G, Dechyeva D, Wenke T, Holtgräwe D, Weisshaar B, Schmidt T. Diversity of a  
1292 complex centromeric satellite and molecular characterization of dispersed sequence  
1293 families in sugar beet (*Beta vulgaris*). *Ann Bot.* 2008;102(4):521–30.

1294 77. Belser C, Baurens FC, Noel B, Martin G, Cruaud C, Istace B, et al. Telomere-to-telomere  
1295 gapless chromosomes of banana using nanopore sequencing. *Commun Biol.*  
1296 2021;4(1):1047.

1297 78. Li B, Choulet F, Heng Y, Hao WW, Paux E, Liu Z, et al. Wheat centromeric  
1298 retrotransposons: the new ones take a major role in centromeric structure. *Plant J.*  
1299 2013;73(6):952–65.

1300 79. Berrens RV, Yang A, Laumer CE, Lun ATL, Bieberich F, Law CT, et al. Transposable  
1301 element expression at unique loci in single cells with CELLO-seq. *bioRxiv*. 2020;322073.

1302 80. Hirsch CD, Springer NM. Transposable element influences on gene expression in plants.  
1303 *Biochim Biophys Acta*. 2017;1860(1):157–65.

1304 81. Richert-Pöggeler KR, Vijverberg K, Alisawi O, Chofong GN, Heslop-Harrison JS,  
1305 Schwarzacher T. Participation of multifunctional RNA in replication, recombination and  
1306 regulation of Endogenous Plant Pararetroviruses (EPRVs). *Front Plant Sci.*  
1307 2021;12:689307.

1308 82. Cheng F, Sun C, Wu J, Schnable J, Woodhouse MR, Liang J, et al. Epigenetic regulation  
1309 of subgenome dominance following whole genome triplication in *Brassica rapa*. *New*  
1310 *Phytol*. 2016;211(1):288–99.

1311 83. Ahokas H. Unfecund, gigantic mutant of oats (*Avena sativa*) shows fecundity  
1312 overdominance and difference in DNA methylation properties. In: Tigerstedt PMA, editor.  
1313 *Adaptation in plant breeding*. Jyväskylä: Springer Science Business Media B.V.; 1997. p.  
1314 21–26.

1315 84. Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. Polyploidy and genome evolution in  
1316 plants. *Curr Opin Genet Dev*. 2015;35:119–25.

1317 85. Alix K, Gérard PR, Schwarzacher T, Heslop-Harrison JS. Polyploidy and interspecific  
1318 hybridization: partners for adaptation, speciation and evolution in plants. *Ann Bot*.  
1319 2017;120(2):183–94.

1320 86. Yu F, Zhoa XW, Chai J, Ding XE, Li XT, Huang YJ, et al. Chromosome-specific painting  
1321 unveils chromosomal fusions and distinct allopolyploid species in the *Saccharum*  
1322 complex. *New Phytol.* 2022;233(4):1953–65.

1323 87. Johnson MG, Pokorny L, Dodsworth S, Botigue LR, Cowan RS, Devault A, et al. A  
1324 universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant  
1325 designed using k-methods clustering. *Syst Biol.* 2018;68(4):594–606.

1326 88. Katsiotis A, Loukas M, Heslop-Harrison JS. Repetitive DNA, genome and species  
1327 relationships in *Avena* and *Arrhenatherum*. *Ann Bot.* 2000;86(6):1135–42.

1328 89. Birchler JA, Han FP. Barbara McClintock's unsolved chromosomal mysteries: parallels  
1329 to common rearrangements and karyotype evolution. *Plant Cell.* 2018;30(4):771–9.

1330 90. Yang XF, Gao SH, Guo L, Wang B, Jia YY, Zhou J, et al. Three chromosome-scale  
1331 *Papaver* genomes reveal punctuated patchwork evolution of the morphinan and noscapine  
1332 biosynthesis pathway. *Nat Commun.* 2021;12(1):6030.

1333 91. Ladizinsky G. My research findings in *Avena*. In: Ladizinsky G, editor. *Studies in oat*  
1334 *evolution a man's life with Avena*. Heidelberg: Springer; 2012. p. 19–66.

1335 92. Rai A, Hirakawa H, Nakabayashi R, Kikuchi S, Hayashi K, Rai M, et al. Chromosome-  
1336 level genome assembly of *Ophiorrhiza pumila* reveals the evolution of camptothecin  
1337 biosynthesis. *Nat Comm.* 2021;12(1):405

1338 93. Meng W, Yuan JR, Qin LM, Shi WM, Xia GM, Liu SW. *TaCYP81D5*, one member in a  
1339 wheat cytochrome P450 gene cluster confers salinity tolerance via reactive oxygen  
1340 species scavenging. *Plant Biotechnol J.* 2019;18(3):791–804.

1341 94. Jiao YN, Wickett NJ, Ayyampalayam A, Chanderbali AS, Landherr L, Ralph PE, et al.  
1342 Ancestral polyploidy in seed plants and angiosperms. *Nature.* 2011;473(7345):97–100.

1343 95. Hunter CT, Kirienko DH, Sylvester AW, Peter GF, McCarty DR, Koch KE. Cellulose  
1344 synthase-like D1 is integral to normal cell division, expansion, and leaf development in  
1345 maize. *Plant Physiol.* 2012;158(2):708–24.

