

# The mosaic oat genome gives insights into a uniquely healthy cereal crop

<https://doi.org/10.1038/s41586-022-04732-y>

Received: 28 July 2021

Accepted: 6 April 2022

Published online: 18 May 2022

Open access

 Check for updates

Nadia Kamal<sup>1,22</sup>, Nikos Tsardakas Renhuldt<sup>2,22</sup>, Johan Bentzer<sup>2</sup>, Heidrun Gundlach<sup>1</sup>, Georg Haberer<sup>1</sup>, Angéla Juhász<sup>3</sup>, Thomas Lux<sup>1</sup>, Utpal Bose<sup>3,4</sup>, Jason A. Tye-Din<sup>5,6</sup>, Daniel Lang<sup>1,20</sup>, Nico van Gessel<sup>7</sup>, Ralf Reski<sup>7</sup>, Yong-Bi Fu<sup>8</sup>, Peter Spégel<sup>9</sup>, Alf Cęplitis<sup>10</sup>, Axel Himmelbach<sup>11</sup>, Amanda J. Waters<sup>12</sup>, Wubishet A. Bekele<sup>13</sup>, Michelle L. Colgrave<sup>3,4</sup>, Mats Hansson<sup>14</sup>, Nils Stein<sup>11,15</sup>, Klaus F. X. Mayer<sup>11,16</sup>, Eric N. Jellen<sup>17</sup>, Peter J. Maughan<sup>17</sup>, Nicholas A. Tinker<sup>13</sup>, Martin Mascher<sup>11,18</sup>, Olof Olsson<sup>19</sup>, Manuel Spannagl<sup>1,20</sup> & Nick Sirijovski<sup>2,19,21</sup>✉

Cultivated oat (*Avena sativa* L.) is an allohexaploid (AACCCDD,  $2n = 6x = 42$ ) thought to have been domesticated more than 3,000 years ago while growing as a weed in wheat, emmer and barley fields in Anatolia<sup>1,2</sup>. Oat has a low carbon footprint, substantial health benefits and the potential to replace animal-based food products. However, the lack of a fully annotated reference genome has hampered efforts to deconvolute its complex evolutionary history and functional gene dynamics. Here we present a high-quality reference genome of *A. sativa* and close relatives of its diploid (*Avena longiglumis*, AA,  $2n = 14$ ) and tetraploid (*Avena insularis*, CCDD,  $2n = 4x = 28$ ) progenitors. We reveal the mosaic structure of the oat genome, trace large-scale genomic reorganizations in the polyploidization history of oat and illustrate a breeding barrier associated with the genome architecture of oat. We showcase detailed analyses of gene families implicated in human health and nutrition, which adds to the evidence supporting oat safety in gluten-free diets, and we perform mapping-by-sequencing of an agronomic trait related to water-use efficiency. This resource for the *Avena* genus will help to leverage knowledge from other cereal genomes, improve understanding of basic oat biology and accelerate genomics-assisted breeding and reanalysis of quantitative trait studies.

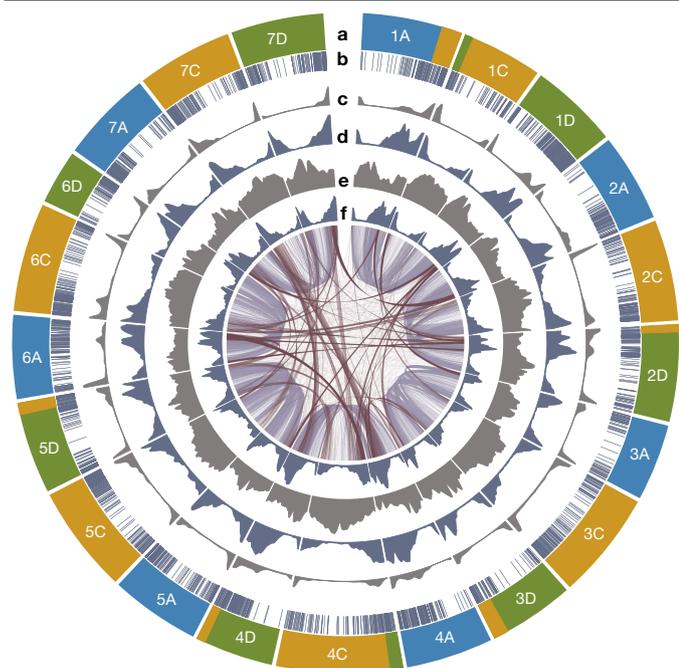
Oat is a member of Poaceae, an economically important grass family that includes wheat, rice, barley, common millet, maize, sorghum and sugarcane. *Avena* species exist in nature as diploids, tetraploids and hexaploids and exhibit the greatest genetic diversity around the Mediterranean, Middle East, Canary Islands and Himalayas. Currently, oat is a global crop with production ranking seventh among cereals (<http://www.fao.org/faostat/en/>, accessed May 2021). Compared with that of other cereals, oat cultivation requires fewer treatments with insecticides, fungicides or fertilizers. Whole-grain oats are a healthy source of antioxidants, polyunsaturated fatty acids, proteins and dietary fibre such as  $\beta$ -glucan, which is important in post-meal glycaemic responses and for preventing cardiovascular disease<sup>3–5</sup>. Cereals such as wheat,

barley and rye store high amounts of gluten proteins in their grain; by contrast, oat and rice store globular proteins in their grain.

## Genome assembly and composition

We produced a chromosome-scale reference sequence of oat cv. ‘Sang’ comprising 21 pseudochromosomes (Fig. 1, Extended Data Fig. 1a and Supplementary Table 1), with a BUSCO (v5.1.2; ref. <sup>6</sup>) score of 98.7% (Extended Data Fig. 2a), following the short-read strategy used for wheat<sup>7</sup>, barley<sup>8</sup> and rye<sup>9</sup>. Inspection of Hi-C contact matrices (Supplementary Fig. 1) and the consensus genetic map<sup>10</sup> (Supplementary Fig. 2a) and their comparison with the independent assembly

<sup>1</sup>Plant Genome and Systems Biology, German Research Center for Environmental Health, Helmholtz Zentrum München, Neuherberg, Germany. <sup>2</sup>ScanOats Industrial Research Centre, Department of Chemistry, Division of Pure and Applied Biochemistry, Lund University, Lund, Sweden. <sup>3</sup>Australian Research Council Centre of Excellence for Innovations in Peptide and Protein Science, School of Science, Edith Cowan University, Joondalup, Western Australia, Australia. <sup>4</sup>Agriculture and Food, Commonwealth Scientific and Industrial Research Organisation, St Lucia, Queensland, Australia. <sup>5</sup>Immunology Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia. <sup>6</sup>Department of Gastroenterology, Royal Melbourne Hospital, Parkville, Victoria, Australia. <sup>7</sup>Plant Biotechnology, Faculty of Biology, University of Freiburg, Freiburg, Germany. <sup>8</sup>Plant Gene Resources of Canada, Agriculture and Agri-Food Canada, Saskatoon, Saskatchewan, Canada. <sup>9</sup>Department of Chemistry, Centre for Analysis and Synthesis, Lund University, Lund, Sweden. <sup>10</sup>Plant Breeding, Lantmännen, Svalöv, Sweden. <sup>11</sup>Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Germany. <sup>12</sup>Research and Development Division, PepsiCo, St Paul, MN, USA. <sup>13</sup>Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, Ontario, Canada. <sup>14</sup>Molecular Cell Biology, Department of Biology, Lund University, Lund, Sweden. <sup>15</sup>Department of Crop Sciences, Center of Integrated Breeding Research (CiBreed), Georg-August-University, Göttingen, Germany. <sup>16</sup>School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany. <sup>17</sup>Department of Plant and Wildlife Sciences, Brigham Young University, Provo, UT, USA. <sup>18</sup>German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Leipzig, Germany. <sup>19</sup>CropTailor AB, Department of Chemistry, Division of Pure and Applied Biochemistry, Lund University, Lund, Sweden. <sup>20</sup>Present address: Department of Microbial Genomics and Bioforensics, Bundeswehr Institute of Microbiology, Munich, Germany. <sup>21</sup>Present address: Food Science Organisation, Oatly AB, Lund, Sweden. <sup>22</sup>These authors contributed equally: Nadia Kamal, Nikos Tsardakas Renhuldt. ✉e-mail: manuel.spannagl@helmholtz-muenchen.de; nick.sirijovski@biokem.lth.se



**Fig. 1 | Structural and functional landscape of the 21 oat pseudochromosomes.** a–f, The tracks from the outer circles towards the centre of the diagram display the chromosome name and subgenome origin (A, blue; C, gold; D, green) with major translocations (a); anchored oat genetic markers<sup>51</sup> (b); distribution of recombination rates (c); density and genomic distribution of high-confidence genes (d); age distribution of long terminal repeat retrotransposons (e); and median gene expression in 1-Mb windows (f). Inner connections show the best bidirectional BLAST hits between genes on homoeologous chromosomes (grey) and between genes on non-homoeologous chromosomes (dark red). Figure generated with Circa (<http://omgenomics.com/circa>).

(long-read) of hexaploid oat OT3098 (ref. <sup>11</sup>; version 2; Supplementary Table 2) verified the integrity of the assembly (Extended Data Fig. 2b and Supplementary Fig. 2). We also assembled pseudochromosomes of the diploid *Avena longiglumis* and tetraploid *Avena insularis*, which are presumed A and CD subgenome progenitors of *Avena sativa*<sup>12</sup> (Extended Data Figs. 1a and 2a and Supplementary Figs. 3 and 4). Phylogenomic analyses (Supplementary Fig. 5) used to assign *A. sativa* chromosomes to subgenomes showed that gene order is conserved in the proximal chromosomal regions. The 21 *A. sativa* chromosomes, named 1A–7D following the subgenome assignments of ref. <sup>13</sup>, were oriented to preserve the orientations of core regions across homoeologues and possibly between *Avena* and Triticeae. Alignments to barley (Extended Data Fig. 1b), *Avena eriantha*<sup>14</sup> (Supplementary Fig. 6), *A. longiglumis* (Supplementary Fig. 7a) and *A. insularis* (Supplementary Fig. 7b) confirmed the validity of this revised nomenclature, which is accepted by the International Oat Nomenclature Committee<sup>15</sup>.

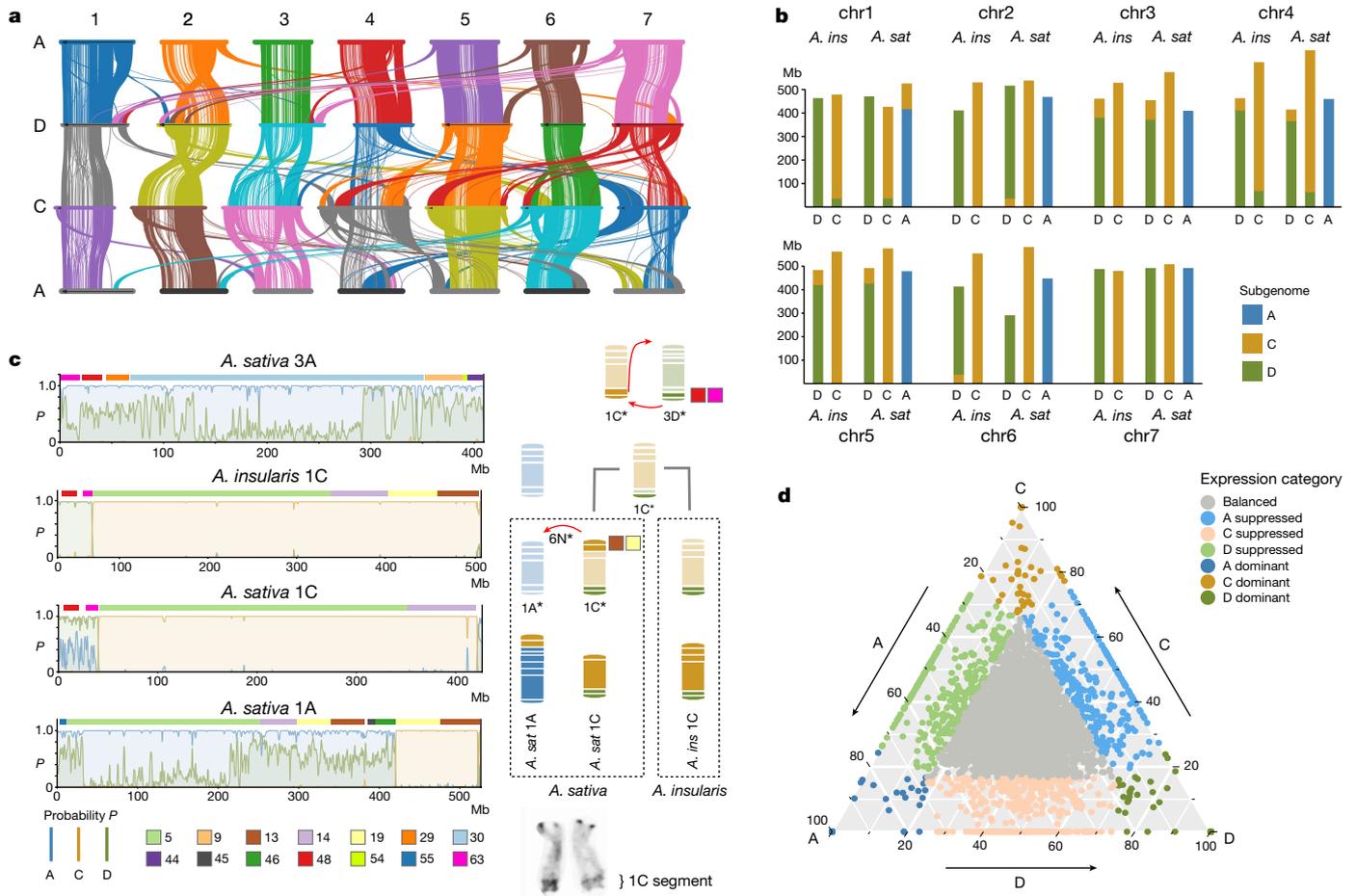
We predicted gene models in the oat genome using an automated annotation pipeline<sup>16</sup>, assisted by RNA-sequencing (RNA-seq) and Iso-seq transcriptome data, protein homology and ab initio prediction. This yielded 80,608 high-confidence protein-coding loci (98.5% BUSCO; Extended Data Fig. 2c and Supplementary Table 3), 83.5% of which showed evidence of transcription in at least one condition. Another 71,277 low-confidence protein-coding loci primarily represent gene fragments, pseudogenes and gene models with weak support. The overall amount and composition of transposable elements is very similar between the Sang and OT3098 assemblies (Supplementary Tables 4 and 5 and Supplementary Fig. 8). Transposable elements accounted for 64% of the oat genome sequence. The size difference of about 1 Gb

observed between the C and A or D subgenome probably reflects higher transposon activity in the diploid ancestor of the C subgenome, as evidenced by a 1.3-fold increase in the number of full-length long terminal repeat retrotransposons, an enrichment in specific transposable element-related Pfam domains and C subgenome specific transposon families, higher repetitiveness, more tandem repeats and higher numbers of transposable element and low-confidence genes (Extended Data Fig. 1c). Several tandem repeat subfamilies were unequally distributed across the subgenomes, highlighting potentially rearranged genomic regions (Extended Data Fig. 1d). However, limitations of the short-read assembly arising from lower contiguity (Supplementary Table 1) were apparent in the overall reduced representation of tandem repeats and ribosomal DNA loci (Supplementary Tables 4 and 6) as well as in regions of reduced gene density mainly in centromeric and pericentromeric regions and unplaced genes (Supplementary Fig. 9, Supplementary Table 7 and Supplementary Methods).

### Mosaic chromosome architecture of oat

The overall structure of the oat genome is similar to that of Triticeae genomes, although frequent genomic rearrangements in oat have resulted in a mosaic-like genome architecture. In many oat chromosomes, gene and recombination density is not a monotonic function of distance from the centromere (Extended Data Fig. 3), as is mostly observed in the Triticeae<sup>17</sup>. Examination of whole-genome alignments, subgenome-specific *k*-mers and orthologous and homoeologous genes clustering as syntenic blocks in genomic neighbourhoods in four *Avena* species (Extended Data Figs. 1d and 4) revealed numerous large-scale genomic rearrangements affecting the order of these blocks within and between subgenomes (Fig. 2a). We detected seven large-scale genomic rearrangements in *A. sativa* and traced them back to eight translocation events between the A, C and D subgenomes (Fig. 2b, c, Extended Data Figs. 4a and 5a, Supplementary Fig. 10 and Supplementary Table 8), spanning 4.3% of the genome and approximately 7.9% of the high-confidence genes. Two of the translocation events were specific to *A. sativa*. Unlike those in wheat<sup>7</sup>, the oat subgenomes exhibit unbalanced gene counts; specifically, the C subgenome appears to have 12% fewer genes than the A or D subgenome (Extended Data Fig. 2d and Supplementary Table 9). Analysis of orthologous gene groups (Supplementary Table 10 and Supplementary Figs. 11 and 12) showed that unbalanced gene families were associated with significant spatial clustering (Supplementary Fig. 13) in genomic rearrangements. Ancestral state reconstruction of the oat chromosomes revealed a loss of at least 226 Mb of gene-rich regions from the C subgenome to the A and D subgenomes (Supplementary Table 9). This implies that the translocations fully account for the lower gene count in the C subgenome and not gene loss or subfractionation after formation of the hexaploid.

Previous molecular marker studies using oat mapping and breeding populations have provided independent evidence for frequent translocations among oat subgenomes<sup>18,19</sup>. Using the oat genome to reanalyse the data (Extended Data Fig. 6), we observed inter-chromosomal pseudo-linkage in a population that segregates for the 1C translocation on 1A. Such pseudo-linkage has been implicated in the propensity for cold hardiness to remain associated with non-carriers of this translocation<sup>20</sup>. An accompanying study<sup>21</sup> details similar associated opportunities and barriers in genomic breeding strategies. The mosaic nature of the oat genome may be associated with the apparent lack of an orthologue of *TaZIP4-B2* (located within the *Ph1* locus), which in bread wheat stabilizes the genome structure during meiosis and suppresses crossovers between homoeologues<sup>22–24</sup> (Extended Data Fig. 5b and Supplementary Figs. 14 and 15). In contrast to wheat<sup>25</sup>, interploidy crosses and alien introgressions have been extremely challenging in *Avena*<sup>26</sup>, suggesting that incompatible genome architecture is an additional barrier preventing genetic gains in oat.



**Fig. 2 | Genome organization, rearrangement and subgenome interplay in oat.**

**a**, Overview of syntenic blocks across the three subgenomes<sup>52</sup>. **b**, Predicted borders of the seven major inter-subgenomic translocations in hexaploid oat *A. sativa* (*A. sat*) and its closest tetraploid ancestor, *A. insularis* (*A. ins*). Blue, gold and green colours represent the A-, C- and D-subgenomic regions, respectively. **c**, Reconstruction of translocations in *A. sativa* and *A. insularis* using subgenome-specific *k*-mers and syntenic blocks and orthoblocks. Left side, probabilities of the A-, C- and D-subgenome classification by *k*-mers for chromosome 1C of *A. insularis* and chromosomes 3A, 1C and 1A of *A. sativa*. The boxes above each plot show the order and identity of colour-coded blocks of the respective orthologous homoeologous genes according to the colour bar at the bottom left. Right side, illustration of two translocation events deduced from the information at left: translocation

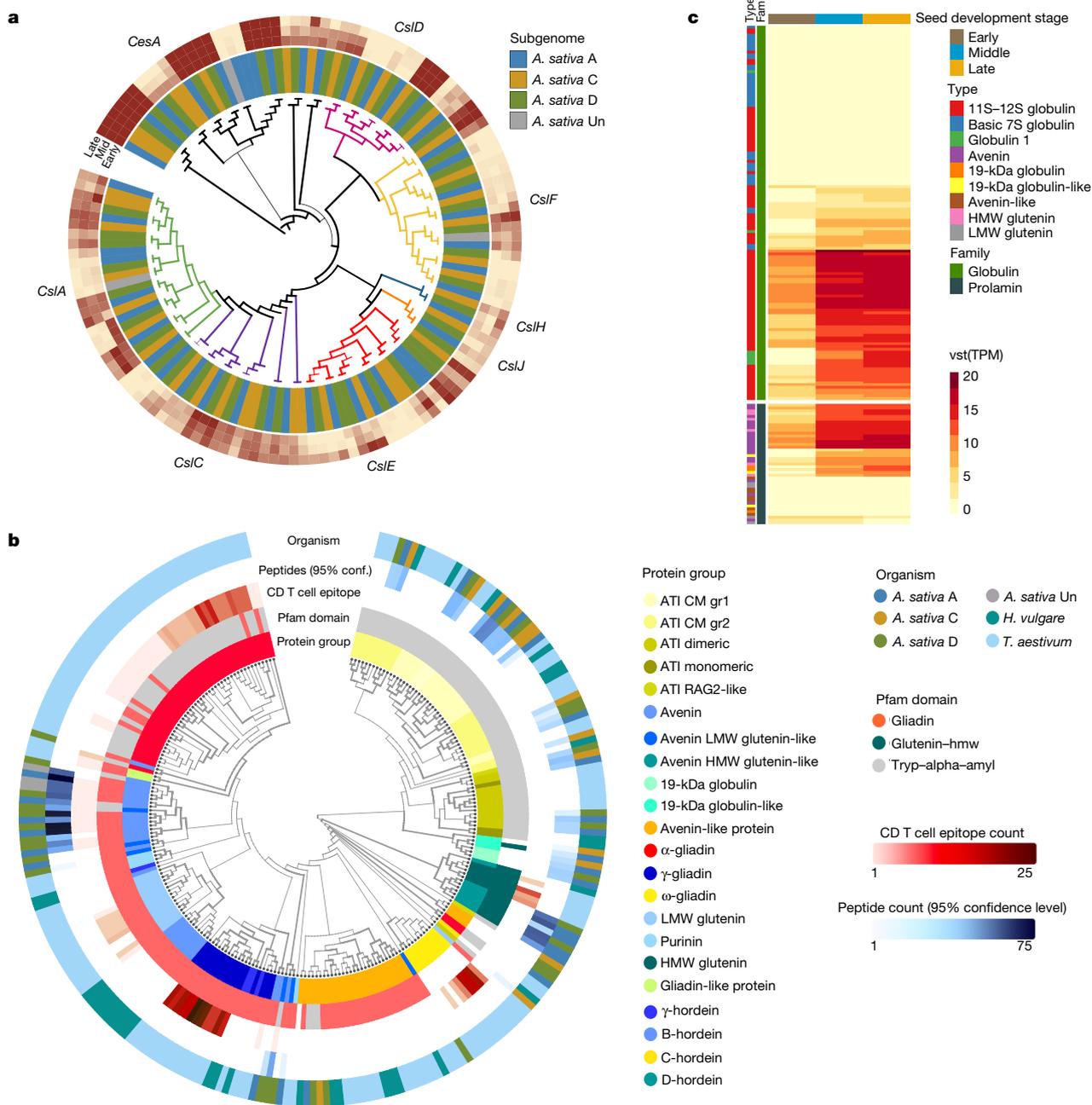
of blocks 48 and 63 from chromosome 3D to 1C, which is shared by *A. sativa* and *A. insularis* and occurred in their tetraploid ancestor, and the transfer of blocks 13 and 19 from chromosome 1C to 1A in *A. sativa*, resulting in a duplication pattern of these blocks. The ancestral location of blocks 48 and 63 on chr3 is supported by chr3A of *A. sativa* (top left); chr3C of *A. sativa*, *A. insularis* and *A. eriantha*; and chr3A of *A. longiglumis*. Bottom right, the 1C segment of chr1A in *A. sativa* is cytologically highlighted. Asterisks refer to the ancestral state of chromosomes. **d**, Ternary plot of the relative expression levels of 7,726 ancestral triads (23,178 genes) in hexaploid oat in a combined analysis of all transcriptome samples. Each dot represents a gene triad with an A, C and D coordinate. Subgenome-dominant categories are defined by triads in vertices, whereas suppressed categories are associated with triads near edges and between vertices. Grey dots in the centre indicate balanced triads.

### Oat subgenome expression is balanced

After polyploidization, sub- and neofunctionalization and gene loss modify the gene content in the new species<sup>27,28</sup>. Systematic differences in subgenome/homoeologue gene expression (homoeologue expression bias<sup>29</sup>) may also be prevalent. In fact, quantitative variation for many agronomic traits may reflect genetic interactions between homoeologues such as functional redundancy (buffering) or dominant phenotypes attributed to one homoeologue<sup>30</sup>. To investigate homoeologue expression bias in hexaploid oat, we defined 7,726 homoeologous gene triads with a 1:1:1 correspondence across the three oat subgenomes (Supplementary Table 11), referred to as ancestral triads. Average expression values across transcriptome samples from six tissues showed that C-subgenome genes were slightly less expressed (32.32%) than those in the D (33.53%) and A (33.76%) subgenomes (Kruskal–Wallis,  $P = 0.054$ ). We considered six homoeologous expression categories<sup>31</sup> and found that most ancestral triads (84.1%) showed balanced expression, 3.4% showed single-homoeologue

dominance and 12.6% showed single-homoeologue suppression. The relative contributions of the different categories (Extended Data Fig. 7a) indicated no major overall bias for one of the subgenomes (Fig. 2d). A co-expression network approach revealed that genes from the C subgenome were found in divergent expression modules more frequently than their A- and D-subgenome homoeologues ( $\chi^2$  test,  $P = 2.085 \times 10^{-6}$ ; Extended Data Fig. 7b and Supplementary Table 12).

In another 1,508 triad gene clusters containing at least one member positioned in a translocated region (relocated triads; Supplementary Table 13), the overall expression patterns were similar to those of the ancestral triads (Extended Data Fig. 7c). The C-suppressed category was slightly larger (5.1%) in the ancestral triads compared with the A-suppressed (3.5%) or D-suppressed (4.1%) triads, but the subgenome suppression patterns were reversed (4.5% A, 4.2% C and 5.2% D) in the relocated triads ( $\chi^2$  test,  $P = 0.019$ ; Extended Data Fig. 7c). Our results indicate that translocations and rearrangements in the oat genome may affect global and homoeologous gene expression patterns. Understanding how homoeologues interact



**Fig. 3 | Analysis of cellulose synthase and seed storage protein gene families in *A. sativa*.** **a**, Phylogeny of the cellulose synthase protein superfamily in *A. sativa* highlighting the eight subfamilies. Outer tracks represent the variance stabilizing-transformed transcripts per million (TPM) values determined for early, middle and late seed-development stages. The TPM level correlates with the intensity of burgundy colouring; the branch thickness corresponds to the bootstrap values and increases with higher bootstrap. **b**, Schematic representation and phylogeny of cereal storage proteins. The protein types used for the analysis were: wheat gliadins, glutenins, avenin-like proteins and ATIs; barley hordeins, avenin-like proteins

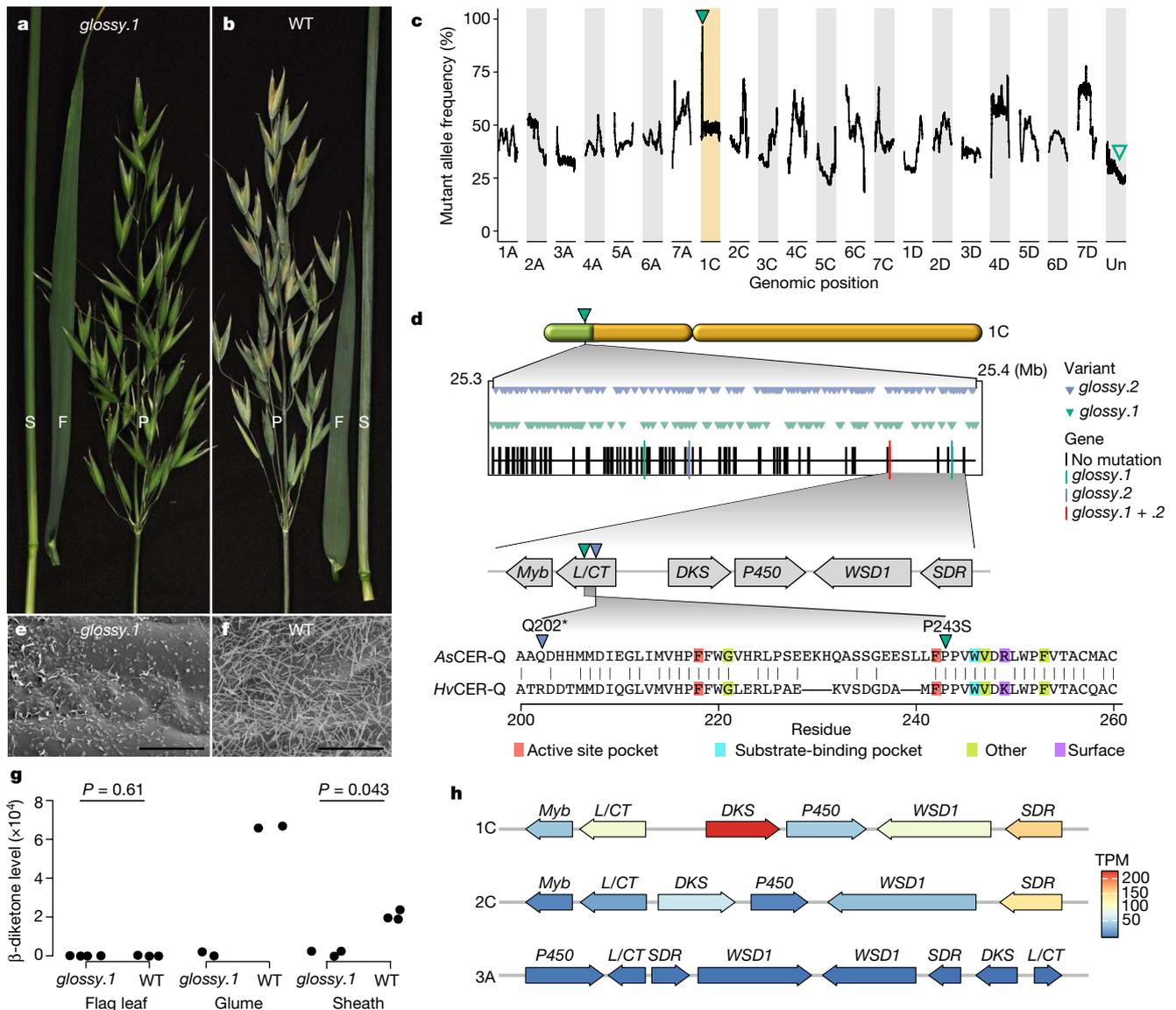
and ATIs<sup>53</sup>; and oat avenins, HMW glutenins, avenin-like proteins and ATIs identified in this study. Pfam domains and the identified protein groups are highlighted in separate layers. Epitopes used for the analysis included coeliac disease (CD)-associated T cell epitopes<sup>47</sup>. The numbers of T cell epitopes are labelled in the red colour scale. The number of peptides identified at the 95% confidence level are labelled in the blue colour scale; the branch thickness corresponds to bootstrap values and increases with higher bootstrap. LMW, low molecular weight. **c**, Expression of oat prolamin and globulin genes in three stages of seed development. The variance stabilizing transformed (vst) TPM levels correlate with the intensity of yellow to red colouring.

to influence gene expression and identifying functional single-copy genes showing non-balanced expression will inform crop improvement in oats.