1346 96. Karas BJ, Ross L, Novero M, Amyot L, Shrestha A, Inada S, et al. Intragenic  
1347 complementation at the *Lotus japonicas* *CELLULOSE SYNTHASE-LIKE D1* locus  
1348 rescues root hair defects. *Plant Physiol.* 2021;186(4):2037–50.

1349 97. Kraemer FJ, Lunde C, Koch M, Kuhn BM, Ruehl C, Brown PJ, et al. A mixed-linkage  
1350 (1,3;1,4)- $\beta$ -D-glucan specific hydrolase mediates dark-triggered degradation of this cell  
1351 wall polysaccharide. *Plant Physiol.* 2021;185(4):1559–73.

1352 98. Kozarewa I, Ning ZM, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-  
1353 free illumina sequencing-library preparation facilitates improved mapping and assembly  
1354 of (G+C)-biased genomes. *Nat Methods.* 2009;6(4):291–5.

1355 99. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of  
1356 occurrences of  $k$ -mers. *Bioinformatics.* 2011;27(6):764.

1357 100. Vulture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al.  
1358 GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics.*  
1359 2017;33(14):2202–4.

1360 101. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford  
1361 Nanopore sequencing. *Genome Biol.* 2019;20(1):129.

1362 102. Hu J, Fan JP, Sun ZY, Liu SL. NextPolish: a fast and efficient genome polishing tool for  
1363 long-read assembly. *Bioinformatics.* 2020;36(7):2253–5.

1364 103. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods.* 2020;  
1365 17(2):155–8.

1366 104. Istance B, Friedrich A, d'Agata L, Faye S, Payen E, Beluche O, et al. *De novo* assembly  
1367 and population genomic survey of natural yeast isolates with the Oxford Nanopore  
1368 MinION sequencer. *Gigascience*. 2017;6(2):1–13.

1369 105. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor.  
1370 *Bioinformatics*. 2018;34(17):i884–90.

1371 106. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, et al. HiC-Pro: an  
1372 optimized and flexible pipeline for Hi-C data processing. *Genome Biol*. 2015;16:259.

1373 107. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*.  
1374 2012;9(4):357–9.

1375 108. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al. Juicer  
1376 provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst*.  
1377 2016;3(1):95–8.

1378 109. Burton JN, Adey A, Patwardhan RP, Qiu RL, Kitzman JO, Shendure J. Chromosome-  
1379 scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat*  
1380 *Biotechnol*. 2013;31(12):1119–25.

1381 110. Lieberman-Aiden E, van Berkum N, Williams L, Imakaev M, Ragoczy T, Telling A, et  
1382 al. Comprehensive mapping of long-range interactions reveals folding principles of the  
1383 human genome. *Science*. 2009;326(5950):289–93.

1384 111. R Core Team. R: A language and environment for statistical computing. R Foundation  
1385 for statistical Computing, Vienna, Austria. 2020. URL <https://www.R-project.org/>.

1386 112. Doležel J, Greilhuber J, Suda J. Estimation of nuclear DNA content in plants using flow  
1387 cytometry. *Nat Protoc*. 2007;2(9):2233–44.

1388 113. Otto F. DAPI staining of fixed cells for high-resolution flow cytometry of nuclear DNA.  
1389 *Methods Cell Biol*. 1990;33:105–10.

1390 114. Luo RB, Liu BH, Xie YL, Li ZY, Huang WH, Yuan JY, et al. SOAPdenovo2: an  
1391 empirically improved memory-efficient short-read *de novo* assembler. *Gigascience*.  
1392 2012;1(1):18.

1393 115. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*.  
1394 2018;34(18):3094–100.

1395 116. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence  
1396 alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9.

1397 117. Ou SJ, Jiang N. LTR\_retriever: a highly accurate and sensitive program for identification  
1398 of long terminal repeat retrotransposons. *Plant Physiol*. 2018;176(2):1410–22.

1399 118. Ellinghaus D, Kurtz S, Willhöft U. LTRharvest, an efficient and flexible software for *de*  
1400 *novo* detection of LTR retrotransposons. *BMC Bioinformatics*. 2008;9:18.

1401 119. Shi JM, Liang C. Generic repeat finder: a high-sensitivity tool for genome-wide *de novo*  
1402 repeat detection. *Plant Physiol*. 2019;180(4):1803–15.

1403 120. Su W, Gu X, Peterson T. TIR-learner, a new ensemble method for TIR transposable  
1404 element annotation, provides evidence for abundant new transposable elements in the  
1405 maize genome. *Mol Plant*. 2016;12(3):447–60.

1406 121. Xiong W, He L, Lai J, Dooner HK, Du C. HelitronScanner uncovers a large overlooked  
1407 cache of *Helitron* transposons in many plant genomes. *Proc Natl Acad Sci U S A*. 2014;  
1408 111(28):10263–8.

1409 122. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2  
1410 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U*  
1411 *S A*. 2020;117(17):9451–7.

1412 123. Smit AF, Hubley R, Green P. RepeatModeler Open-1.0. 2008–2015. Seattle: Institute for  
1413 Systems Biology; 2015.

1414 124. Zhang RG, Wang ZX, Ou S, Li GY. TEsorter: lineage-level classification of transposable  
1415 elements using conserved protein domains. bioRxiv. 2019;800177.

1416 125. Chen JH, Hao ZD, Guang XM, Zhao CX, Wang PK, Xue LJ, et al. *Liriodendron* genome  
1417 sheds light on angiosperm phylogeny and species–pair differentiation. Nature Plants.  
1418 2019;5(1):18–25.