### Soluble fibre-related gene families

Mixed-linkage  $\beta$ -glucans are soluble fibres present at high levels in oat endosperm cell walls (3.8–6.1 g per 100 g dry weight) that reduce

blood cholesterol and post-meal glycaemic responses<sup>3,4</sup>. The cellulose synthase-like gene *CslF6* is central for  $\beta$ -glucan biosynthesis in cereals<sup>32,33</sup>. We catalogued the cellulose synthase (*GT2*) and callose synthase (*GT48*) families of glycosyltransferases to identify the genetic foundation underlying oat  $\beta$ -glucan biosynthesis. The hexaploid oat genome encodes 134 members of the cellulose synthase gene superfamily (Fig. 3a), representing the cellulose synthase (*CesA*) subfamily (Supplementary Fig. 16) and



**Fig. 4 | Single-gene mapping of an epicuticular wax mutant.**

**a, b**, Epicuticular wax phenotypes of the *glossy.1* mutant (**a**) and the glaucous parental cultivar (**b**) at the early grain filling stage. S, sheath; F, flag leaf; P, panicle. **c**, Sliding window of allele frequency for variants unique to *glossy.1*. A window of 100 variants (total allelic depth  $\geq 30$ ) was used. Green triangle, chromosomal region shared by the *glossy.1* pool; empty triangle, location of the contig with the candidate gene in the assembly. Hi-C data anchor the contig to the 1C peak. **d**, Mapping of the *glossy.1* locus. Top, genes and variants (total allelic depth  $\geq 30$ ) at the 1C peak. Middle, the candidate gene (Gene-ID: *AVESA.00010b.r2.UnGI403470*) encoding a lipase/carboxyltransferase (L/CT), indicated by the red vertical bar, is located in the putative biosynthetic gene cluster orthologous to the barley *Cer-cqu* cluster. The genes encoding diketone synthase (DKS) and L/CT are orthologous to the barley *Cer-c* and *Cer-q* genes,

respectively (Supplementary Figs. 22 and 28). Green triangle, *glossy.1* mutation; blue triangle, *glossy.2* mutation. Bottom, alignment of *HvCER-Q* and *AsCER-Q*. Known deleterious single-amino acid substitutions from barley<sup>49</sup> are indicated. **e, f**, Scanning electron micrographs of the glume cuticle surface in *glossy.1* (**e**) and the glaucous parental cultivar (**f**) at  $\times 4,000$  magnification; scale bars, 10  $\mu\text{m}$ . **g**, Hentriacontane-14,16-dione is the major metabolite not detected in *glossy.1* (two-sided Welch *t*-test, *P* values adjusted using Benjamini-Hochberg procedure; glaucous flag leaf,  $n = 4$ ; *glossy* flag leaf,  $n = 3$ ; sheath,  $n = 3$ ; glume,  $n = 2$ ). **h**, Homoeologous gene clusters on chromosomes 1C, 3A and 2C. Genes are coloured according to the mean TPM value (four biological replicates) in glaucous glumes. *Myb*, Myb factor; *P450*, cytochrome P450; *WSD1*, wax ester synthase/diacylglycerol acyltransferase 1; *SDR*, short-chain dehydrogenase/reductase.

seven cellulose synthase-like subfamilies, including *Cs1A*, *Cs1C*, *Cs1D*, *Cs1E*, *Cs1F*, *Cs1H* and *Cs1J*. The *GT48* family comprised 28 members (Supplementary Fig. 17). Genes within the *CesA* and *Cs1F* subfamilies were most highly expressed over multiple stages of seed development (Fig. 3a and Supplementary Fig. 18). Investigation of differentially expressed genes between stages indicated specific roles for particular subfamilies such as *Cs1E* and *Cs1F* (including the C-subgenome copy of *Cs1F6*), which were upregulated in late stages of seed development (Supplementary Fig. 19), as shown in barley<sup>34</sup>. Compared with other grasses, the oat cellulose synthase superfamily showed no significant expansions apart from duplication events in the

*CesA*, *Cs1C*, *Cs1E* and *Cs1J* subfamilies (Supplementary Fig. 20 and Extended Data Fig. 8). These findings suggest that the high content and quality of  $\beta$ -glucan in oat are not driven by major differences in the copy number of cellulose synthase superfamily genes relative to other grasses but rather by allelic variation and transcription factors, as previously reported<sup>34</sup>.

## Oat storage proteins and human health

Oat globulins constitute 75–80% of grain protein content, with prolamins (avenins) accounting for approximately 10–15%. Prolamin

superfamily members trigger coeliac disease, food allergies and baker's asthma<sup>35</sup>. We identified genes encoding 25 avenins, 6 high-molecular-weight glutenins (HMW-GS) and 61 genes representing  $\alpha$ -amylase/trypsin inhibitors (ATIs) and other prolamin superfamily members related to protein accumulation and immunogenicity (Supplementary Table 14). Hexaploid oat has avenin loci on chromosomes 1D, 3D and 7A; seed storage globulin loci (135 genes) on chromosomes 1A, 1D, 3D, 7A, 4A and 4D; and no storage protein loci mapping to the C subgenome (Fig. 3b and Supplementary Table 14).

Unlike that of wheat, the oat genome harboured no  $\alpha$ - or  $\omega$ -gliadin genes, and the identified avenins co-clustered with  $\gamma$ -gliadins, low-molecular-weight glutenins and B-hordeins (Fig. 3b). We detected four complete, highly conserved oat HMW-GS gene models as two distinct loci on 1A and one locus pair on 1D, with no HMW-GS genes mapping to 1C. We identified a prolamin type, the 19-kDa globulin-like proteins, with an unknown function that is distinct from the avenins yet shares sequence similarity with HMW-GS and 19-kDa globulins (Fig. 3b). The predicted oat HMW-GS and avenins were highly conserved in their Pfam domains (Fig. 3b) and cysteine patterns (Extended Data Fig. 9). Glutamine- and proline-rich repetitive peptides were fewer in these oat proteins, making them shorter than those in wheat or barley (Extended Data Fig. 9).

We detected transcripts for most of the avenin genes, which showed gene expression patterns that aligned with their wheat orthologues, with increased transcript levels from the middle phase of seed development<sup>36</sup> (Fig. 3c), and protein levels by using liquid chromatography with tandem mass spectrometry (Fig. 3b). We identified inactive genes and pseudogenes among avenin-encoding genes (Fig. 3b, c and Supplementary Table 14) in a similar proportion as in wheat  $\gamma$ -gliadins<sup>37,38</sup>. This indicates a lower level of gene expansion and pseudogenization compared with the highly immunogenic wheat  $\alpha$ -gliadin genes<sup>39</sup>. Moreover, the expression of 11S globulin genes initiated early in seed development and was higher than that of the avenin genes (Fig. 3c). Discovery proteomics detected thirty-six distinct 11S globulins, five globulin-1 proteins and two 7S globulins, with an average of 83% protein sequence coverage at a 1% false discovery rate.

The oat avenins and globulins showed opposite trends compared with their wheat orthologues in gene copy number, protein length and enrichment in glutamine and asparagine residues that serve as a nitrogen storage sink (Extended Data Fig. 10a). Together with pronounced differences in transcription factor-binding sites specific to the nitrate response (Extended Data Fig. 10b and Supplementary Table 15), this may contribute to the primary role of oat globulins in nitrogen storage. These results confirm that the genomic organization, sequence characteristics and expression patterns of oat storage proteins share more similarities with rice and dicotyledonous plants than with wheat and other gluten-rich cereals<sup>40,41</sup>.

We mapped previously reported coeliac disease-associated T cell epitopes to the predicted oat avenin proteins and compared them with the T cell epitope patterns of wheat and barley prolamins<sup>42</sup>. The results showed that only a subset of encoded avenin proteins contain coeliac disease-associated immune-reactive regions compared with the high prevalence found in wheat or barley (Fig. 3b). Taken together, the low copy number of genes encoding coeliac disease epitopes, low frequency of detected T cell epitopes in the protein sequence, low occurrence of other highly immunogenic proteins, proportion of avenins within total oat protein and relative immunogenicity of avenin epitopes<sup>43</sup> all support the inclusion of oats in gluten-free diets<sup>35</sup>.

### Single-gene mapping of a wax mutant

To demonstrate how an annotated reference genome enables greater use of resources such as TILLING populations<sup>44,45</sup>, we mapped the causal mutation in the epicuticular wax mutant *glossy.1* (Fig. 4a, b). Epicuticular waxes have a role in biotic and abiotic stress resistance<sup>46,47</sup> and

provide an important target for oat breeding. We identified homozygous polymorphisms unique to the mutant, which mapped to chromosome 1C (Fig. 4c and Supplementary Fig. 21), and identified a single gene annotated as an  $\alpha/\beta$ -hydrolase (*AVESA.00010b.r2.UnG1403470*) as a likely candidate that is orthologous to barley *Cer-q* (*HORVU.MORE.X.r3.2HG0097460*) (Supplementary Fig. 22). An independent mutant line (*glossy.2*) exhibited the same glossy phenotype (Fig. 4d, Supplementary Fig. 23 and Supplementary Table 16). Barley *Cer-q* mutants<sup>48</sup> are deficient in the same  $\beta$ -diketone (hentriacontane-14,16-dione) and wax tubules that are absent in the *glossy* mutants (Fig. 4e–g and Supplementary Figs. 24–26). The scaffold containing the candidate gene was localized to the region of chromosome 1C (Fig. 4c and Supplementary Table 17). The presumed *glossy.1* mutation introduced a P243S substitution in the encoded protein adjacent to a deleterious F219L substitution known to inactivate barley CER-Q<sup>49</sup> (Fig. 4d and Supplementary Fig. 27). We identified gene clusters on oat chromosomes 1C, 2C and 3A and in wild *Avena* species (Supplementary Figs. 22 and 28–33) that are homologous to the barley *Cer-cqu* cluster<sup>49,50</sup>. We also noted genes encoding proteins with similarity to *Arabidopsis* wax ester synthase/diacylglycerol acyltransferase 1 (*WSD1*), a Myb-domain transcription factor and a short-chain dehydrogenase/reductase (*SDR*) protein near the *Cer-cqu* homologues in the *Avena* genomes. All genes from the 1C cluster except *SDR* were expressed at levels 3–6 times higher than those of the 3A cluster, with very low expression from 2C cluster genes and with no differential expression between the *glossy* and glaucous glume tissue (two-sided Wald test, null hypothesis logarithmic fold change = 0, adjusted  $P < 0.01$ ; Fig. 4h and Supplementary Fig. 34). Together, these results suggest that *AVESA.00010b.r2.UnG1403470* is the oat *Cer-q* gene. The reference genome thus facilitated a major advance in understanding  $\beta$ -diketone biosynthesis in oat and can help breeders manipulate tissue-specific epicuticular wax composition in future oat cultivars adapted for hotter climates.

### Discussion

In summary, this fully annotated hexaploid oat reference genome lays the foundation for advances in oat breeding and basic oat biology and for the ongoing pan-genome project. With the chromosome rearrangements in a typical spring oat cultivar now delineated, breeders and researchers will have access to a resource equal in calibre to Triticeae genomes, which may help them to overcome the breeding barriers and segregation anomalies described in numerous mapping studies. Using the reference genome to map genes associated with agronomic and human nutrition-related traits is a viable approach for precisely adapting oat varieties. Known quantitative trait loci can be anchored to the Sang reference, and the transcriptome atlas co-expression networks can be leveraged to identify candidate genes in the vicinity of specific quantitative trait loci. Modern breeding strategies such as genome editing and gene pyramiding can now more easily be applied in oat to develop varieties that meet the increasing global demand for oat-derived products. Our proteogenomic investigation of oat storage proteins confirms qualitative and quantitative differences in the expression of proteins compared with the more abundant and immunogenic sequences in wheat, barley and rye, which supports the safety of oats in gluten-free diets. The detailed genome annotation and case studies presented here provide examples of the myriad possibilities for the discovery and exploitation of functional genetic mechanisms in oat.

### Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-022-04732-y>.