1419 126. Avni R, Nave M, Barad O, Baruch K, Twardziok SO, Gundlach H, et al. Wild emmer  
1420 wheat genome architecture and diversity elucidate wheat evolution and domestication.  
1421 Science. 2017;357(6346):93–7.

1422 127. Yandell M, Ence D. A beginner’s guide to eukaryotic genome annotation. Nat Rev Genet.  
1423 2012;13(5):329–42.

1424 128. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: *ab*  
1425 *initio* prediction of alternative transcripts. Nucleic Acids Res. 2006;34(Web  
1426 Server):W435–9.

1427 129. Keilwagen J, Hartung F, Grau J. GeMoMa: Homology-based gene prediction utilizing  
1428 intron position conservation and RNA-seq data. Methods Mol Biol. 2019;1962:161–77.

1429 130. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. *De novo*  
1430 transcript sequence reconstruction from RNA-seq using the Trinity platform for reference  
1431 generation and analysis. Nat Protoc. 2013;8(8):1494–512.

1432 131. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic  
1433 gene structure annotation using EVidenceModeler and the program to assemble spliced  
1434 alignments. Genome Biol. 2008;9(1):R7.

1435 132. Le Mercier P, Bougueret L. The universal protein resource (UniProt). Nucleic Acids  
1436 Res. 2007;36(Database issue):D190–5.

1437 133. Conesa A, Götz S, García-Gómez J, Terol J, Talon M, Robles M. BLAST2GO: a  
1438 universal tool for annotation, visualization and analysis in functional genomics research.  
1439 *Bioinformatics*. 2005;21(18):3674–6.

1440 134. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam:  
1441 annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*  
1442 2005;33(Database issue): D121–4.

1443 135. Chan PP, Lowe TM. tRNAscan-SE: searching for tRNA genes in genomic sequences.  
1444 *Methods Mol Biol.* 2019;1962:1–14.

1445 136. Betel D, Wilson M, Gabow A, Marks DS, Sander C. The microRNA.org resource: targets  
1446 and expression. *Nucleic Acids Res.* 2008;36(Database issue):D149–53.

1447 137. Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW. RNAmmer:  
1448 consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*  
1449 2007;35(9):3100–8.

1450 138. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast  
1451 universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.

1452 139. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie  
1453 enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol.*  
1454 2015;33(3):290–5.

1455 140. David MG, Shu SQ, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a  
1456 comparative platform for green plant genomics. *Nucleic Acids Res.* 2012;40(Database  
1457 issue):D1178–86.

1458 141. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND.  
1459 *Nat Methods.* 2015;12(1):59–60.

1460 142. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of  
1461 intersecting sets and their properties. *Bioinformatics*. 2017;33(18):2938–40.

1462 143. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7:  
1463 improvements in performance and usability. *Mol Bio Evol*. 2013;30(4):772–80.

1464 144. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of  
1465 large phylogenies. *Bioinformatics*. 2014;30(9):1312–3.

1466 145. Yang ZH. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*.  
1467 2007;24(8):1586–91.

1468 146. De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study  
1469 of gene family evolution. *Bioinformatics*. 2006;22(10):1269–71.

1470 147. Chen CJ, Chen H, Zhang Y, Thomas HR, Frank MH, He YH, et al. TBtools: an integrative  
1471 toolkit developed for interactive analyses of big biological data. *Mol Plant*.  
1472 2020;13(8):1194–202.

1473 148. Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for  
1474 detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*.  
1475 2012;40(7):e49.

1476 149. Wang J, Yu J, Sun P, Li Y, Xia R, Liu Y, et al. Comparative genomics analysis of rice  
1477 and pineapple contributes to understand the chromosome number reduction and genomic  
1478 changes in grasses. *Front Genet*. 2016;7:174.

1479 150. Sun PC, Jiao BB, Yang YZ, Shan LX, Li T, Li XN, et al. WGDI: a user-friendly toolkit  
1480 for evolutionary analyses of whole-genome duplications and ancestral karyotypes.  
1481 *bioRxiv*. 2021;441969.

1482 151. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs\_Calculator 2.0: a toolkit incorporating  
1483 gamma-series methods and sliding window strategies. *Genomics Proteomics  
1484 Bioinformatics*. 2010;8(1):77–80.

1485 152. El Baidouri M, Murat F, Veyssiere M, Molinier M, Flores R, Burlot L, et al. Reconciling  
1486 the evolutionary origin of bread wheat (*Triticum aestivum*). *New Phytol.*  
1487 2016;213(3):1477–86.

1488 153. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for  
1489 RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550.

1490 154. Kautsar SA, Duran HGS, Blin K, Osbourn A, Medema MH. PlantSMASH: automated  
1491 identification, annotation and expression analysis of plant biosynthetic gene clusters.  
1492 *Nucleic Acids Res.* 2017;45(Web Server):W55–63.

1493 155. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al.  
1494 Pfam: the protein families database in 2021. *Nucleic Acids Res.* 2021;49(Database  
1495 issue):D412–9.

1496 156. Kaur S, Dhugga KS, Beech R, Singh J. Genome-wide analysis of the cellulose synthase-  
1497 like (Csl) gene family in bread wheat (*Triticum aestivum* L.). *BMC Plant Biol.*  
1498 2017;17:193.

1499 157. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for  
1500 large alignments. *PLoS One.* 2010;5:e9490.

## 1501 **Figure legends**

1502 **Fig. 1** Genomic landscape of seven assemble chromosomes ALO01 to ALO07 of *Avena*  
1503 *longiglumis* (ALO). **a** Chromosome names and sizes (100 Mbp intervals indicated) with  
1504 centromere position marked in pink and major 5S (ALO07 in green) and 45S (ALO01 and  
1505 ALO07 in red) rDNA sites indicated. **b** Transposable element (TE, pink) density along each  
1506 chromosome. **c** LTR TE (purple) density (1 Mbp nonoverlapping windows) along each  
1507 chromosome. **d** Long interspersed nuclear element (*LINE*) density (orange) along each  
1508 chromosome. **e** *Helitron* density (cyan) along each chromosome. **f** Expanded gene locations in  
1509 each chromosome. **g** Contracted gene locations in each chromosome. **h** Single copy orthologue  
1510 gene locations in each chromosome. **i** High-confidence gene locations in each chromosome. **j**  
1511 Purified selection gene locations in each chromosome (these genes with  $P$ -value  $\leq 0.05$ ). **k**  
1512 Expression profiling of genes on each chromosome in ALO roots. **l** Expression profiling of  
1513 genes on each chromosome in ALO leaves. **m** Links between syntenic genes. Orientation in  
1514 outward in circles **b, c, d, e, k** and **l**.

1515 **Fig. 2** Analysis of TEs in *Avena longiglumis* (ALO) genome. **a** Genomic constituent in ALO  
1516 in comparison with those in ATA, BDI, OSA, SBI, SIT and ZMA. Note that the six  
1517 constituents, especially *Gypsy*, *Copia* and unclassified LTR TEs, were much more abundant in  
1518 ALO than in other grasses. **b** Top 10 TE families in ALO and the percentages of these families  
1519 in ATA, BDI, OSA, SBI, SIT and ZMA. Five *Gypsy* families, *Angela*, *Tekay*, *Retland*, *Athila*  
1520 and *CRM*, showed increased abundance in ALO relative to those in ATA, BDI, OSA, SBI, SIT  
1521 and ZMA. **c** Temporal patterns of LTR-RT insertion bursts in ALO as compared to those in  
1522 ATA, BDI, OSA, SBI, SIT and ZMA. The number of intact LTR-RTs used for each species is  
1523 given in parentheses. **d** Insertion bursts of *Gypsy* and *Copia* elements in ALO. The numbers of  
1524 intact elements used for this analysis are provided in parentheses. *Avena longiglumis* (ALO),

1525 *Aegilops tauschii* (ATA) [32], *Brachypodium distachyon* (BDI) [33], *Oryza sativa* (OSA) [9],  
1526 *Sorghum bicolor* (SBI) [10], *Setaria italica* (SIT) [11] and *Zea mays* (ZMA) [34].

1527 **Fig. 3** Evolution of the *Avena longiglumis* (ALO) genome. **a** Phylogenetic relationship of ALO  
1528 with ten plant species. C<sub>3</sub> species are shown with yellow background and C<sub>4</sub> species with blue  
1529 background. Divergence times are labelled in blue; gene family expansion and contraction are  
1530 enumerated below the species names in green and red. **b** Gene categories are shown for all the  
1531 species in Fig. 4a. **c** Distribution of ks distance between syntenic orthologous genes for ALO,  
1532 AER, AST, OSA and ZMA genomes. **d** UpSetR diagram of shared orthologous gene families  
1533 in five species. The number of gene families is listed for each component. *A. atlantica* (AAT)  
1534 [12], *A. eriantha* (AER) [12], *A. longiglumis* (ALO), *A. strigosa* (AST) [7], *Aegilops tauschii*  
1535 (ATA) [32], *Arabidopsis thaliana* (ATH) [50], *Brachypodium distachyon* (BDI) [33], *Oryza*  
1536 *sativa* (OSA) [9], *Sorghum bicolor* (SBI) [10], *Setaria italica* (SIT) [11], *Zea mays* (ZMA)  
1537 [34].

1538 **Fig. 4** Syntenic relationships of chromosomes of the ancestral grass karyotype (AGK) and  
1539 analysed species. **a** Syntenic analysis of *Avena strigosa* (AST), *A. longiglumis* (ALO) and *A.*  
1540 *eriantha* (AER). Subterminal regions are frequently involved in interspecific evolutionary  
1541 translocations. **b** Reconstruction of ancestral chromosomes for the seven species showing  
1542 conservation of major syntenic blocks from the ancestral grass karyotype (AGK) with fusions  
1543 and insertions leading to the reduced chromosome numbers (chromosomes numbered by  
1544 published linkage groups, some are upside down to display features of evolutionary  
1545 conservation). **c** Deep syntenic analysis of *Oryza sativa* (OSA), ALO and *Brachypodium*  
1546 *distachyon* (BDI) showing detailed conservation of syntenic block and the expansions between  
1547 OSA ( $x = 12,389$  Mbp), BDI ( $x = 5,260$  Mbp) and ALO ( $x = 7,3,960$  Mbp). Genes from the  
1548 ancestral linkage groups are indicated by colours, with pairs of similar colours representing the  
1549 pre-rho whole genome duplication.