1. Vavilov, N. I. in *Origin and Geography of Cultivated Plants* (ed. Dorofeyev, V. F.) 22–135 (Cambridge Univ. Press, 1992).
2. Malzew, A. I. *Wild and Cultivated Oats, Sectio Euavena Griseb.* (Publ. of the All-Union Inst. of Appl. Botany and New Cultures under the Council of People's Commissars of the USSR, 1930).
3. EFSA Panel on Dietetic Products, Nutrition and Allergies (NDA). Scientific opinion on the substantiation of a health claim related to oat  $\beta$ -glucan and lowering blood cholesterol and reduced risk of (coronary) heart disease pursuant to Article 14 of Regulation (EC) no. 1924/2006. *EFSA J.* <https://doi.org/10.2903/j.efsa.2010.1885> (2010).
4. EFSA Panel on Dietetic Products, Nutrition and Allergies (NDA). Scientific opinion on the substantiation of health claims related to  $\beta$ -glucans from oats and barley and maintenance of normal blood LDL-cholesterol concentrations (ID 1236, 1299), increase in satiety leading to a reduction in energy intake (ID 851, 852), reduction of post-prandial glycaemic responses (ID 821, 824), and 'digestive function' (ID 850) pursuant to Article 13(1) of Regulation (EC) no. 1924/2006. *EFSA J.* <https://doi.org/10.2903/j.efsa.2011.2207> (2011).
5. Mathews, R., Kamil, A. & Chu, Y. Global review of heart health claims for oat  $\beta$ -glucan products. *Nutr. Rev.* **78**, 78–97 (2020).
6. Manni, M., Berkeley, M. R., Seppely, M. & Zdobnov, E. M. BUSCO: assessing genomic data quality and beyond. *Curr. Protoc.* **1**, e323 (2021).
7. International Wheat Genome Sequencing Consortium (IWGSC). et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191 (2018).
8. Monat, C. et al. TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. *Genome Biol.* **20**, 284 (2019).
9. Rabanus-Wallace, M. T. et al. Chromosome-scale genome assembly provides insights into rye biology, evolution and agronomic potential. *Nat. Genet.* **53**, 564–573 (2021).
10. Bekele, W. A., Wight, C. P., Chao, S., Howarth, C. J. & Tinker, N. A. Haplotype-based genotyping-by-sequencing in oat genome research. *Plant Biotechnol. J.* **16**, 1452–1463 (2018).
11. GrainGenes: a database for Triticeae and *Avena*, *Avena sativa*, OT3098 v2, PepsiCo, <https://wheat.pw.usda.gov/jb?data=ggds/oat-ot3098v2-pepsico> (accessed 15 January 2022).
12. Ladizinsky, G. A new species of *Avena* from Sicily, possibly the tetraploid progenitor of hexaploid oats. *Genet. Resour. Crop Evol.* **45**, 263–269 (1998).
13. Sanz, M. J. et al. A new chromosome nomenclature system for oat (*Avena sativa* L. and *A. byzantina* C. Koch) based on FISH analysis of monosomic lines. *Theor. Appl. Genet.* **121**, 1541–1552 (2010).
14. Maughan, P. J. et al. Genomic insights from the first chromosome-scale assemblies of oat (*Avena* spp.) diploid species. *BMC Biol.* **17**, 92 (2019).
15. Nomenclature Database, National Oat Conference Newsletter, <https://oatnews.org/nomenclature> (accessed 8 March 2021).
16. Mascher, M. et al. Long-read sequence assembly: a technical evaluation in barley. *Plant Cell* **33**, 1888–1906 (2021).
17. Mascher, M. et al. A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**, 427–433 (2017).
18. Chaffin, A. S. et al. A consensus map in cultivated hexaploid oat reveals conserved grass synteny with substantial subgenome rearrangement. *Plant Genome* <https://doi.org/10.3835/plantgenome2015.10.0102> (2016).
19. Yan, H. et al. High-density marker profiling confirms ancestral genomes of *Avena* species and identifies D-genome chromosomes of hexaploid oat. *Theor. Appl. Genet.* **129**, 2133–2149 (2016).
20. Santos, A. G., Livingston, D. P., Jellen, E. N., Wooten, D. R. & Murphy, J. P. A cytological marker associated with winterhardness in oat. *Crop Sci.* **46**, 203–208 (2006).
21. Tinker, N. A. et al. Genome analysis in *Avena sativa* reveals hidden breeding barriers and opportunities for oat improvement. *Commun. Biol.* <https://doi.org/10.1038/s42003-022-03256-5> (2022).
22. Alabdullah, A. K., Moore, G. & Martin, A. C. A duplicated copy of the meiotic gene *ZIP4* preserves up to 50% pollen viability and grain number in polyploid wheat. *Biology* **10**, 290 (2021).
23. Griffiths, S. et al. Molecular characterization of *Ph1* as a major chromosome pairing locus in polyploid wheat. *Nature* **439**, 749–752 (2006).
24. Rey, M.-D. et al. Exploiting the *ZIP4* homologue within the wheat *Ph1* locus has identified two lines exhibiting homoeologous crossover in wheat-wild relative hybrids. *Mol. Breed.* **37**, 95 (2017).
25. Roskal, K., Kowalik, S., Podyma, W., Łapiński, B. & Boczkowska, M. The pros and cons of rye chromatin introgression into wheat genome. *Agronomy* **11**, 456 (2021).
26. Dilkova, M., Jellen, E. N. & Forsberg, R. A. C-banded karyotypes and meiotic abnormalities in germplasm derived from interploidy crosses in *Avena*. *Euphytica* **111**, 175–184 (2000).
27. Otto, S. P. The evolutionary consequences of polyploidy. *Cell* **131**, 452–462 (2007).
28. Jiao, Y. & Paterson, A. H. Polyploidy-associated genome modifications during land plant evolution. *Philos. Trans. R. Soc. Lond. B* **369**, 20130355 (2014).
29. Grover, C. E. et al. Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytol.* **196**, 966–971 (2012).
30. van de Peer, Y., Mizrahi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).
31. Ramirez-González, R. H. et al. The transcriptional landscape of polyploid wheat. *Science* **361**, eaar6089 (2018).
32. Garcia-Gimenez, G. et al. Targeted mutation of barley (1,3;1,4)- $\beta$ -glucan synthases reveals complex relationships between the storage and cell wall polysaccharide content. *Plant J.* **104**, 1009–1022 (2020).
33. Nemeth, C. et al. Down-regulation of the *CSLF6* gene results in decreased (1,3;1,4)- $\beta$ -D-glucan in endosperm of wheat. *Plant Physiol.* **152**, 1209–1218 (2010).
34. Wong, S. C. et al. Differential expression of the *HvCSLF6* gene late in grain development may explain quantitative differences in (1,3;1,4)- $\beta$ -glucan concentration in barley. *Mol. Breed.* **35**, 20 (2015).
35. Pinto-Sánchez, M. I. et al. Safety of adding oats to a gluten-free diet for patients with celiac disease: systematic review and meta-analysis of clinical and observational studies. *Gastroenterology* **153**, 395–409 (2017).
36. Vinje, M. A., Walling, J. G., Henson, C. A. & Duke, S. H. Comparative gene expression analysis of the  $\beta$ -amylase and hordein gene families in the developing barley grain. *Gene* **693**, 127–136 (2019).
37. Juhász, A. et al. Genome mapping of seed-borne allergens and immunoresponsive proteins in wheat. *Sci. Adv.* **4**, eaar8602 (2018).
38. Zhang, Y. et al. Genome-, transcriptome- and proteome-wide analyses of the gliadin gene families in *Triticum urartu*. *PLoS ONE* **10**, e0131559 (2015).
39. Huo, N. et al. Dynamic evolution of  $\alpha$ -gliadin prolamin gene family in homeologous genomes of hexaploid wheat. *Sci. Rep.* **8**, 5181 (2018).
40. Shewry, P. R. & Halford, N. G. Cereal seed storage proteins: structures, properties and role in grain utilization. *J. Exp. Bot.* **53**, 947–958 (2002).
41. Black, M. & Derek Bewley, J. *Seed Technology and Its Biological Basis* (CRC Press, 2000).
42. Sollid, L. M. et al. Update 2020: nomenclature and listing of celiac disease-relevant gluten epitopes recognized by CD4<sup>+</sup> T cells. *Immunogenetics* **72**, 85–88 (2020).
43. Hardy, M. Y. et al. Ingestion of oats and barley in patients with celiac disease mobilizes cross-reactive T cells activated by avenin peptides and immuno-dominant hordein peptides. *J. Autoimmun.* **56**, 56–65 (2015).
44. Adamski, N. M. et al. A roadmap for gene functional characterisation in crops with large genomes: lessons from polyploid wheat. *eLife* **9**, e55646 (2020).
45. Chawade, A. et al. Development and characterization of an oat TILLING-population and identification of mutations in lignin and  $\beta$ -glucan biosynthesis genes. *BMC Plant Biol.* **10**, 86 (2010).
46. von Wettstein-Knowles, P. Ecophysiology with barley *eceriferum* (*cer*) mutants: the effects of humidity and wax crystal structure on yield and vegetative parameters. *Ann. Bot.* **126**, 301–313 (2020).
47. Wang, X., Kong, L., Zhi, P. & Chang, C. Update on cuticular wax biosynthesis and its roles in plant disease resistance. *Int. J. Mol. Sci.* **21**, 5514 (2020).
48. von Wettstein-Knowles, P. The polyketide components of waxes and the *Cer-cqu* gene cluster encoding a novel polyketide synthase, the  $\beta$ -diketone synthase, DKS. *Plants* **6**, 28 (2017).
49. Schneider, L. M. et al. The *Cer-cqu* gene cluster determines three key players in a  $\beta$ -diketone synthase polyketide pathway synthesizing aliphatics in epicuticular waxes. *J. Exp. Bot.* **67**, 2715–2730 (2016).
50. Hen-Avivi, S. et al. A metabolic gene cluster in the wheat *W1* and the barley *Cer-cqu* loci determines  $\beta$ -diketone biosynthesis and glaucousness. *Plant Cell* **28**, 1440–1460 (2016).
51. Tinker, N. A. et al. A SNP genotyping array for hexaploid oat. *Plant Genome* **7**, 1–8 (2014).
52. Bandi, V. & Gutwin, C. Interactive exploration of genomic conservation. In *Proceedings of the 46th Graphics Interface Conference 2020 (GI'20)* <https://synvisio.github.io> (2020).
53. Mascher, M. et al. Pseudomolecules and annotation of the third version of the reference genome sequence assembly of barley cv. Morex [Morex V3]. *eDAL Plant Genomics and Phenomics Research Data Repository (PGP)* <https://doi.org/10.5447/ipk/2021/3> (accessed 3 December 2020).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022

## Reporting summary

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

## Data availability

The raw sequence data used for de novo whole-genome assembly are available from the European Nucleotide Archive (ENA) under accession number PRJEB44810 (*A. sativa* cv. Sang) and from the Sequence Read Archive under accession numbers PRJNA727490 (*A. insularis* BYU209) and PRJNA726919 (*A. longiglumis* CN58138). Chromosome conformation capture (Hi-C) sequencing data are available from ENA under accession numbers PRJEB43668 (*A. sativa* cv. Sang), PRJEB43670 (*A. insularis* BYU209) and PRJEB43669 (*A. longiglumis* CN58138). Chromosome-scale sequence assemblies (pseudomolecules) are available from ENA under accession numbers PRJEB44810 (*A. sativa* cv. Sang), PRJEB45088 (*A. insularis* BYU209) and PRJEB45087 (*A. longiglumis* CN58138). The raw RNA-seq and genome-sequencing data generated in this study are available under ENA accession number PRJEB46365. Pseudomolecules, annotation data and analysis results are available at the Plant Genomics and Phenomics Research Data Repository at <https://doi.org/10.5447/ipk/2022/2>. The DOI was registered using eDAL (<https://edal.ipk-gatersleben.de/>). Pseudomolecules, annotation data and associated analyses for *A. sativa* cv. Sang, *A. longiglumis* and *A. insularis* are also available from GrainGenes<sup>54</sup>. Sang genome browser, <https://wheat.pw.usda.gov/jb/?data=/ggds/oat-sang>; Sang data download, <https://wheat.pw.usda.gov/GG3/content/avena-sang-download>; *A. longiglumis* genome browser, <https://wheat.pw.usda.gov/jb/?data=/ggds/oat-longiglumis>; *A. longiglumis* data download, <https://wheat.pw.usda.gov/GG3/content/avena-longiglumis-download>; *A. insularis* genome browser, <https://wheat.pw.usda.gov/jb/?data=/ggds/oat-insularis>; *A. insularis* data download, <https://wheat.pw.usda.gov/GG3/content/avena-insularis-download>. The mass spectrometry proteomics data and ProteinPilot search result files have been deposited to MassIVE (<https://massive.ucsd.edu>) under accession number MSV000088727. The publicly available OT3098 oat genome data were generated by PepsiCo and Corteva Agriscience. This dataset (annotation version 2) has been obtained and is available from GrainGenes at <https://wheat.pw.usda.gov/GG3/content/pepsico-ot3098-hexaploid-oat-version-2-genome-assembly-release-collaboration-graingenes>. Databases used in this study included PTREP release 19, Uniref download 2019-09-03, Pfam download 2019-09-03, Swiss-Prot, TAIR, TrEMBL, REdat\_9.9\_Poaceae section of the PGSB transposon library, Immune Epitope Database and Analysis Resource (<https://www.iedb.org>), PLACE and PlantCare promoter motif databases and pfam2GO.

## Code availability

All relevant code developed in this study was deposited in the GitHub node at <https://github.com/PGSB-HMGU/oatkmers>.

54. Blake, V. C. et al. GrainGenes: centralized small grain resources and digital platform for geneticists and breeders. *Database* **2019**, baz065 (2019).

**Acknowledgements** We are grateful for funding from Stiftelsen för Strategisk Forskning (SSF) (ScanOats: IRC15-0068) and the members of the ScanOats Industrial Research Center: Lund University; the Research Institute of Sweden (RISE); the Swedish University of Agricultural Sciences (SLU, Alnarp); Lantmännen ek för; Oatly AB; and Swedish Oat Fiber AB. We thank the Lantmännen Research Foundation for supporting the raw data collection for the Sang reference genome and mutated oat lines (2016H033, 2018F005). The pre-sequencing phase of the Sang genome project was supported by the Swedish Farmers Supply and Crop Marketing Co-operative (grant no. O-14-20-177). Funding was also provided by the German Federal Ministry of Education and Research (De.NBI, FKZ 031A536), German Federal Ministry of Food and Agriculture (FUGE, FKZ 28AIN02D20), Agriculture and Agri-Food Canada and Genome Quebec (Targeted and Useful Genomics for Barley and Oat), Australian Research Council Centre of Excellence for Innovations in Peptide and Protein Science (CE200100012), National Health and Medical Research Council (investigator grant APP1176553) and Coeliac Australia (G1005443). The authors acknowledge support from the National Genomics Infrastructure funded by the Science for Life Laboratory, the Knut and Alice Wallenberg Foundation and the Swedish Research Council. The SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science provided assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure. We acknowledge V. Llaca and K. Fengler for sequencing and assembly of OT3098 v2 and D. Eickholt for OT3098 project support. We thank D. Ahren for supervision of N.T.R., and bioinformatics guidance to J.B. and N. Sirijovski. We thank L. Sjögren for isolating Sang genomic DNA for shotgun and massively parallel libraries, O. Van Aken and E. Darwish for contributing RNA-seq data used in the atlas, and A. Hansson for support during the initial planning phase of the Sang project. We are grateful to M. Knauff and I. Walde for assistance with Hi-C library preparation and sequencing; A. Fiebig for support in sequence data submission; E. Wilcox for support and expertise with PacBio sequencing; the staff at the Centre d'expertise et de services (Genome Quebec); A. Itaya and C. Wight for biological materials and sample preparation; S. Stockwell for proteomics sample preparation; O. Gustafsson for assistance with scanning electron microscopy; and R. Öste, K. E. Andersson and L. Bülow for major contributions to the SFF ScanOats grant application. We thank the GrainGenes database (<https://wheat.pw.usda.gov>) for providing a long-term home for the datasets, genome browsers and BLAST functionality for the oat genomes. The opinions in this study are those of the authors and do not necessarily represent the opinions or policies of PepsiCo, Inc.