1550 **Fig. 5** Identification of biosynthetic gene clusters (BGCs) and *cellulose synthase A (CesA)* and  
1551 *cellulose-like (Csl)* gene families in ALO. **a** Total 109 BGCs identified in ALO chromosomes  
1552 by plantiSMASH. The cluster types, including alkaloid, lignin, polyketide, saccharide, terpene,  
1553 lignin\_polyketide, lignin\_saccharide, lignin\_terpene, saccharide\_alkaloid,  
1554 saccharide\_polyketide, saccharide\_terpene, saccharide\_terpene\_alkaloid, and  
1555 terpene\_polyketide biosynthesis genes, labeled as different colour. The cluster position shown  
1556 by blue band. The centromere position shown by pink band. A scale in the left represented  
1557 length of chromosome in megabases (Mbp). **b** Maximum likelihood phylogenetic tree of CesA  
1558 and CSL proteins from ALO, rice, wheat and arabidopsis. Nongroup: ALO CSL proteins are  
1559 not clustered with any known CesA and Csl proteins. **c** Heat map showing hierarchical  
1560 clustering of *CesA* and *Csl* gene families in roots, salt-treated roots, leaves and salt-treated  
1561 leaves of ALO. Expression values were normalized by  $\log_2(\text{FPKM} + 1)$ . Highly and weakly  
1562 expressed genes were colored by red and blue boxes, respectively.

1563

## Tables

1564

**Table 1** *Avena longiglumis* genome statistics and gene predictions

	Number	Size
Assembly feature		
Estimated genome size		4.60 Gbp
Assembled sequences		3,960,768,570 bp
N50 contig length		12,682,464 bp
Longest contig		99,445,397 bp
N50 scaffold length		527,343,613 bp
N90 scaffold length		6,968,329 bp
Number of scaffolds (> N90)	9	
Longest scaffold (bp)		594,546,470 bp
Repetitive DNAs		
Retrotransposons		3,198,067,781 bp (80.74%)
DNA transposons		137,389,012 bp (3.47%)
Total		3,447,484,807 bp (87.04%)
Gene annotation (pseudo-chromosomes and unanchored)		
Gene models (high confidence)	33,271	115,042,134 bp
Gene models (low confidence)	7,574	18,590,004 bp
Total genes	40,845	133,632,138 bp
Non-coding RNAs	16,439	2,222,342 bp

1565

**Table 2 Pseudo-chromosome length and gene content of *Avena longiglumis* (ALO)**

Chromosome number	Pseudomolecular length (bp)	Arm ratio*	Gene number**	Mean length (bp)	Median length (bp)	Minimum length (bp)	Maximum length (bp)	High confidence gene number***
ALO01	594,546,470	1.24	6,371	3370.96	1932	163	254,605	5,186
ALO02	587,543,788	1.14	5,976	3229.50	1928.5	165	158,444	4,814
ALO03	587,190,583	1.07	6,417	3298.18	2069	150	105,522	5,323
ALO04	583,925,327	1.07	5,846	3262.89	2018.5	163	169,716	4,895
ALO05	527,343,613	1.39	4,212	3339.13	2005	201	145,603	3,493
ALO06	513,337,126	1.08	5,413	3272.85	1997	163	154,734	4,375
ALO07	453,691,697	2.01 NOR	5,256	3193.56	1975.5	158	165,822	4,100
Anchored genome	3,847,578,604		39,491	22967.07	13925.5	1163	1,154,446	32,186
Unanchored	113,189,966		1,354					1,085

1567 \* Arm ratio = (long arm length)/(short arm length); NOR has secondary constriction at the major Nucleolar Organizer Region on short arm.

1568 \*\* Number of genes anchored on chromosomes.

1569 \*\*\* Number of high confidence genes supported by transcriptome data with FPKM value larger than zero or homology. For the genes without transcriptome  
1570 transcript abundance support, the alignment was performed with *A. atlantica* (identity > 95%, coverage > 95%), *A. eriantha* (identity > 90%, coverage > 90%),  
1571 *Hordeum vulgare* and *Triticum aestivum* (identity > 80%, coverage > 80%) by BLASTP (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>; *E* value = 1e-5), respectively.  
1572 Those supported by alignment results of two or more species alignments were defined as high confidence genes.