**Author contributions** O.O. conceived the Sang genome study and secured funding. N. Sirijovski and M.S. conceived and coordinated the study. N. Sirijovski performed molecular biology and coordinated sequencing and assembly for Sang, N.A.T., W.A.B., E.N.J., P.J.M. and Y.-B.F. contributed *A. insularis* and *A. longiglumis* assemblies. A.H. and N. Stein prepared Hi-C libraries and performed sequencing. M.M. performed pseudomolecule assemblies. M.M. and H.G. determined chromosome nomenclature. N. Sirijovski prepared RNA for atlas samples presented in this work. A.C. contributed sequence data: RNA-seq, hexaploid markers and pedigree background. A.J.W. contributed Iso-seq data and was principal investigator of the OT3098 genome project. M.S. and K.F.X.M. supervised annotation, with assistance from N.K., N.T.R. and T.L. N.T.R. performed confidence classification. TE and repeat analyses were performed by H.G. Ontology annotations were provided by D.L., N.V.G. and R.R. J.B. was responsible for data management and submission, with assistance from T.L. N. Sirijovski performed experiments with mutants with support from M.H. and P.S. for profiling of wax metabolites. N.T.R. performed variant calling and single-gene mapping. N.K. conducted expression analysis with assistance from J.B. N.A.T. and W.A.B. performed molecular marker and recombination landscape analysis. E.N.J. contributed cytogenetic data. N.K., N.T.R. and A.J. performed gene family analysis and manual annotations. G.H. and H.G. analysed genome structure, organization and translocations. A.J. and M.L.C. supervised proteomics: U.B. performed proteomic data acquisition; A.J., U.B. and M.L.C. conducted proteomic data analysis; A.J. performed storage protein analysis; J.A.T.-D. performed storage protein data interpretation. M.S. and N. Sirijovski drafted the manuscript with input from N.K., N.T.R., J.B., M.M., A.J., U.B., J.A.T.-D., M.L.C., H.G., G.H. and N.A.T. All co-authors contributed to and edited the final version.

**Funding** Open access funding provided by Lund University.

**Competing interests** N. Sirijovski was an employee of CropTailor AB, a commercial enterprise, and became employed by Oatly AB during manuscript revision. O.O. is an employee of CropTailor AB. A.C. is an employee at Lantmännen. A.J.W. is an employee of PepsiCo, Inc. These authors are not expected to benefit financially from publication of these results. All other authors declare no competing interests.

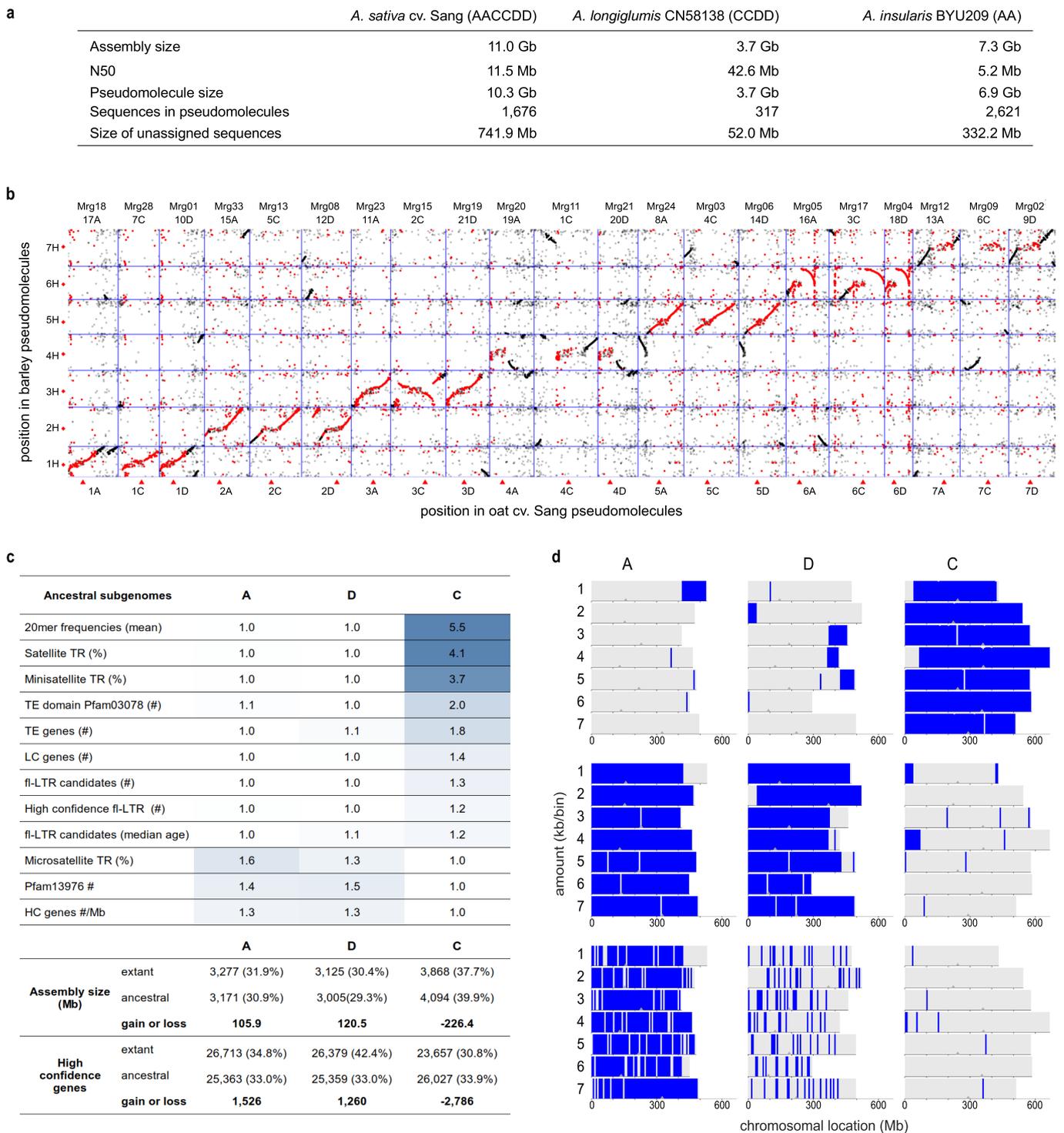
## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41586-022-04732-y>.

**Correspondence and requests for materials** should be addressed to Manuel Spannagl or Nick Sirijovski.

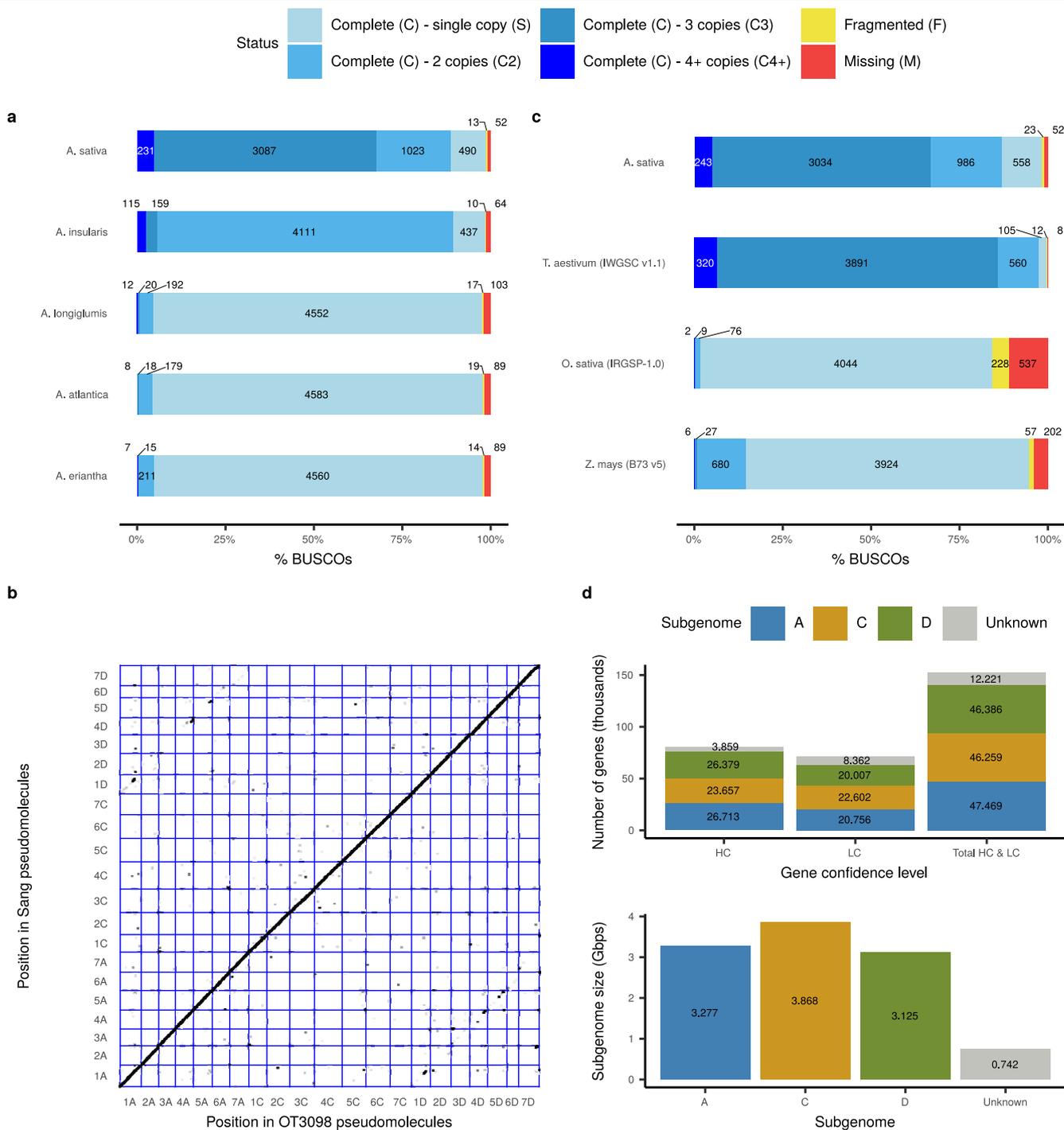
**Peer review information** Nature thanks John Lovell for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



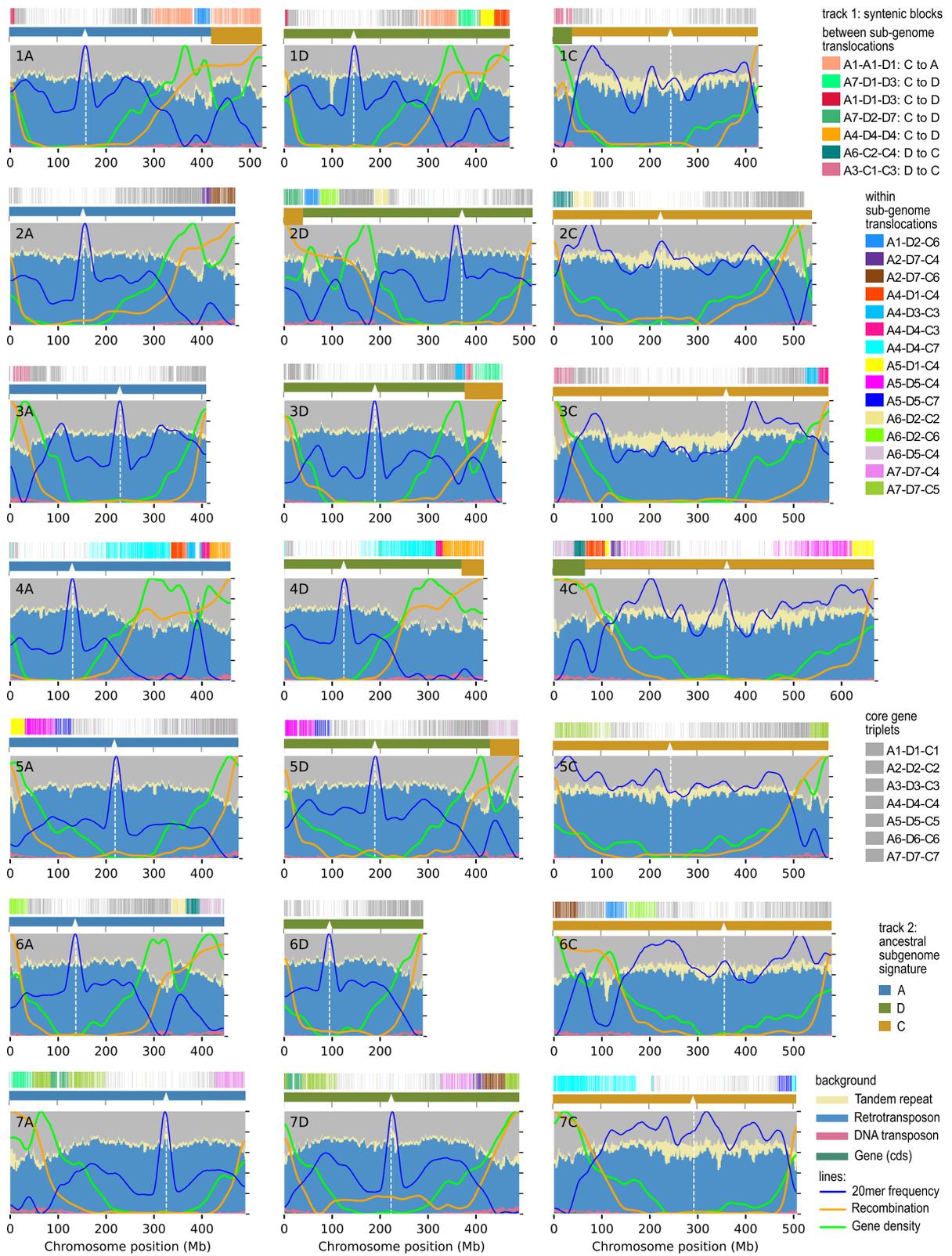
**Extended Data Fig. 1 | Assembly of three *Avena* species and chromosome organisation of hexaploid oat.** **a**, Summary of assembly statistics of the hexaploid *A. sativa* cv. Sang, diploid *A. longiglumis* CN58138 and tetraploid *A. insularis* BYU209 genomes. **b**, Gene-based collinearity of oat (*A. sativa*) to barley (*H. vulgare*). Each data point is an aligned gene. Genes in core regions are shown in red. The phylogenetically informed nomenclature is used on the bottom x-axis. The top axis shows the chromosome names according to Bekele et al. (2018)<sup>10</sup>, (Mrg groups) and according to Sanz et al. (2010)<sup>13</sup>, respectively. Centromere positions in oat and barley are indicated by red triangles (x axis)

and red diamonds (y axis). **c**, Subgenome composition of hexaploid oat cv. Sang. Subgenome-specific features related to independent transposon histories in the diploid ancestors. The upper part shows the x-fold increase compared to the lowest of the three values across subgenomes A–C. The lower part summarises differences in genome size and gene number between the extant hexaploid and the ancestral state predating the seven translocations between the tetraploid and hexaploid subgenomes. **d**, Tandem repeat families in hexaploid oat cv. Sang specific to either the C or the A/D diploid ancestor or enriched in the extant A subgenome (from top to bottom).



**Extended Data Fig. 2 | Validation and statistics of *A. sativa* cv. Sang assembly and gene annotation.** **a**, BUSCO (lineage poales\_odb10, created 2020-08-05) scores of the genome assemblies of *A. sativa* cv. Sang, *A. insularis* and *A. longiglumis* as well as the diploid progenitors *A. eriantha* and *A. atlantica* as a comparison. **b**, Colinearity plot of the pseudomolecules of *A. sativa* cv. Sang (short read assembly, y-axis) and OT3098 (long read assembly, x-axis). **c**, BUSCO (lineage poales\_odb10, created 2020-08-05) scores for the oat

reference gene annotation v1.1 in comparison to the results of recent gene predictions for maize, rice and bread wheat. **d**, *Top*: overview of the number of predicted genes (HC - high confidence; LC - low confidence) for the three oat subgenomes and unplaced/unknown scaffolds. *Bottom*: total number of basepairs assembled for the respective three subgenomes and the number of basepairs in unplaced scaffolds.



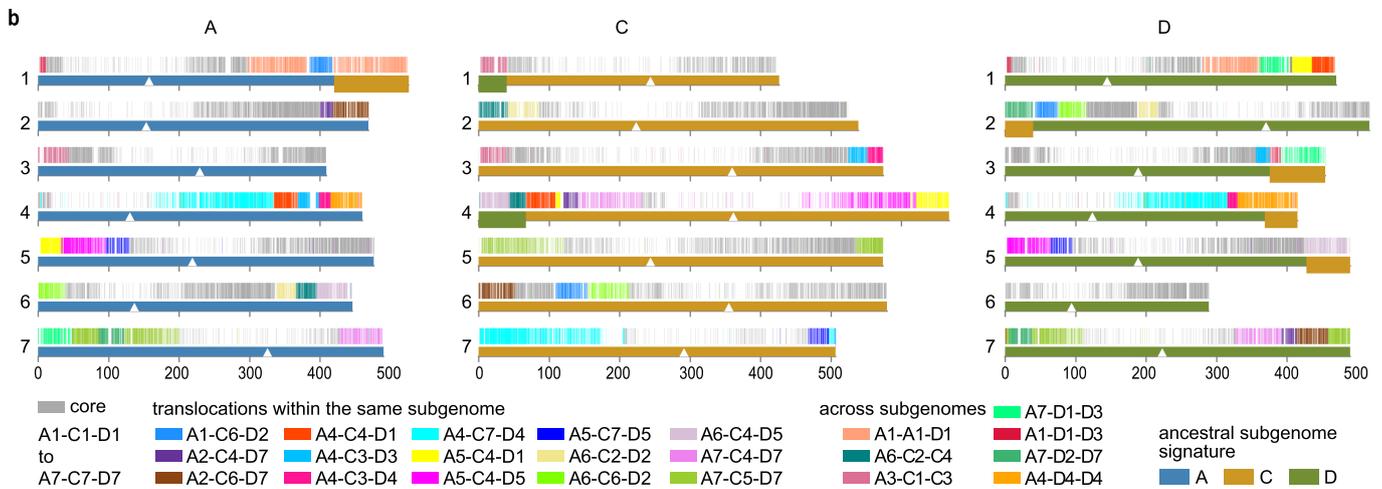
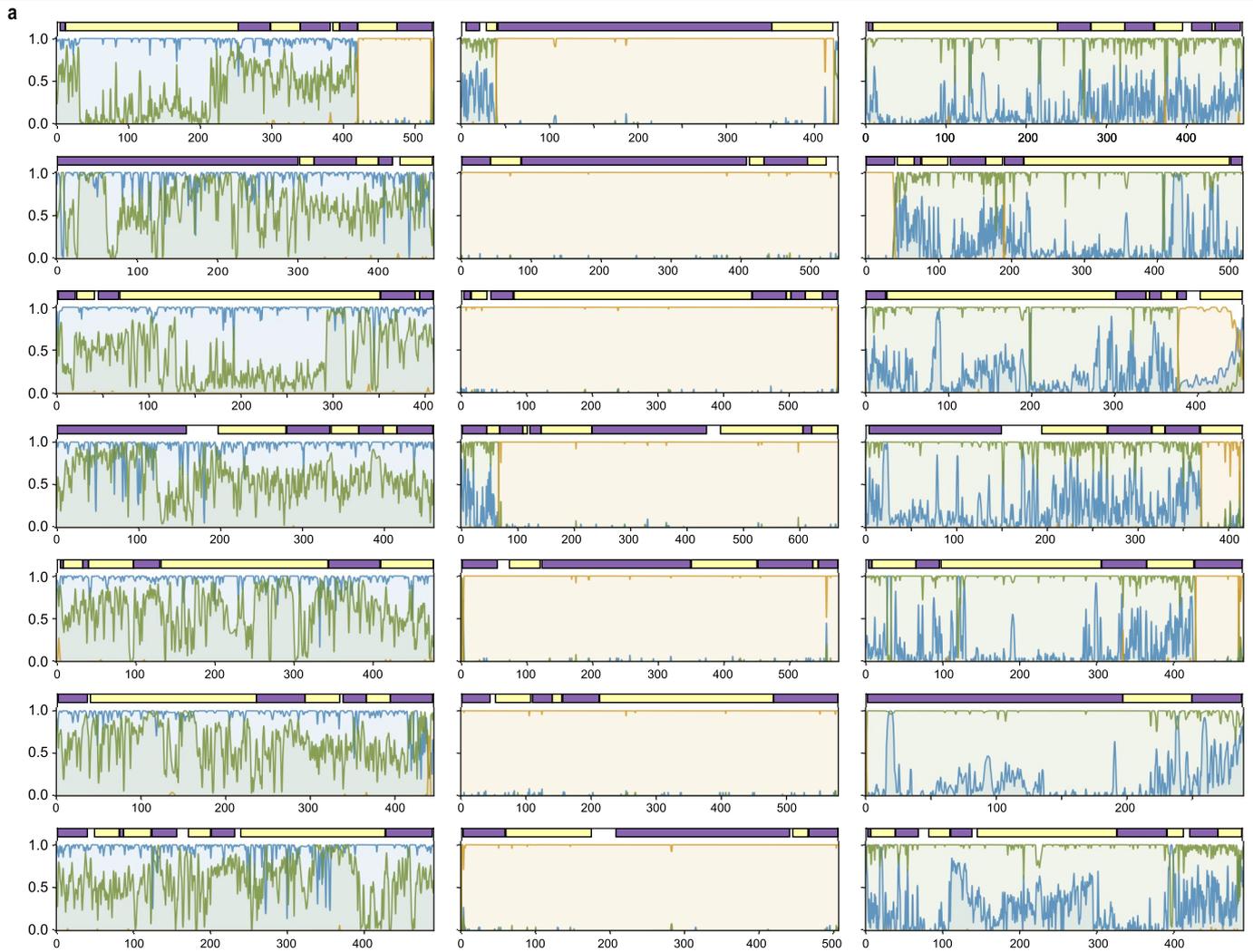
**Extended Data Fig. 3** | See next page for caption.

# Article

## Extended Data Fig. 3 | Chromosome architecture of hexaploid oat.

Distribution of main features along the 21 oat chromosomes. Track 1 displays syntenic blocks based on the chromosome pattern of gene clusters with exactly three members. Track 2 denotes the sub genome assignment. The background of the bottom part consists of a stacked bar chart for the major genome components. The lines in the front show smoothed chromosomal distributions for mean 20-mer frequency, average recombination rate (cM/Mb) and gene density (#/Mb), each scaled min to max per chromosome. The population-level analysis of crossover frequency is explored in (N.T., W.B. et al. <sup>21</sup>). Therein we focus on average recombination rates that may result from global

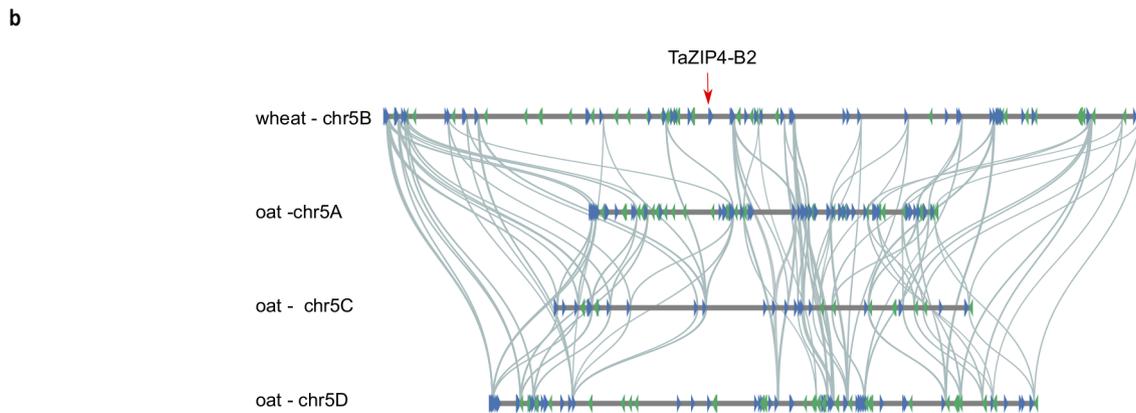
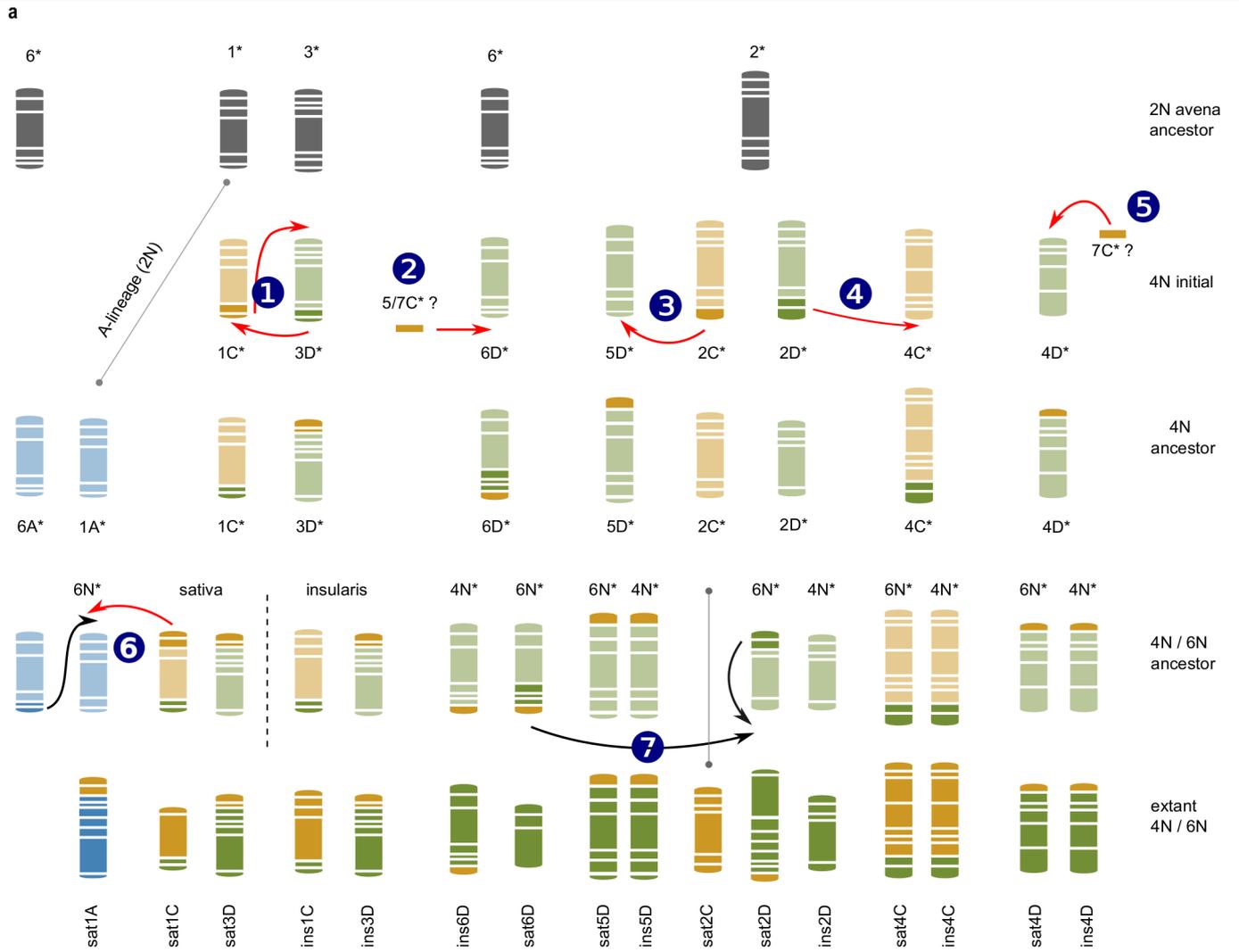
chromosome restructuring within *Avena*. Of particular interest are the long stretches of increased recombination on the long arms of chromosomes 4A and 4D. These regions coincide with stretches of high gene density and reduced transposon frequency. Other regions containing multiple ancestral telomeric regions showed corresponding multiple peaks of gene density and recombination. Also of interest is a region of suppressed recombination on chromosome 7D which is speculated to result from intraspecific rearrangements on this chromosome that impede meiotic pairing and crossovers.



**Extended Data Fig. 4 | Mosaic composition of hexaploid oat chromosomes.**

**a**, Subgenome-specific kmers and syntenic blocks in *A. sativa*. Probabilities of A- (blue), C- (yellow) and D- (green) subgenome classification by subgenomic-kmers are shown. Top row in each subplot displays identified synteny shown in alternative colours to emphasise block borders. From top to bottom, chromosomes 1 to 7 for subgenomes A (left column), C (mid column) and D (right column), Y-axis: kmer-probability, x-axis: chromosomal position in Mb. **b**, Ancestral subgenome origin. In the extant Sang genome seven peripheral regions ranging from 40 Mb to 106 Mb differ in their subgenome signature from their current location. Five of these regions are of C genome origin: chr1A

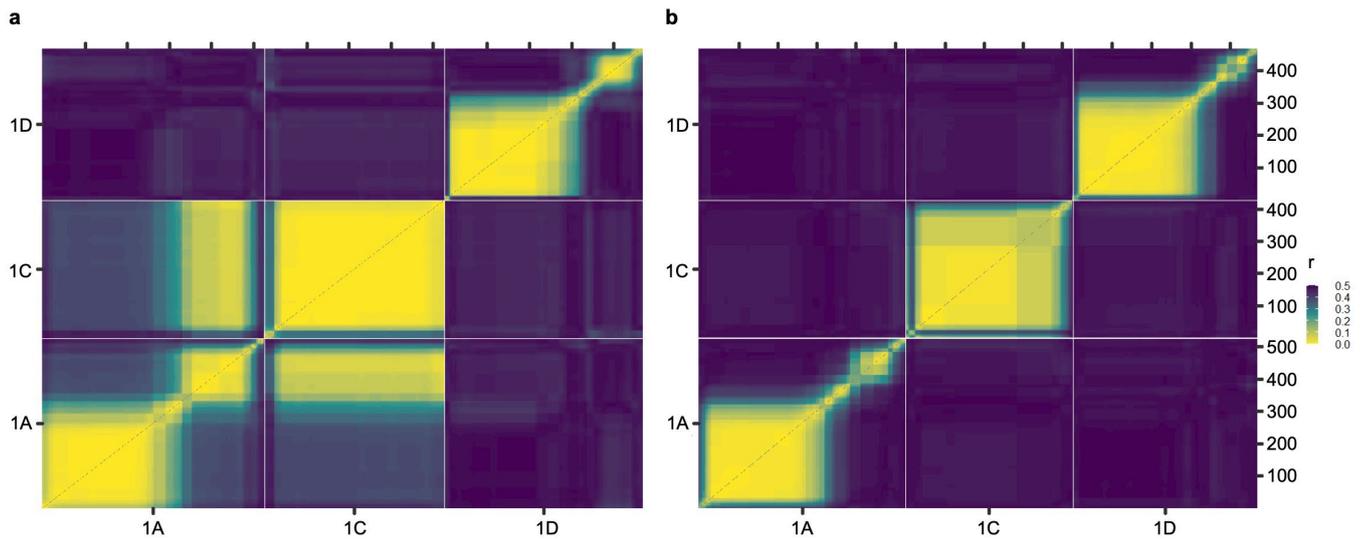
(106 Mb), chr2D (40Mb), chr3D (79 Mb), chr4D (46 Mb), chr5D (62 Mb). Two regions on the extant C subgenome are of D genome origin: chr1C (40 Mb) and chr4C (67 Mb). All in all, the transfers between subgenomes add up to 441 Mb, with a net loss in the C subgenome of 226 Mb. The upper tracks display homoeologous syntenic blocks within the Sang genome based on the chromosome pattern of gene clusters with exactly three members. The pattern A1-A1-D1 corresponds for example to a cluster with two members from chr1A and one member from chr1D. The core denotes all seven symmetric cluster patterns from A1-D1-C1 up to A7\_D7\_C7 with one member on each homoeolog chromosome (equivalent to the red dots in Extended Data Fig. 1b).