Figure 1

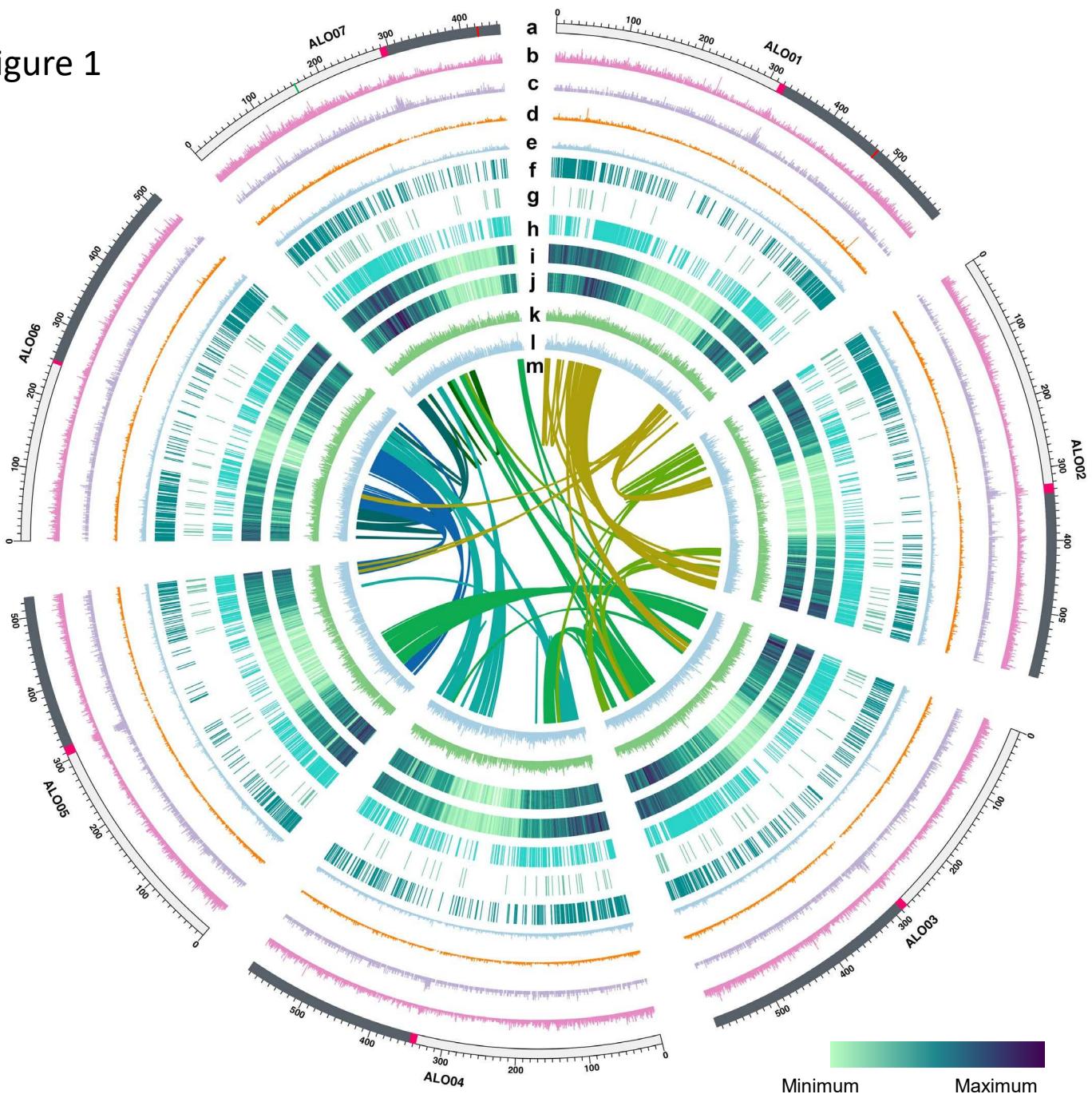
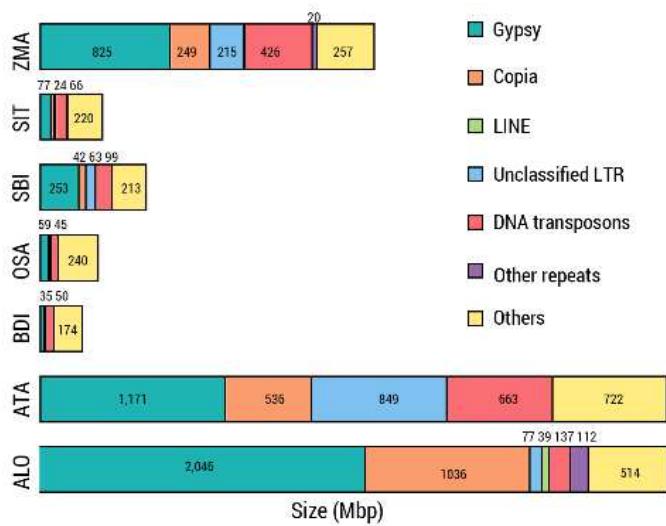
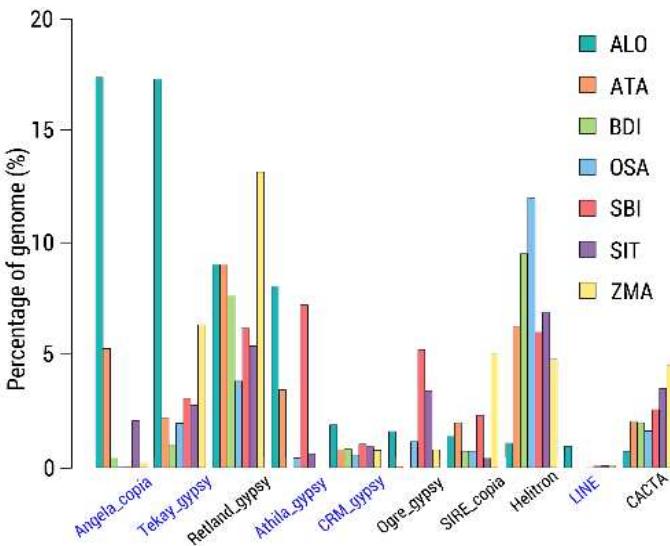


Figure 2

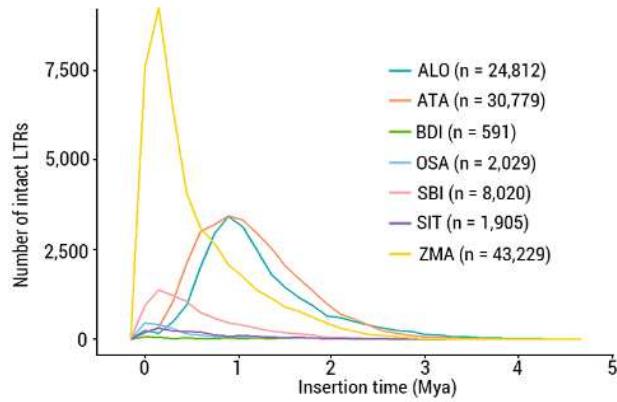
a



b



c



d

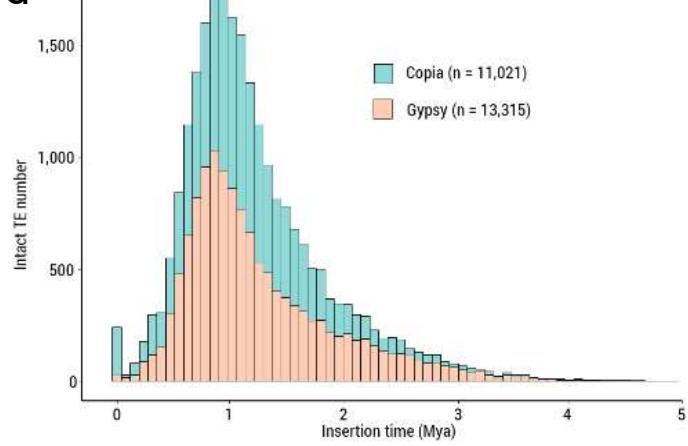


Figure 3

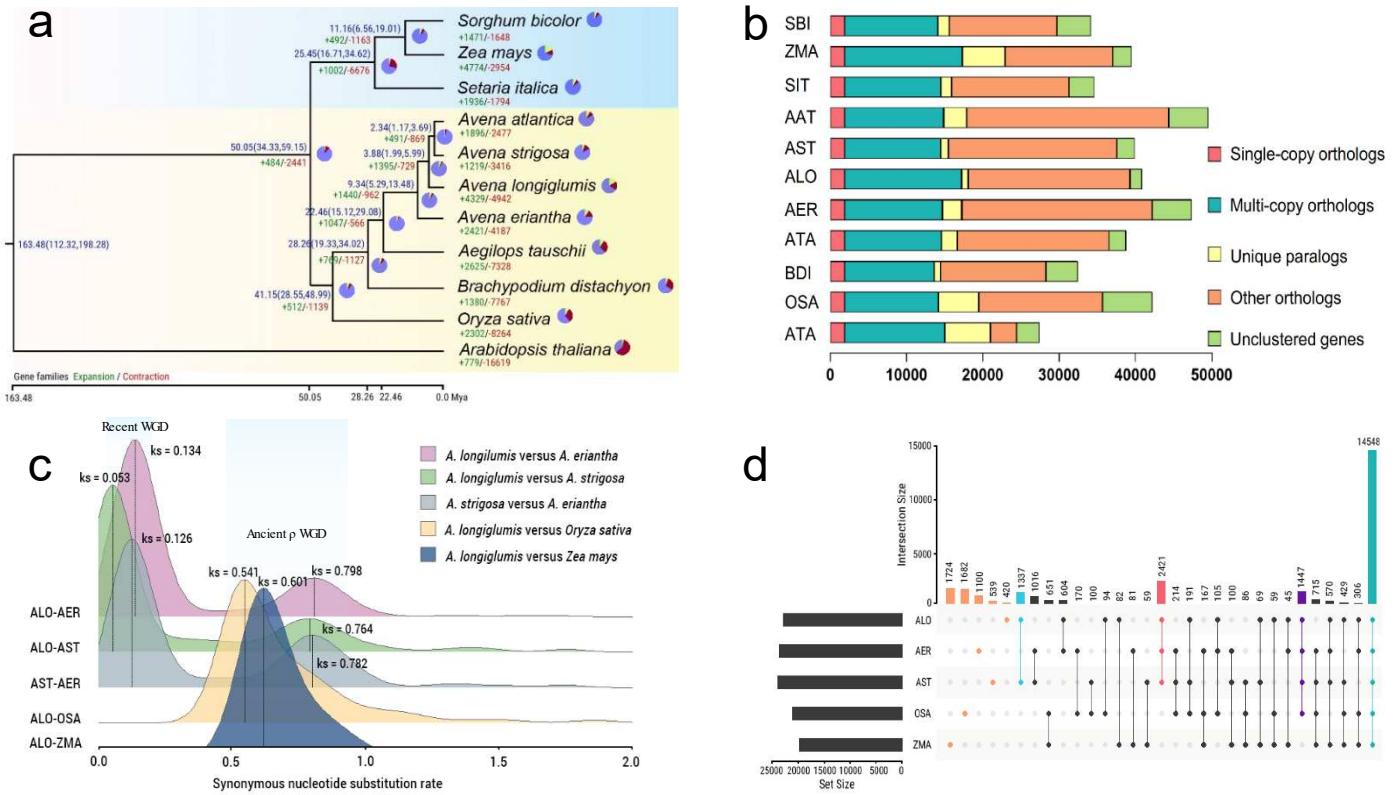
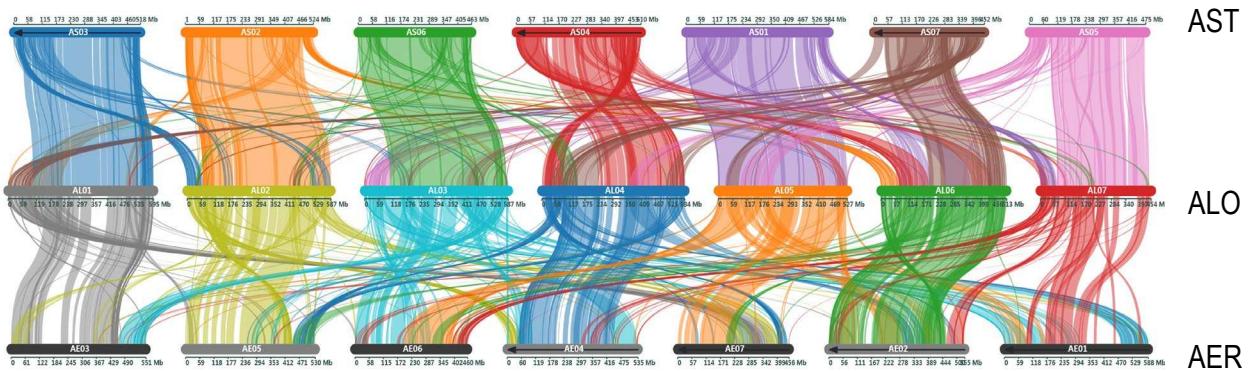
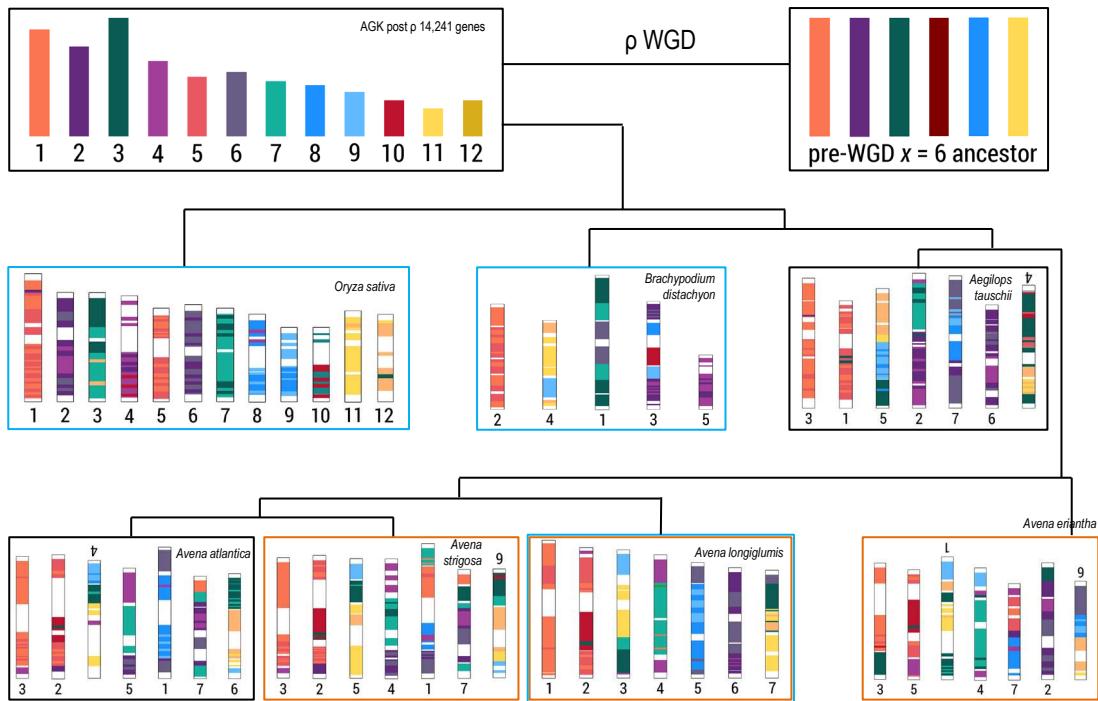


Figure 4

a



b



c

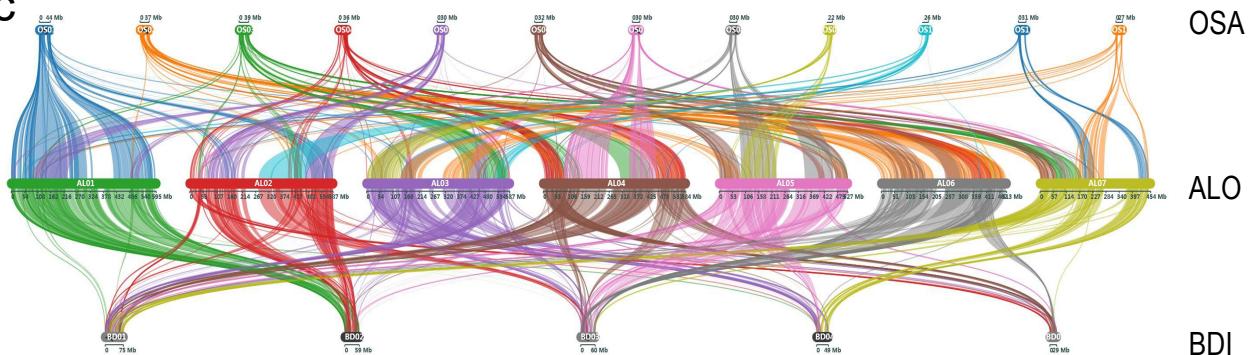


Figure 5

a