Extended Data Fig. 5 | See next page for caption.

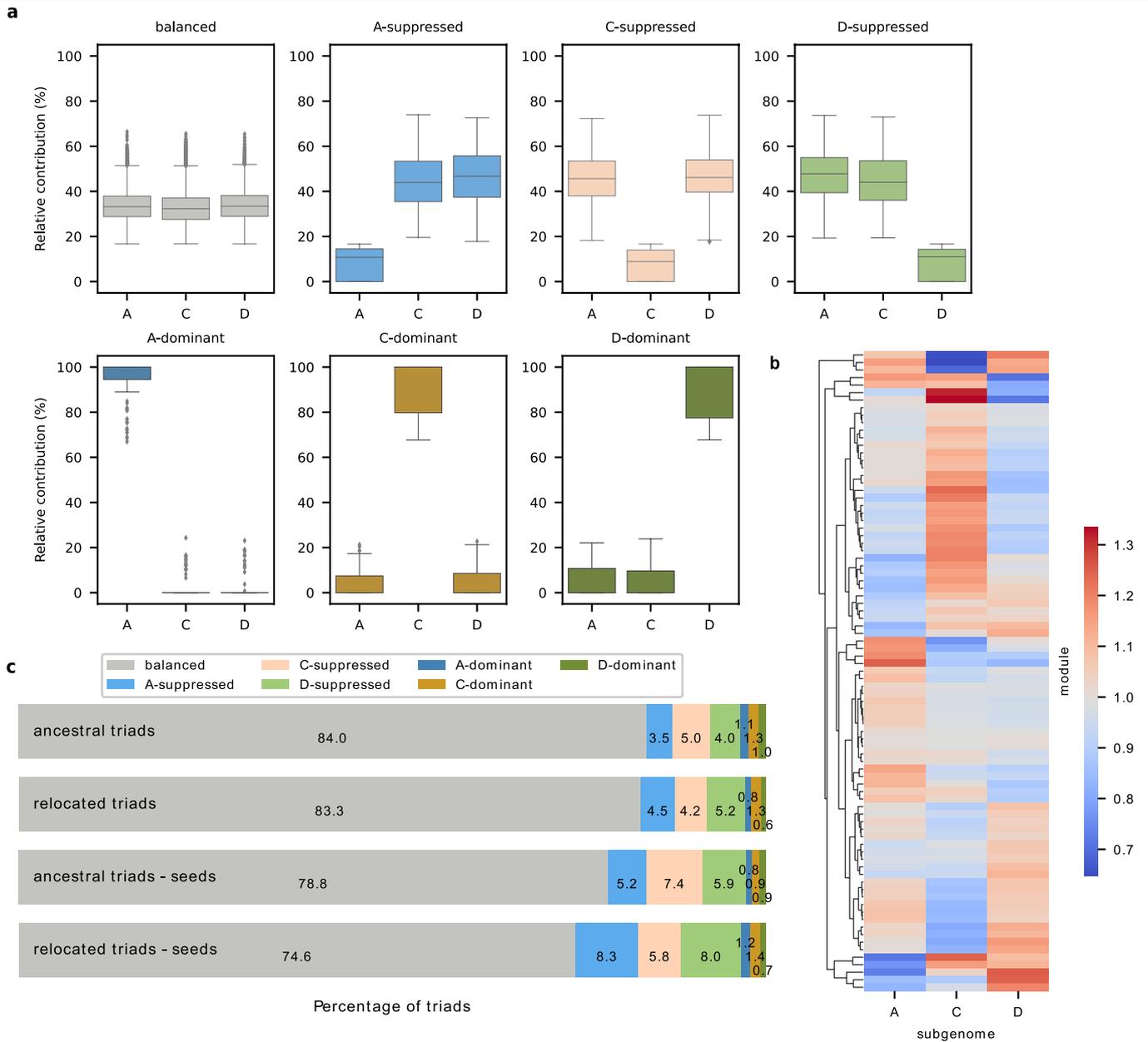
**Extended Data Fig. 5 | Genome rearrangements in oat. a**, Translocation history in *A. insularis* and *A. sativa* cv. *Sang*. Top panel displays likely ancestral syntenic blocks for *Avena* chromosomes 1, 2, 3 and 6, for chromosomes 4, 5 and 7 we only provide ancestral states in the A-, D- and C-lineages separately. From top to bottom, subsequent panels show chromosomes and translocation operations for the initial tetraploid (directly after the hybridization of the diploid ancestors), the ancestral tetraploid after translocations, translocation events occurring after the split of *A. insularis* and oat, and the extant states. Genomic positions for each inter-subgenomic translocation in the extant tetra- and hexaploid are provided in Supplementary Table 8. A-, C- and D-subgenomic

blocks are coloured in blue, yellow and green, respectively. Blocks contributing to translocations are accented in the mid panels by darker colours. Circled numbers represent translocation events: (1) represents two non-reciprocal translocations between the ancestral 1C and 3D chromosomes, (2) to (5) are additional translocations shared by *A. sativa* and *A. insularis*, while (6) and (7) are translocation differentiating extant oat and *A. insularis*. Asterisks refer to the ancestral state of chromosomes. **b**, Schematic representation of syntenic genomic regions between bread wheat (chromosome 5B harbouring *TaZIP4-B2*) and oat (chromosomes 5A, 5C, 5D). At the syntenic position, a *TaZIP4-B2* ortholog is absent in the oat genome.



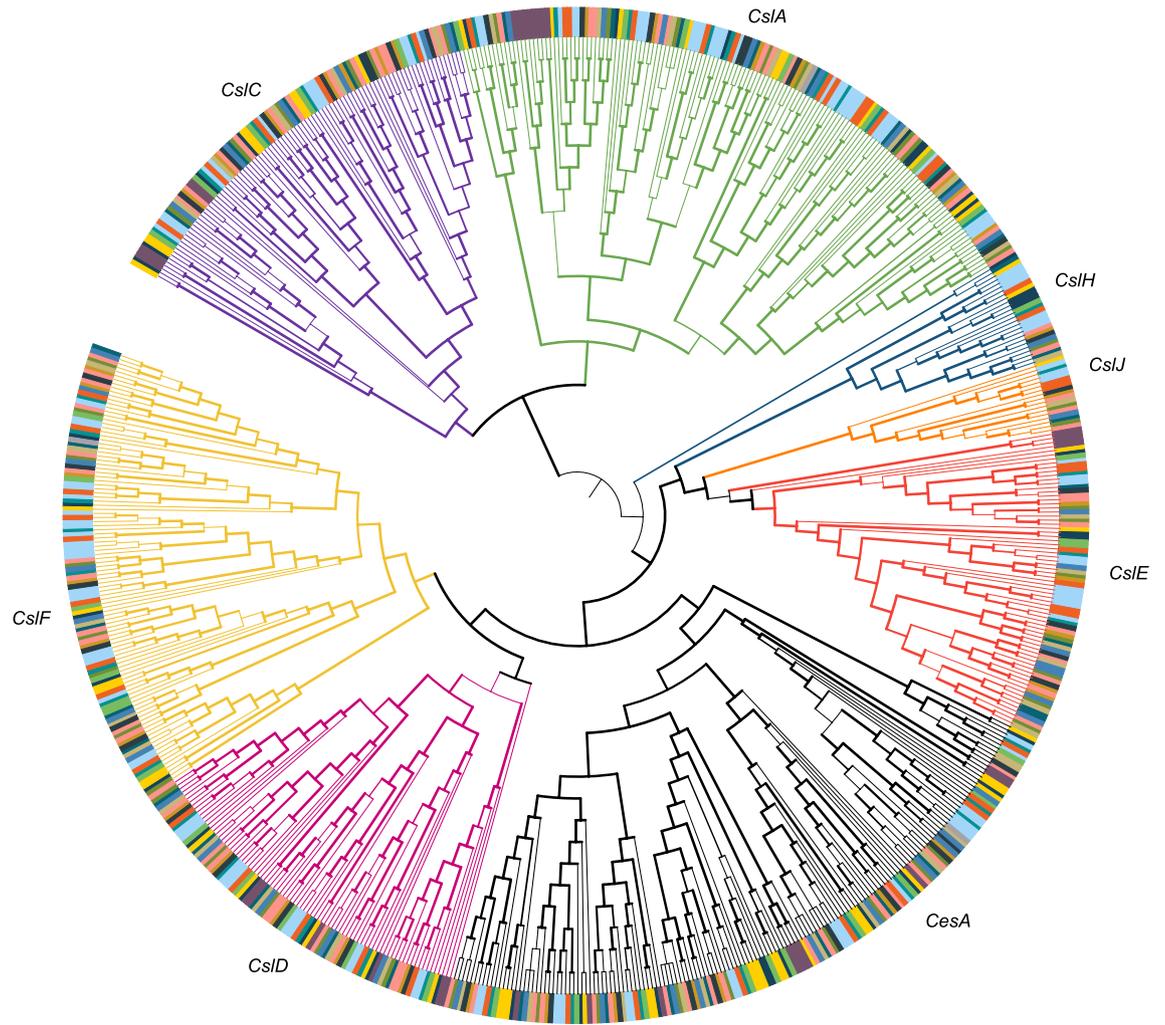
**Extended Data Fig. 6 | An example of a hidden breeding barrier in *Avena sativa* revealed by recombination matrices in two oat populations.** Average pairwise recombination frequencies ( $r$ ) among chromosomes 1A, 1C and 1D were computed in 16 Mbp windows at 1 Mbp increments for two RIL populations and are visualised as blended colours of yellow ( $r = 0$ ) to cyan ( $r = 0.25$ ) to burgundy ( $r = 0.5$ ). Blocks of yellow along the diagonal dashed lines indicate recombination suppression within a chromosome. Blocks of yellow off the diagonal represent pseudo-linkage between chromosomes. **a**, recombination in the population 'Goslin' x 'HiFi', where Goslin carries a 1C-A

translocation (like Sang) and HiFi is a non-carrier. The result is that a large proportion of 1C shows pseudo linkage to the part of 1A where the translocation is present (large yellow rectangle in the intersection of 1A and 1C). **b**, both parents of the population ('TX07CS-1948' x 'Hidalgo') carry the same translocation as Sang thus no pseudo linkage is observed. The 1C-1A translocation is associated with adaptation-related traits<sup>20</sup>, with pseudo linkage limiting the possibility of recombining different traits affected by genes in the translocated region in crosses between translocated and non-translocated germplasm.



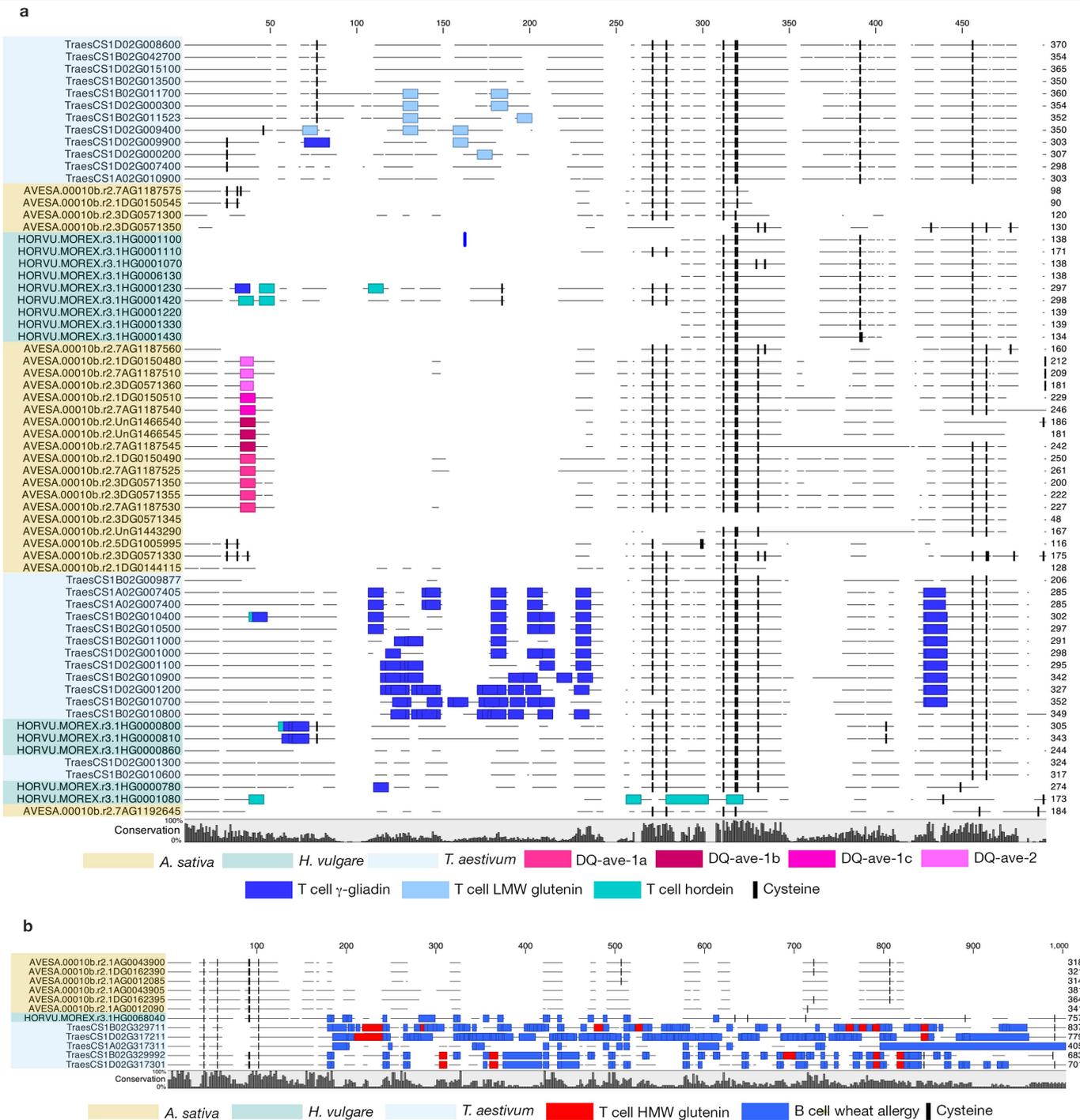
**Extended Data Fig. 7 | Gene expression analysis in the hexaploid oat genome.** **a**, Box plots representation of the relative contribution of each subgenome based on triad assignment to the seven categories defined, balanced  $n = 6,308$ , A-suppressed  $n = 263$ , C-suppressed  $n = 378$ , D-suppressed  $n = 303$ , A-dominant  $n = 83$ , C-dominant  $n = 98$ , D-dominant  $n = 74$ ,  $n$  = number of triads. Values from the first to third quartiles are shown within the boxplots boxes (inter-quartile range) with the median represented by the middle line. The upper and lower whiskers extend from the edge to the largest and smallest

value of the edge but no further than  $1.5 \times$  the inter-quartile range, the data beyond the end of the whiskers are outliers plotted individually. **b**, Heatmap representation of WGCNA modules showing the relation between the expected representation of each subgenome in the module based on the overall number of genes per subgenome and the observed one.  $>1$ : higher than expected,  $<1$  lower than expected,  $1$  = as expected. **c**, Gene expression patterns in ancestral triads versus triads with genes positioned in translocated regions across all samples (upper two panels) and in seed tissue (lower two panels).



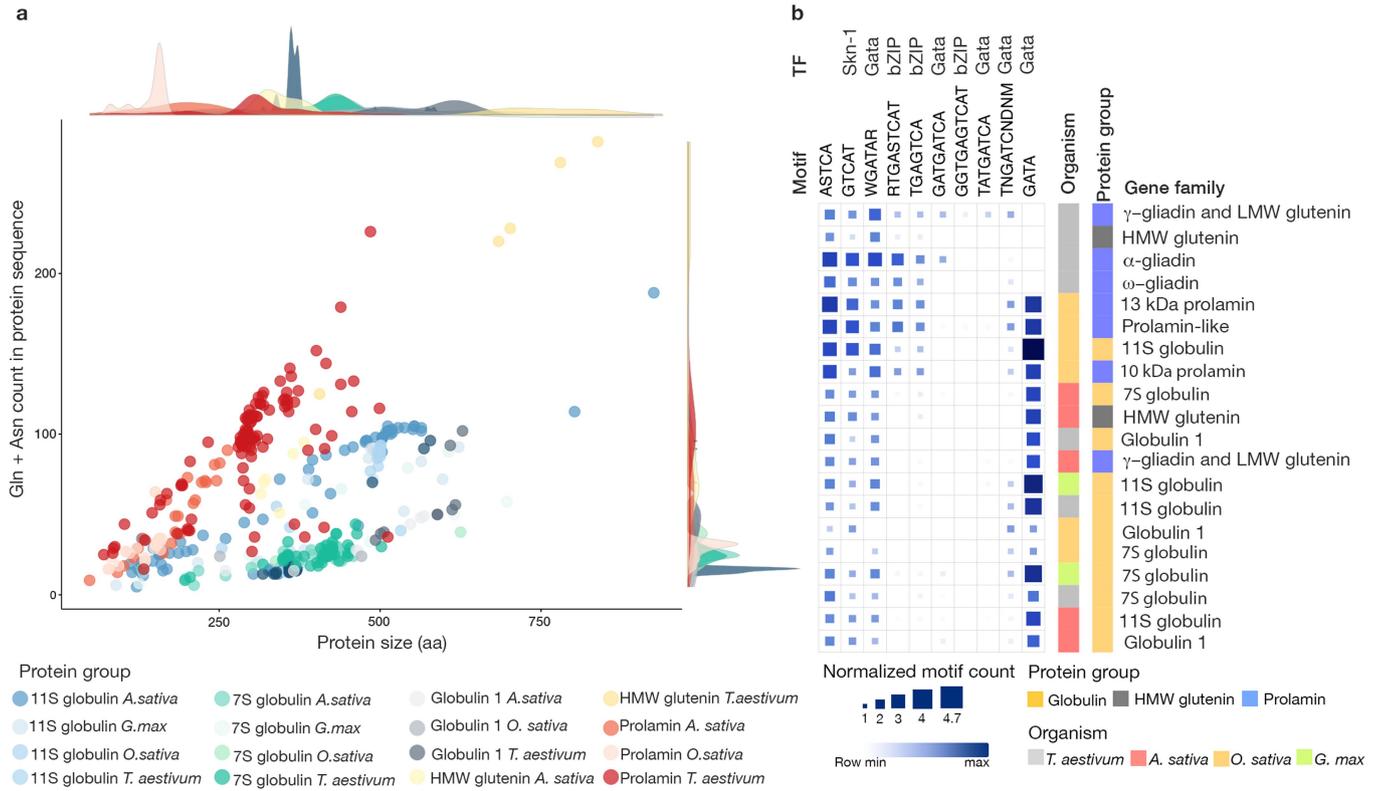
**Extended Data Fig. 8 | Phylogeny of the Cesa-subfamily of cellulose synthase genes and the Cellulose synthase-like subfamilies.** *CslA, C, D, E, F, H, and J* in 11 different species including oat. Tree branches for different

subfamilies are shown in different colours. Branch thickness corresponds to bootstrap values and increases with higher bootstrap.



**Extended Data Fig. 9 | Comparative analysis of coeliac disease and allergy related epitopes in oat, wheat and barley. a,** Protein sequence alignment of avenins with  $\gamma$ -glutamins, B-hordeins, and LMW glutenins shows the conserved position of cysteine residues (black bars) and differences in the T cell epitope prevalence and variability. Blue bars represent T cell epitopes characteristic at  $\gamma$ -glutamins, green bars indicate epitopes characteristic at LMW glutenins, pink bars show positions of avenin-specific epitopes. Colour code on the left side

indicates the species: oat (highlighted in pale yellow); wheat (highlighted in pale blue); and barley (highlighted in jagged ice). **b,** Sequence alignment of HMW glutenins show the conserved position of cysteine residues in the N- and C-terminal regions in oat and wheat HMW glutenins and barley D-hordein and shows the complete absence of coeliac disease (highlighted in red) and wheat allergy-related epitopes (shown as blue bars) in the sequences.



**Extended Data Fig. 10 | Comparative analysis of nitrogen storing capacity and transcription factor binding site motifs. a**, Size distribution of prolamin and globulin protein sequences in amino acid residues (aa) identified in the genomes of hexaploid oat cultivar Sang, the IWGSC v1 wheat reference genome, *Oryza sativa* Japonica IRGSP-1.0 rice genome and soybean reference genome, *Glycine max* v2.1 in relation to their nitrogen storing capacity, measured as a sum of asparagine and glutamine content in their sequence.

**b**, Comparison of nitrogen-responsive transcription factor binding site patterns in oat, wheat and rice prolamins and oat, wheat, rice and soybean globulins. Consensus motifs identified in the 0 to -500 region and the related transcription factor families are labelled in each column; total numbers of detected motifs normalised by the number of gene models are indicated in each row.

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

**Data collection** Custom code developed in this study can be found in the GitHub node at <https://github.com/PGSB-HMGU/oatkmers>. Code developed in other studies (but used here) can be found at <https://github.com/PGSB-HMGU/plant.annot>. Software for data collection included HiSeq Control Software HD 3.4.0.38/RTA 2.7.7, NovaSeq Control Software 1.7.0/RTA v3.4.4, and bcl2fastq\_v2.20.0.422.

**Data analysis** A myriad software was used in this study, all of which have been listed and cited. These include DeNovoMAGIC v3.0, TRITEX pipeline, EMBOS restrict, Minimap2, SAMtools, Novosort (<http://www.novocraft.com/products/novosort/>), BEDTools, FASTQC v0.11.8, Canu v1.9, Arrow from the GenomicConsensus package in the Pacific BioSciences SMRT portal v5.1.0, PILON v0.22, BBDuk (<https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/>), cactus pipeline v1.0, tallymer subtools from the genome tools package v1.6.1, GenomeThreader, v1.7.1, HISAT2 v2.1.0, Stringtie v1.2.3, Transdecoder v3.0.0, BLAST+, gffread v0.11.6, Biopython, Diamond v0.9.25 and v0.9.29.130, STAR, v2.7.6a, Trinity v2.8.5, PASApipeline v2.3.3 and v2.4.1, GMAP v2017.11.15 and v2020.06.01, TransDecoder v5.5.0, HMMER3 v3.1b2 and v3.2.1, Augustus v3.3.1 and v3.3.3, EvidenceModeler v73350ce, BUSCO v3.0.2 liliopsida\_odb10 created on 2017-12-01 and embryophyta\_odb9 created on 2017-02-13 (protein mode), BUSCO v4.0.4 liliopsida\_odb10 created on 2019-11-20 (protein mode), AHRD v3.3.3, TandemRepeatsFinder v4.07b, vmatch dbcluster, R, Haplotag, FSHap as implemented in TASSEL v5.0, KMC tools v3, Louvain algorithm (<https://github.com/taynaud/python-louvain>), Orthofinder v2.4, SciPy v1.6.1, networkx v2.5, InterProScan5, CLC Genomics Workbench v21, gpubr, FileMaker Pro Advanced v17, MEME suit, Morpheus R package, ProteinPilot™ 5.0.3 software (SCIEX), McScan of the jvci utility library (<https://github.com/tanghaibao/jvci>), MUSCLE v3.8.155, fasttree v2.1.10 and v2.1.11, iTol v6.3, Trimmomatic, fastp v0.20.0, Salmon v1.1.0, tximport package v1.12.3, DESeq2 v1.24.0, ggtern, WGCNA, tidyverse v1.3.1, ggplot2 v3.3.5, vcfR v1.12.0, svglite v2.0.0, fs v1.5.0, slider v0.2.2, tidymodels v0.1.4, treeio v1.16.1, ggtree v3.0.1, patchwork v1.1.1, gggenes v0.4.1, BWA-MEM2 v2.2.1, MultiQC v1.10.1, DeepVariant v1.1.0, GNU Parallel v20210422, GLNexus v1.3.1, Snakemake v6.5.1, bcftools v1.12, McScanX, Matlab, and SnpEff v4.3.1t.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The raw sequence data used for de novo whole genome assembly are available from the European Nucleotide Archive (ENA) under accession number PRJEB44810 (*A. sativa* cv. Sang) and from the Sequence Read Archive (SRA) under accession number PRJNA727490 (*A. insularis* BYU209) and PRJNA726919 (*A. longiglumis* CN58138). Chromosome conformation capture (Hi-C) sequencing data are available from ENA under accession PRJEB43668 (*A. sativa* cv. Sang), PRJEB43670 (*A. insularis* BYU209) and PRJEB43669 (*A. longiglumis* CN58138). Chromosome-scale sequence assemblies (pseudomolecules) are available from ENA under accession PRJEB44810 (*A. sativa* cv. Sang), PRJEB45088 (*A. insularis* BYU209) and PRJEB45087 (*A. longiglumis* CN58138). The raw RNA-seq and WGS data generated in this study are available under ENA accession number PRJEB46365. Pseudomolecules, annotation data and analysis results are available in the Plant Genomics & Phenomics (PGP) Research Data Repository at <http://dx.doi.org/10.5447/ipk/2022/2>. The DOI was registered using eDAL (<https://edal.ipk-gatersleben.de/>). Pseudomolecules, annotation data and associated analyses for *A. sativa* cv. Sang, *A. longiglumis*, and *A. insularis* are also available from GrainGenes60: Sang genome browser: <https://wheat.pw.usda.gov/jb/?data=/ggds/oat-sang>; Sang data download: <https://wheat.pw.usda.gov/GG3/content/avena-sang-download>; *A. longiglumis* genome browser: <https://wheat.pw.usda.gov/jb/?data=/ggds/oat-longiglumis>; *A. longiglumis* data download: <https://wheat.pw.usda.gov/GG3/content/avena-longiglumis-download>; *A. insularis* genome browser: <https://wheat.pw.usda.gov/jb/?data=/ggds/oat-insularis>; *A. insularis* data download: <https://wheat.pw.usda.gov/GG3/content/avena-insularis-download>. The mass spectrometry proteomics data and ProteinPilot search result files have been deposited to MassIVE (UCSD, San Diego, CA, USA; <https://massive.ucsd.edu>) under accession number MSV000088727. The publicly available OT3098 oat genome data was generated by PepsiCo and Corteva Agriscience. This dataset (annotation version 2) has been obtained and is available from GrainGenes: <https://wheat.pw.usda.gov/GG3/content/pepsico-ot3098-hexaploid-oat-version-2-genome-assembly-release-collaboration-graingenes>. Databases used in this study include PTREP release 19, Uniref download 2019-09-03, Pfam download 2019-09-03, Swiss-Prot, TAIR, TrEMBL, REdat\_9.9\_Poaceae section of the PGSB transposon library, Immune Epitope Database and Analysis Resource (<https://www.iedb.org>), PLACE and PlantCare promoter motif databases, and pfam2GO.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to establish sample size for genome sequencing and assembly. The two progenitor Avena accessions were chosen as the likely descendant of the hexaploid A, C and D subgenomes based on previous marker data analysis. The Sang cultivar was chosen as a representative Spring oat cultivar and to facilitate single gene mapping in a closely related TILLING population.
Data exclusions	All sequencing data generated and reported as raw data was used in the genome assembly and analyses.
Replication	In all analyses that support the genome assemblies, gene expression, proteomics, GC/MS and SEM, the number of replicates or iterations are indicated in materials and methods or supplemental tables. In each case, replications were successful and used. The genome assemblies themselves were validated using multiple methods i.e. BUSCO, genetic maps, HiC, and for <i>A. sativa</i> multiple comparisons to oat long-read assembly OT3098 were performed. This helped validate the other approaches.
Randomization	Randomization does not directly apply to the genome sequencing and assembly. However it does apply to some of the analyses conducted. In these cases, the group design and data seeding for computational analysis are described in the materials and methods and adhere to widely accepted standards. For example, bootstrapping was applied to all phylogenies computed (e.g. Fig. 3a).
Blinding	Blinding does not apply to this study, as the study focuses on genome sequencing. This study focuses on plants genomics and the results of the study are not impacted by the concealment of treatment, data, or groups.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

- | n/a                                 | Involvement in the study                               |
|-------------------------------------|--|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies                    |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines         |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Human research participants   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data                 |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern  |

## Methods

- | n/a                                 | Involvement in the study                        |
|-------------------------------------|---|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq               |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry         |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |