

1 **Chromosome-scale genome assembly of bread wheat's wild relative *Triticum timopheevii***

2
3 Surbhi Grewal¹, Cai-yun Yang¹, Duncan Scholefield¹, Stephen Ashling¹, Sreya Ghosh², David
4 Swarbreck², Joanna Collins³, Eric Yao^{4,5}, Taner Z. Sen^{4,5}, Michael Wilson⁶, Levi Yant⁶, Ian P. King¹ and
5 Julie King¹

6

7 1. Wheat Research Centre, Department of Plant and Crop Sciences, School of Biosciences, University
8 of Nottingham, Loughborough, LE12 5RD, UK

9 2. Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, UK

10 3. Genome Reference Informatics Team, Wellcome Sanger Institute, Wellcome Trust Genome
11 Campus, Hinxton, CB10 1RQ, UK

12 4. University of California, Department of Bioengineering, Berkeley, CA, 94720, USA

13 5. United States Department of Agriculture—Agricultural Research Service, Western Regional
14 Research Center, Crop Improvement and Genetics Research Unit, 800 Buchanan St., Albany, CA
15 94710, USA

16 6. University of Nottingham, University Park, Nottingham, NG7 2RD

17 Corresponding author: Surbhi Grewal (surbhi.grewal@nottingham.ac.uk)

18

19 **Abstract**

20

21 Wheat (*Triticum aestivum*) is one of the most important food crops with an urgent need for increase
22 in its production to feed the growing world. *Triticum timopheevii* ($2n = 4x = 28$) is an allotetraploid
23 wheat wild relative species containing the A^t and G genomes that has been exploited in many pre-
24 breeding programmes for wheat improvement. In this study, we report the generation of a
25 chromosome-scale reference genome assembly of *T. timopheevii* accession PI 94760 based on PacBio
26 HiFi reads and chromosome conformation capture (Hi-C). The assembly comprised a total size of
27 9.35 Gb, featuring a contig N50 of 42.4 Mb, and 166,325 predicted gene models. DNA methylation
28 analysis showed that the G genome had on average more methylated bases than the A^t genome. The
29 G genome was also more closely related to the S genome of *Aegilops speltoides* than to the B
30 genome of hexaploid or tetraploid wheat. In summary, the *T. timopheevii* genome assembly provides
31 a valuable resource for genome-informed discovery of agronomically important genes for food
32 security.

33

34 **Background and Summary**

35

36 The *Triticum* genus comprises many wild and cultivated wheat species including diploid, tetraploid
37 and hexaploid forms. The polyploid species originated after hybridisation between *Triticum* and the
38 neighbouring *Aegilops* genus (goatgrass). The tetraploid species, *Triticum turgidum* ($2n = 4x = 28$,
39 AABB), also known as emmer wheat, and *Triticum timopheevii* ($2n = 4x = 28$, A^tA^tGG) are
40 polyphyletic. *Triticum urartu* Thum. ex Gandil ($2n = 2x = 14$, AA) is the A genome donor for both
41 these species¹ whereas, the B and G genomes are closely related to the S genome of *Aegilops*
42 *speltoides*². Both tetraploid species have wild and domesticated forms, i.e., *T. turgidum* L. ssp.
43 *dicoccoides* (Körn. ex Asch. & Graebn.) Thell. and ssp. *dicoccum* (Schrank ex Schübl.) Thell.,
44 respectively, and *T. timopheevii* (Zhuk.) Zhuk. ssp. *armeniacum* (Jakubz.) Slageren and ssp.
45 *timopheevii*, respectively. Additionally, tetraploid durum wheat *T. turgidum* L. ssp. *durum* (Desf.)
46 Husn. ($2n = 4x = 28$, AABB), used for pasta production, and hexaploid bread wheat *Triticum aestivum*
47 L. ($2n = 6x = 42$, AABBDD) evolved from domesticated emmer wheat with the latter originating
48 through hybridisation with *Aegilops tauschii* (D genome donor) 6,000-7,000 years ago. Hexaploid
49 *Triticum zhukovskyi* (AAGGA^mA^m) originated from hybridisation of cultivated *T. timopheevii* and
50 cultivated einkorn *Triticum monococcum*³ ($2n = 2x = 14$, A^mA^m).

51

52 The G genome is only found in *T. timopheevii* and *T. zhukovskyi* and is virtually identical to the S
53 genome on a molecular level^{4,5} but differs from it, and the B genome, due to a number of
54 chromosomal rearrangements and translocations involving the A^t genome⁶. The most studied are the
55 6A^t/1G/4G and 4G/4A^t/3A^t translocations in *T. timopheevii*⁷⁻¹⁰.

56
57 *Triticum timopheevii* ssp. *timopheevii* has been exploited in various studies for wheat improvement
58 as it has been shown to be an abundant source for genetic variation for many traits such as
59 resistance to leaf rust¹¹⁻¹³, stem rust¹⁴⁻¹⁶, powdery mildew¹⁶⁻¹⁸, fusarium head blight^{19,20} Hessian fly,
60 Septoria blotch, wheat curl mite and tan spot²¹. It has also been shown to have tolerance to abiotic
61 stresses such as salinity^{22,23} and be a good source for traits affecting grain quality such as milling yield
62 and grain protein²⁴ and grain mineral content²⁵. During sequence analysis of reference quality
63 assemblies (RQA) of 10 wheat cultivars, recent studies found two of them, cv. LongReach Lancer and
64 cv. Julius, contained major introgressions on Chr2B (among others) potentially originating from *T. timopheevii*^{26,27}. Introgressions from *T. timopheevii* have also been found in many other wheat
65 accessions present in genebank collections²⁸. Pre-breeding programmes involving the introgression
66 of the whole genome of *T. timopheevii*, in small segments, into bread wheat^{10,29} with diagnostic KASP
67 markers that can track these introgressions in wheat^{29,30} have provided promising new germplasm
68 and tools to the wheat research community.

69
70
71 In this study, we report a chromosome-scale reference genome sequence assembly for *T. timopheevii*
72 by integrating chromatin conformation capture (Hi-C) derived short-reads³¹ with PacBio HiFi long-
73 reads³². The assembly was annotated for gene models and repeats. CpG methylation along the
74 chromosomes was inferred from the PacBio CCS data. Known chromosomal translocations within
75 and between the A^t and G genomes were confirmed, and new chromosome rearrangements were
76 found in comparison to wild emmer wheat. The high-quality *T. timopheevii* genome assembly
77 obtained in this study provides a reference for the G genome of the *Triticum* genus. This new
78 resource will form the basis to study chromosome rearrangements across different Triticeae species
79 and will be explored to detect A^t and G genome introgressions in durum and bread wheat allowing
80 future genome-informed gene discoveries for various agronomic traits.

81
82 **Methods**

83

84 **Plant material, nucleic acid extraction and sequencing**

85

86 High molecular weight (HMW) DNA was extracted from a young seedling (dark-treated for 48 hours)
87 of *T. timopheevii* accession PI 94760 (United States National Plant Germplasm System, NPGS
88 available at <https://npgsweb.ars-grin.gov/gringlobal/search>) using a modified Qiagen Genomic DNA
89 extraction protocol (<https://doi.org/10.17504/protocols.io.bafmibk6>)³³. DNA was sheared to the
90 appropriate size range (15–20 kb) and PacBio HiFi sequencing libraries were constructed by
91 Novogene (UK) Company Limited. Sequencing was performed on 10 SMRT cells of the PacBio Sequel
92 II system in CCS mode with kinetics option to generate ~267 Gb (~28-fold coverage) of long HiFi reads
93 (Table S1). Four Hi-C libraries were prepared using leaf samples (from the same plant used for HMW
94 DNA extraction), at Phase Genomics (Seattle, USA) using the Proximo® Hi-C Kit for plant tissues
95 according to the manufacturer's protocol. The Hi-C libraries were sequenced on an Illumina NovaSeq
96 6000 S4 platform to generate ~2.8 billion of paired end 150bp reads (~842 Gb raw data; ~89-fold
97 coverage; Table S2).

98

99 Total RNA was extracted from seedlings (3-leaf stage), seedlings at dusk, roots, flag leaves, spikes and
100 grains. Flag leaf and whole spike were collected at 7 days post-anthesis and whole grains were
101 collected at 15 days post-anthesis. In brief, 100 mg of ground powder from each tissue was used for
102 RNA isolation using the RNeasy Plant Mini Kit (#74904, QIAGEN Ltd UK) following manufacturer's

103 instructions. The RNA samples were split into 2 aliquots, one for mRNA sequencing (RNA-Seq) and
104 one for Iso-Seq³⁴. Library construction for both types of sequencing was carried out by Novogene
105 (UK) Company Limited. Illumina NovaSeq 6000 S4 platform was used for mRNA sequencing to
106 generate on average 450 million reads (~67 Gb of 2 x 150bp reads) for each sample (Table S3). The
107 second set of RNA aliquots from each of the six tissues were pooled into one sample and sequenced
108 on the PacBio Sequel II system using the Iso-Seq pipeline to generate 4.47 Gb of Iso-Seq data (Table
109 S4) which was analysed using PacBio Iso-Seq analysis pipeline (SMRT Link v12.0.0.177059).

110

111 Plants were grown in a glasshouse in 2L pots containing John Innes No. 2 soil and maintained at 18 –
112 25 °C under 16 h light and 8 h dark conditions. All sequencing was carried out by Novogene (UK)
113 Company Limited.

114

115 Cleaning of sequencing data

116

117 The HiFi sequencing read files in BAM format were converted and combined into one fastq file using
118 bam2fastq v1.3.1 (available at <https://github.com/jts/bam2fastq>). Reads with PacBio adapters were
119 removed using cutadapt v4.1³⁵ with parameters: --error-rate=0.1 --times=3 --overlap=35 --
120 action=trim --revcomp --discard-trimmed. Hi-C reads were trimmed to remove Illumina adapters
121 using Trimmomatic v0.39³⁶ with parameters ILLUMINACLIP:TruSeq3-PE-
122 2.fa:2:30:10:2:keepBothReads SLIDINGWINDOW:4:20 MINLEN:40 CROP:150.

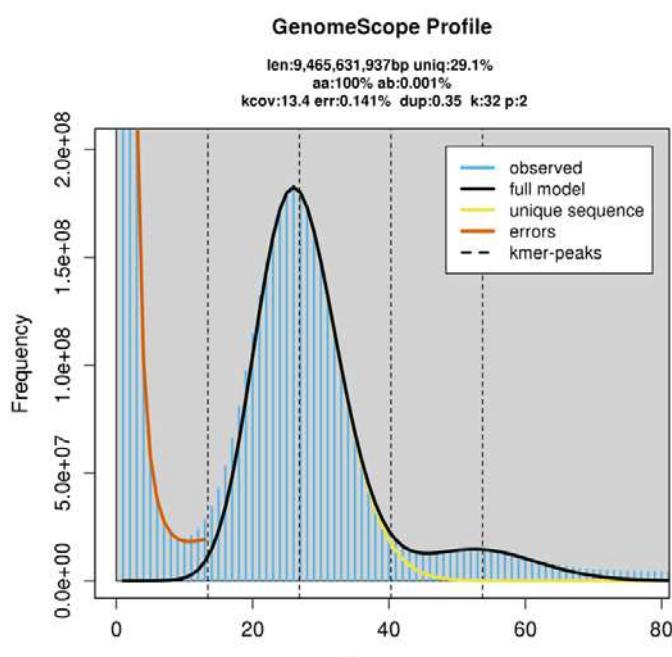
123

124 Genome size estimation

125

126 The size of the *T. timopheevii* genome was estimated by using k-mer (k=32) distribution analysis
127 with Jellyfish v2.2.10³⁷ on the cleaned HiFi reads³⁸. A k-mer count histogram was generated and the
128 size of the *T. timopheevii* genome was estimated as ~9.46 Gb with heterozygosity of 0.001% (Fig. 1),
129 using GenomeScope v2.0³⁹ (available at <http://qb.cshl.edu/genomescope/genomescope2.0/>) with
130 parameters: ploidy = 2, k-mer length = 32, max k-mer coverage = 1000000 and average k-mer
131 coverage = 10.

132



133

134

135 **Figure 1.** Genomescope profile for 32-mers based on HiFi reads.

136

137 **Chromosome-scale genome assembly**

138

139 The cleaned HiFi reads were assembled into the initial set of contigs using hifiasm v.0.16.1⁴⁰ with
140 default parameters for an inbred species (-l 0) and the dataset was assessed using gfastats v1.3.1⁴¹.
141 The contig assembly had a total size of ~9.41 Gb, with a contig N50 value of 43.12 Mb. Genome
142 completeness was assessed using the Benchmarking Universal Single-Copy Orthologs (BUSCO
143 v5.3.2)⁴² program with the embryophyta_odb10 database which yielded 99% of the complete BUSCO
144 genes. Contaminants (contigs other than those categorised as Streptophyta or no hit) were identified
145 using BlobTools v1.1.1⁴³ and removed.

146

147 To achieve chromosome-level assembly, the trimmed Hi-C data⁴⁴ was mapped onto the
148 decontaminated contig assembly using the Arima Genomics® mapping pipeline (available at
149 https://github.com/ArimaGenomics/mapping_pipeline) and chromosome construction was
150 conducted using the Salsa2⁴⁵ pipeline (available at <https://github.com/marbl/SALSA>) with default
151 parameters and GATC as the cutting site for the restriction enzyme (DpnII). The Hi-C contact map for
152 the scaffold assembly was constructed using PretextMap v0.1.9 and the chromatin contact matrix
153 was manually corrected using PretextView v0.2.5 by following the Rapid Curation pipeline⁴⁶
154 (<https://gitlab.com/wtsi-grit/rapid-curation>). The curated assembly was assessed using gfastats to
155 consist of 14 pseudomolecules and 1656 unplaced scaffolds with a total length of 9,350,839,849 bp
156 (including gaps) and a contig N50 of 42.4 Mb (Table 1). The orientation and the chromosome name
157 of each pseudomolecule were determined based on homology with the wheat cv. Chinese Spring
158 assembly RefSeq2.1⁴⁷ A and B subgenomes, using dotplot comparison of sequence alignments
159 produced by MUMmer's (v3.23⁴⁸) nucmer aligner and viewed on Dot (available at
160 <https://github.com/marianattestad/dot>). The pseudomolecules were thus, renamed into the 14
161 *T. timopheevii* chromosomes, seven A^t genome chromosomes with a total length of ~4.85 Gb and
162 consisting of 119 contigs and seven G genome chromosomes with a total length of ~4.40 Gb and
163 consisting of 529 contigs (Table 2).

164

165 **Table 1.** Summary statistics for genome assembly of *Triticum timopheevii*.

Assembly characteristics	Value
Number of scaffolds	1,670
Total scaffold length (bp)	9,350,839,849
Scaffold N50 (bp)	671,191,297
Largest scaffold (bp)	771,176,557
No. of contigs	2,304
Total contig length (bp)	9,350,587,949
Average contig length (bp)	4,058,415
Contig N50 (bp)	42,410,373
Largest contig (bp)	311,469,246
GC content (%)	46
BUSCO evaluation (% of complete BUSCO genes)	98.9

167

168

169 **Table 2.** Statistics of the *Triticum timopheevii* pseudomolecules

170

Chromosome	Length (bp)	Number of contigs	Number of gene models
Chr1A ^t	614,431,332	14	9,982
Chr1G	495,016,746	50	8,777
Chr2A ^t	767,071,137	10	12,729
Chr2G	671,256,291	72	13,941
Chr3A ^t	670,741,101	10	9,489
Chr3G	671,191,297	75	13,452
Chr4A ^t	771,176,557	23	12,878
Chr4G	643,128,204	68	9,936
Chr5A ^t	694,350,238	12	11,821
Chr5G	641,290,954	78	13,079
Chr6A ^t	585,824,631	33	9,011
Chr6G	589,079,669	87	11,406
Chr7A ^t	745,638,687	17	12,863
Chr7G	692,654,486	99	14,851
Unplaced scaffolds	97,988,519	1656	2,110
Total	9,350,839,849	2,304	166,325

171

172

Organellar genome assembly

173

174

De novo assembly of the organelle genomes was carried out using the Oatk pipeline (available at <https://github.com/c-zhou/oatk>) with the cleaned HiFi reads. The circular chloroplast and mitochondrial contigs were assembled with a total size of 136,158 bp and 443,464 bp, respectively. Any unanchored contigs that aligned to these extranuclear genomes were removed from the final assembly.

175

176

177

178

Genome annotation

179

180

181

Gene models were generated from the *T. timopheevii* assembly using REAT - Robust and Extendable eukaryotic Annotation Toolkit (<https://github.com/EI-CoreBioinformatics/reat>) and Minos (<https://github.com/EI-CoreBioinformatics/minos>) which make use of Mikado⁴⁹ (<https://github.com/EI-CoreBioinformatics/mikado>), Portcullis (<https://github.com/EI-CoreBioinformatics/portcullis>) and many third-party tools (listed in the above repositories). A consistent gene naming standard⁵⁰ was used to make the gene models uniquely identifiable.

182

183

1. Repeat identification

184

185

186

187

188

189

Repeat annotation was performed using EI-Repeat version 1.3.4 pipeline (<https://github.com/EI-CoreBioinformatics/eirepeat>) which uses third party tools for repeat calling. In the pipeline, RepeatModeler (v1.0.11 - <http://www.repeatmasker.org/RepeatModeler/>) was used for *de novo* identification of repetitive elements from the assembled *T. timopheevii* genome. High copy protein coding genes potentially included in the RepeatModeler library were identified and effectively removed by running RepeatMasker v4.0.7 using a curated set of high confidence *T. aestivum* coding genes to hard mask the RepeatModeler library; transposable element genes were first excluded from the *T. aestivum* coding gene set by running TransposonPSI (r08222010). Unclassified repeats were searched in a custom BLAST database of organellar genomes (mitochondrial and chloroplast

190

191

192

193

194

195

196

197

198

199

200 sequences from *Triticum* in the NCBI nucleotide division). Any repeat families matching organellar
201 DNA were also hard-masked. Repeat identification was completed by running RepeatMasker v4.0.72
202 with a RepBase embryophyte library and with the customized RepeatModeler library (i.e. after
203 masking out protein coding genes), both using the -nolow setting.
204

205 2. Reference guided transcriptome reconstruction

206

207 Gene models were derived from the RNA-Seq reads, Iso-Seq transcripts (122,253 high quality and 82
208 low quality isoforms; Supplementary File 1) and Full-Length Non- Concatamer Reads (FLNC) using the
209 REAT transcriptome workflow. HISAT2 v2.2.1⁵¹ was selected as the short read aligner with Iso-Seq
210 transcripts aligned with minimap2 v2.18-r1015⁵², maximum intron length was set as 50,000 bp and
211 minimum intron length to 20bp. Iso-Seq alignments were required to meet 95% coverage and 90%
212 identity. High-confidence splice junctions were identified by Portcullis v 1.2.4⁵³. RNA-Seq Illumina
213 reads were assembled for each tissue with StringTie2 v2.1.5⁵⁴ and Scallop v0.10.5⁵⁵, while FLNC reads
214 were assembled using StringTie2 (Table S5). Gene models were derived from the RNA-Seq
215 assemblies and Iso-Seq and FLNC alignments with Mikado. Mikado was run with all Scallop,
216 StringTie2, Iso-Seq and FLNC alignments and a second run with only Iso-Seq and FLNC alignments
217 (Table S6).
218

219 3. Cross-species protein alignment

220

221 Protein sequences from 10 Poaceae species (Table S7) were aligned to the *T. timopheevii* assembly
222 using the REAT Homology workflow with options --annotation_filters aa_len
223 --alignment_species Angiosp --filter_max_intron 20000 -- filter_min_exon 10 --alignment_filters
224 aa_len internal_stop intron_len exon_len splicing -- alignment_min_coverage 90 --junction_f1_filter
225 40 --post_alignment_clip clip_term_intron-exon - -term5i_len 5000 --term3i_len 5000 --term5c_len
226 36 --term3c_len 36. The REAT Homology workflow aligns proteins with spaln v2.4.7⁵⁶ and filters and
227 generates metrics to remove misaligned proteins. Simultaneously, the same protein set were also
228 aligned using miniprot v0.3⁵⁷ and similarly filtered as in the REAT homology workflow. The aligned
229 proteins from both methods were clustered into loci and a consolidated set of gene models were
230 derived via Mikado.
231

232 4. Evidence guided gene prediction

233

234 The evidence guided annotation of protein coding genes based on repeats, RNA-Seq mappings,
235 transcript assembly and alignment of protein sequences was created using the REAT prediction
236 workflow. The pipeline has four main steps: (1) the REAT transcriptome and homology Mikado
237 models are categorised based on alignments to UniProt proteins to identify models with likely full-
238 length CDS and which meet basic structural checks i.e., having complete but not excessively long
239 UTRs and not exceeding a minimum CDS/cDNA ratio. A subset of gene models is then selected from
240 the classified models and used to train the AUGUSTUS gene predictor⁵⁸; (2) Augustus is run in both
241 *ab initio* mode and with extrinsic evidence generated in the REAT transcriptome and homology runs
242 (repeats, protein alignments, RNA-Seq alignments, splice junctions, categorised Mikado models).
243 Three evidence guided AUGUSTUS predictions are created using alternative bonus scores and
244 priority based on evidence type; (3) AUGUSTUS models, REAT transcriptome/homology models,
245 protein and transcriptome alignments are provided to EVidenceModeler⁵⁹ (EVM) to generate
246 consensus gene structures; (4) EVM models are processed through Mikado to add UTR features and
247 splice variants.
248

249 5. Projection of gene models from *Triticum aestivum*

250

251 A reference set of hexaploid wheat gene models was derived from public gene sets (IWGSC⁶⁰ and 10+
252 wheat²⁶) projected onto the IWGSC RefSeq v1.0 assembly⁶⁰; a filtered and consolidated set of models
253 was derived with Minos, with a primary model defined for each gene. Models were scored on a
254 combination of intrinsic gene structure characteristics, evidence support (protein and transcriptome
255 data) and consistency in gene structure across the input gene models. The Minos primary models
256 were classified as full-length or partial based on alignment to a filtered magnoliopsida Swiss-Prot
257 TrEMBL database. This assignment, together with criteria for gene structure characteristics and the
258 original confidence classification, was used to classify models into 6 categories (Platinum, Gold,
259 Silver, Bronze, Stone and Paper), with Platinum being the highest confidence category for models
260 assessed as full-length, with an original confidence classification of "high", meeting structural checks
261 for number of UTR and CDS/cDNA ratio and which were assessed as consistently annotated across
262 the input gene sets. Reclassification resulted in 55,319 Platinum, 24,789 Gold, 11,968 Silver, 61,845
263 Bronze, 110,518 Stone and 115,336 Paper genes. The four highest confidence categories Platinum,
264 Gold, Silver and Bronze were projected onto the *T. timopheevii* assembly with Liftoff v1.5.1⁶¹, only
265 those models transferred fully with no loss of bases and identical exon/intron structure were
266 retained (<https://github.com/lucventurini/ei-liftover>). Similarly, high confidence genes annotated in
267 the hexaploid wheat cv. Chinese Spring RefSeq v2.1 assembly⁴⁷ were projected onto the
268 *T. timopheevii* genome assembly with Liftoff, and only those models transferred fully with no loss of
269 bases and identical exon/intron structure were retained. Among these, gene models with the
270 attribute "manually_curated" in the original Refseq v2.1 assembly were extracted as a set.
271

272 6. Gene model consolidation

273
274 The final set of gene models was selected using Minos (Table 3). Minos is a pipeline that generates
275 and utilises metrics derived from protein, transcript, and expression data sets to create a
276 consolidated set of gene models. In this annotation, the following gene models were filtered and
277 consolidated into a single set of gene models using Minos:

- 278 1. The three alternative evidence guided Augustus gene builds described earlier.
- 279 2. The gene models derived from the REAT transcriptome runs described earlier.
- 280 3. The gene models derived from the REAT homology runs described earlier.
- 281 4. The gene models derived from the REAT prediction run (AUGUSTUS and EVM-Mikado)
282 described earlier.
- 283 5. The gene models derived from projecting public and curated *T. aestivum* gene models of
284 varying confidence levels onto the *T. timopheevii* genome as described earlier.
- 285 6. IWGSC Refseq v2.1 models identified as "manually_curated" projected onto the
286 *T. timopheevii* genome as described earlier.

287
288 **Table 3.** Summary statistics for the final structural annotation of the *T. timopheevii* genome.
289

Stat	Value
Number of genes	166,325
Number of Transcripts	218,100
Transcripts per gene	1.31
Number of monoexonic genes	51,702
Monoexonic transcripts	53,192
Transcript mean size cDNA (bp)	1,658.27
Transcript median size cDNA (bp)	1412
Min cDNA	96
Max cDNA	20,589

Total exons	997,779
Exons per transcript	4.57
Exon mean size (bp)	362.47
CDS mean size (bp)	277.55
Transcript mean size CDS (bp)	1,171.61
Transcript median size CDS (bp)	957
Min CDS	0
Max CDS	20,283
Intron mean size (bp)	628.4
5'UTR mean size (bp)	182.93
3'UTR mean size (bp)	294.58

290

291 Gene models were classified as biotypes `protein_coding_gene`, `predicted_gene` and
292 `transposable_element_gene`, and assigned as high or low confidence (Table 3) based on the criteria
293 below:

294 a) **High confidence (HC) `protein_coding_gene`:** Any protein coding gene where any of its
295 associated gene models have a BUSCO v5.4.7⁶² protein status of Complete/Duplicated OR
296 have diamond v0.9.36 coverage (average across query and target coverage) $\geq 90\%$ against
297 the listed Poaceae protein datasets (section 3; Supplemental File 2) or UniProt
298 magnoliopsida proteins. Or alternatively have average blastp coverage (across query and
299 target coverage) $\geq 80\%$ against the listed protein datasets/UniProt magnoliopsida AND have
300 transcript alignment F1 score (average across nucleotide, exon and junction F1 scores based
301 on RNA-Seq transcript assemblies) $\geq 60\%$.

302 b) **Low confidence (LC) `protein_coding_gene`:** Any protein coding gene where all its associated
303 transcript models do not meet the criteria to be considered as high confidence protein
304 coding transcripts.

305 c) **HC `transposable_element_gene`:** Any protein coding gene where any of its associated gene
306 models have coverage $\geq 40\%$ against the combined interspersed repeats (see section 1).

307 d) **LC `transposable_element_gene`:** Any protein coding gene where all its associated transcript
308 models do not meet the criteria to be considered as high confidence and assigned as a
309 `transposable_element_gene` (see c).

310 e) **LC `predicted_gene`:** Any protein coding gene where all the associated transcript models do
311 not meet the criteria to be considered as high confidence protein coding transcripts. In
312 addition, where any of the associated gene models have average blastp coverage (across
313 query and target coverage) $< 30\%$ against the listed protein datasets AND having a protein-
314 coding potential score < 0.25 calculated using CPC2 0.1⁶³.

315 f) **LC `ncRNA_gene`:** Any gene model with no CDS features AND a protein-coding potential score
316 < 0.3 calculated using CPC2 0.1.

317 g) **Discarded models:** Any models having no BUSCO protein hit AND a protein alignment score
318 (average across nucleotide, exon and junction F1 scores based on protein alignments) < 0.2
319 AND a transcript alignment F1 score (average across nucleotide, exon and junction F1 scores
320 based on RNA-Seq transcript assemblies) < 0.2 AND a diamond coverage (target coverage)
321 < 0.3 AND Kallisto v0.44⁶⁴ expression score < 0.2 from across RNA-Seq reads OR having short
322 CDS < 30 bps. Any ncRNA genes (no CDS features) not meeting the ncRNA gene requirements
323 (f) were also excluded.

324

325 **Table 4.** Minos classified gene models.

326

Biotype	Confidence	Gene	Transcript
---------	------------	------	------------

protein_coding_gene	Low	73,844	79,329
protein_coding_gene	High	67,107	112,338
transposable_element_gene	Low	15,871	16,231
predicted_gene	Low	4,974	5,033
transposable_element_gene	High	3,258	3,410
ncRNA_gene	Low	1,271	1,759
Total		166,325	218,100

327

328 Gene model distribution across the pseudomolecules and unplaced scaffolds is shown in Table 2 and
329 gene density of 164,617 protein coding genes across the *T. timopheevii* genome is shown in Fig. 2b.

330

331 7. Functional annotation

332

333 All proteins were annotated using AHRD v.3.3.3 (available at
334 <https://github.com/groupschoof/AHRD/blob/master/README.textile>). Sequences were compared
335 against the reference proteins (*Arabidopsis thaliana* TAIR10,
336 TAIR10_pep_20101214_updated.fasta.gz - <https://www.araport.org>) and the UniProt viridiplantae
337 sequences⁶⁵ (data download 06-May-2023), both Swiss-Prot and TrEMBL datasets using blastp v2.6.0
338 with an e-value of 1e- 5. InterproScan v5.22.61⁶⁶ results were also provided to AHRD. The standard
339 AHRD example configuration file path test/resources/ahrd_example_input_go_prediction.yml,
340 distributed with the AHRD tool, was adapted apart from the location of input and output files. The
341 GOA mapping from UniProt
342 (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/goa_uniprot_all.gaf.gz) was included as
343 parameter 'gene_ontology_result'. The interpro database
344 (ftp://ftp.ebi.ac.uk/pub/databases/interpro/61.0/interpro.xml.gz) was included as parameter
345 'interpro_database'. The parameter 'prefer_reference_with_go_annos' was changed to 'false' and
346 the blast database specific weights used were:
347

blast_dbs:

348 swissprot:
349 weight: 100
350 description_score_bit_score_weight: 0.2
351 trembl:
352 weight: 50
353 description_score_bit_score_weight: 0.4
354 tair:
355 weight: 50
356 description_score_bit_score_weight: 0.4
357

358 Since *T. timopheevii* is known as an important source for genetic variation for resistance against
359 major diseases of wheat as described above and as the majority of cloned disease-resistance genes
360 encode nucleotide-binding leucine-rich repeats (NLRs)^{67,68}, we analysed the genomic distribution of
361 all gene models annotated as NB-ARC domain-containing/disease resistance proteins in the genome
362 assembly (Fig. 2c).

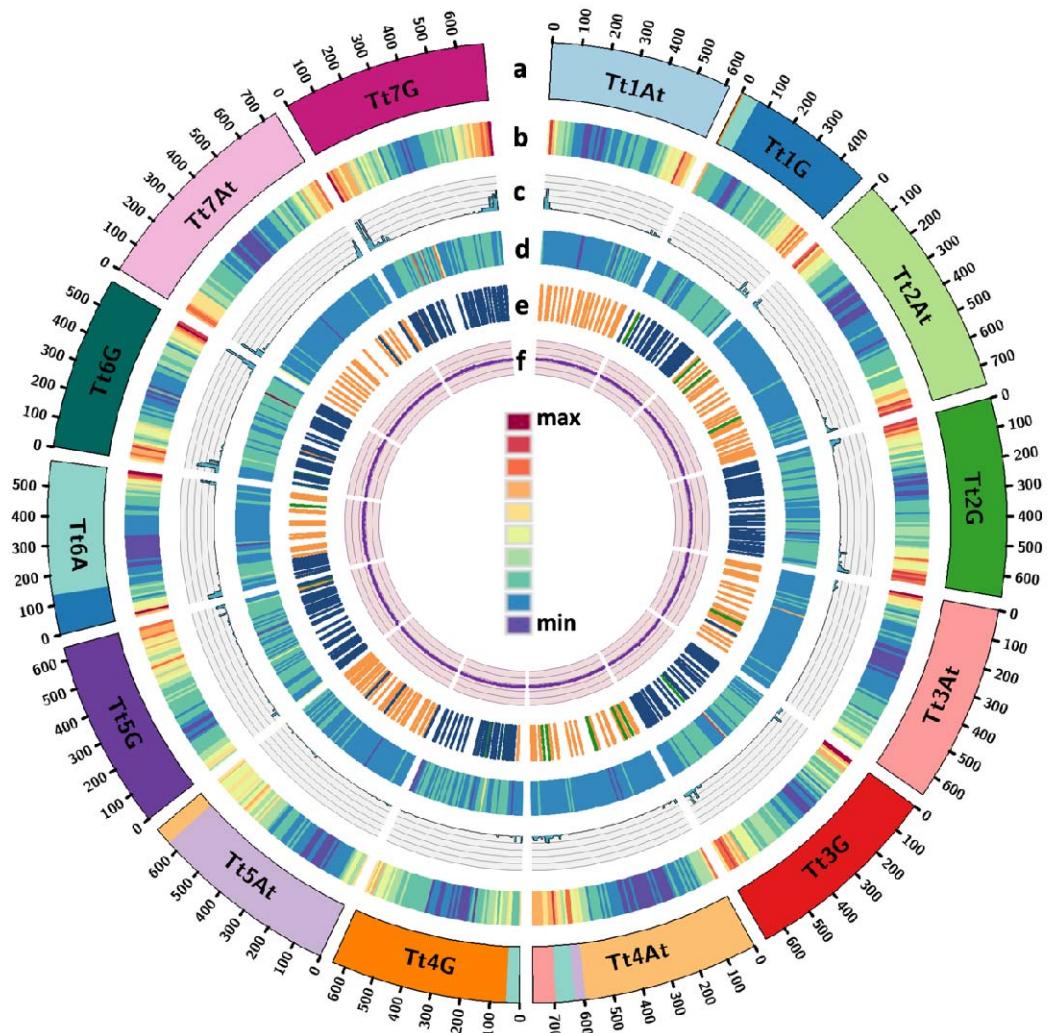
363

364 Generation of PacBio DNA methylation profile

365

366 Methylation in CpG context was inferred with ccsmeth v0.3.2⁶⁹, using the kinetics data from PacBio
367 CCS subreads obtained during HMW DNA sequencing. The methylation prediction for CCS reads were
368 called using the model "model_ccsmeth_5mCpG_call_mods_attbigru2s_b21.v2.ckpt". The reads with
369 the MM+ML tags were aligned to the pseudomolecules in the *T. timopheevii* assembly using BWA

370 v0.7.17⁷⁰. The methylation frequency was calculated at genome level with the modbam files and the
371 aggregate mode of ccsmeth with the model
372 “model_ccsmeth_5mCpG_aggregate_attbigru_b11.v2p.ckpt”. The genomic distribution of 5mC
373 modifications across *T. timopheevii* (Fig. 2d) shows that G genome chromosomes have more
374 methylation with an average of ~401.8 Kb methylated bases per 10 Mb bin as compared to the A^t
375 genome chromosomes with an average of ~385.5 Kb per 10 Mb bin.
376



377
378
379 **Figure 2.** Circos plot⁷¹ of features of the chromosome-scale assembly of *T. timopheevii* showing (a)
380 major translocations with the *T. timopheevii* genome as observed through collinearity analysis
381 against *T. turgidum*, (b) gene density (of all gene models), (c) NLR density (max count 87), (d) DNA
382 methylation (5mC modification) density, (e) distribution of KASP markers based on SNPs with bread
383 wheat cv. Chinese Spring²⁹ and (f) GC content. Tt in chromosome name represents *T. timopheevii*.
384

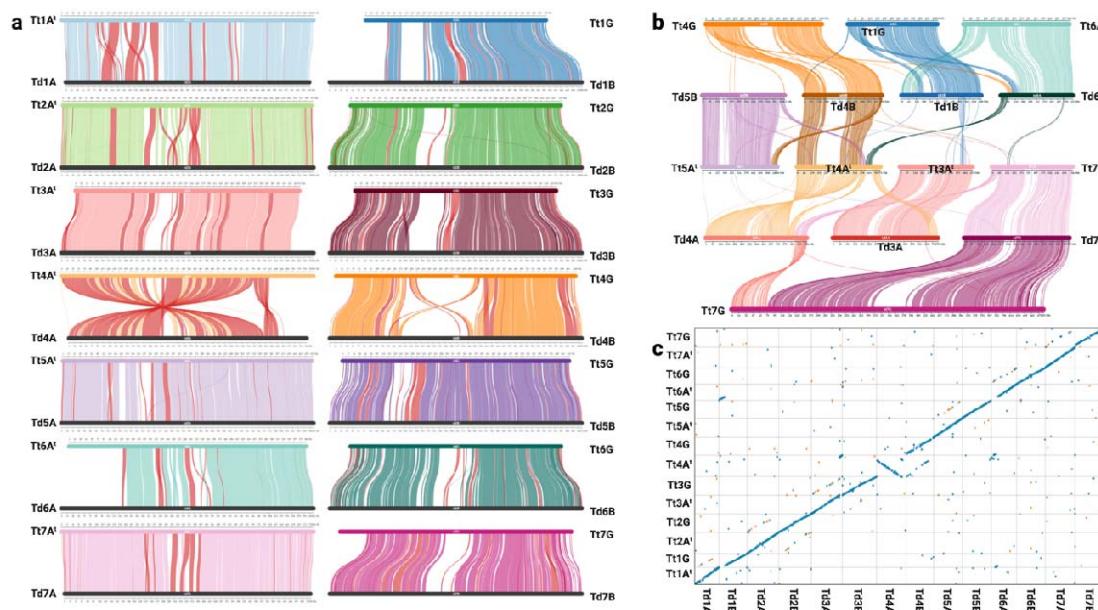
385 Comparative genome analysis

386

387 Synteny and collinearity analysis of the *T. timopheevii* gene set against the reference gene set of wild
388 emmer wheat *T. turgidum* accession Zavitan WEWSeq v1.0^{72,73} (available from Ensembl⁷⁴ plants) was

389 computed using MCScanX⁷⁵ with defaults parameters and results viewed using SynVisio^{76,77}
390 (<https://synvisio.github.io/>) and shown in Fig. 3a-b. A and G genome chromosomes of *T. timopheevii*
391 maintain synteny with the A and B genome chromosomes of tetraploid wheat albeit some inversions,
392 deletions and translocations shown in red. Fig. 3a. Analysis of large chromosome translocations
393 within the *T. timopheevii* genome confirmed previous reports⁷⁻⁹ of 5 translocation events including
394 T4A^tL/5A^tL (1), T6A^tS/1GS/4GS (2-4) and T4A^tL/3A^tL (5). Fig. 3b shows the composition of the
395 chromosomes involved (or suspected to be) in the translocation events as compared to the
396 composition of homoeologous chromosomes in tetraploid wheat (also depicted in Fig. 2a). It shows
397 that Chr4GS had retained a part of Chr6A^tS during the fourth reciprocal translocation event between
398 T4A^tL/4GS⁸ unlike previous reports that indicated that all of Chr6A^tS was translocated to Chr4A^tL. It
399 was also confirmed that unlike tetraploid (and hexaploid) wheat there is no inversion of Chr4A^tL (also
400 shown in Fig. 3c) and no reciprocal translocation between Chr7G and Chr4A^tL^{78,79} indicating that
401 although the T4AL/5AL was inherited from *T. urartu*^{8,80}, the following inversion of Chr4AL and
402 translocation with Chr7B were specific to the tetraploid and hexaploid wheat species.

403
404 Dotplot comparison of sequence alignments (as described earlier) between the *T. timopheevii*
405 pseudomolecules and the reference genome of *T. turgidum* accession Zavitan^{72,73} WEWSeq v1.0 also
406 confirmed the synteny, collinearity and translocations (Fig. 3c) as observed by comparing the gene
407 sets between these two species (Fig. 3a-b).
408

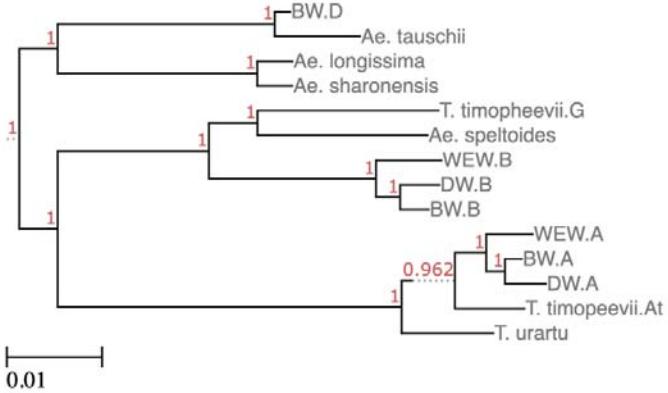


409
410
411 **Figure 3.** Comparative analysis of *T. timopheevii* (Tt) and *T. turgidum* (Td) genomes. (a) SynVisio plots
412 showing synteny and collinearity between the two genomes with rearrangements in red, (b) SynVisio
413 plots showing major translocations within the *T. timopheevii* chromosomes as compared to
414 tetraploid wheat, and (c) Dotplot comparison of the sequence alignments between the
415 chromosomes of the two genomes.

416 417 **Phylogeny analysis**

418
419 Orthofinder⁸¹ (<https://github.com/davidemms/OrthoFinder>) was used to locate orthologous genes
420 between *T. timopheevii* (Tt) and other *Aegilops* and wheat annotations using protein coding genes
421 (HC + LC). We used the S genome annotations from three *Aegilops* species⁸²: *Ae. longissima*, *Ae.*
422 *speltooides* and *Ae. sharonensis*, the A genome annotation of *T. urartu*, the D genome annotation of

423 Ae.*tauschii*, the AB annotation of wild emmer wheat (WEW) accession Zavitan⁷² and durum wheat
424 (DW) cv. Svevo⁸³ and the ABD annotation of hexaploid bread wheat (BW) cv. Chinese Spring⁶⁰
425 (available from Ensembl plants). The polyploid annotations (*T. timopheevii*, WEW, DW and BW) were
426 split into the subgenomes, and each was handled separately. The species tree (Fig. 4) was viewed
427 using the ETE Toolkit tree viewer⁸⁴ (available at <http://etetoolkit.org/treeview/>) and confirms that
428 the G genome of *T. timopheevii* is more closely related to the S genome of *Ae. speltoides* than the B
429 genomes of tetraploid and hexaploid wheats.



430
431
432 **Figure 4.** Phylogenetic tree based on orthofinder analysis of all protein coding genes. Branch values
433 in red correspond to orthofinder support values. BW, bread wheat cv. Chinese Spring; DW durum
434 wheat cv. Svevo; WEW, wild emmer wheat accession Zavitan.

435 436 **Genome visualisation**

437
438 A genome browser for the assembly of *T. timopheevii* generated in this study is currently being
439 hosted at GrainGenes⁸⁵ (<https://wheat.pw.usda.gov/jb?data=ggds/whe-timopheevii>) with tracks for
440 annotated gene models and repeats and BLAST functionality available at
441 <https://wheat.pw.usda.gov/blast/>.

442 443 **Data Records**

444
445 The raw sequence files for the HiFi, Hi-C, RNA-Seq and IsoSeq reads were deposited in the European
446 Nucleotide Archive (ENA) under accession number [PRJEB71660](https://www.ebi.ac.uk/ena/study/PRJEB71660). The final chromosome-scale
447 assembly consisting of the nuclear and organelle genomes was deposited at ENA under the accession
448 number GCA_963921465.2.

449
450 The genome assemblies, gene model and repeat annotations, methylation profile and Hi-C contact
451 map are also available at on DRYAD Digital Repository⁸⁶ (<https://doi.org/10.5061/dryad.mpg4f4r6p>).

452 453 **Technical Validation**

454 455 **Assessment of genome assembly and annotation**

456
457 The final curated assembly was assessed by mapping the trimmed Hi-C reads to the post-curated
458 assembly (as described above for scaffolding) and generating a final Hi-C contact map using
459 PretextMap v0.1.9 and viewed using PretextView v0.2.5. It showed a dense dark blue pattern along
460 the diagonal revealing no potential mis-assemblies (Fig. 5). The anti-diagonals in the Hi-C contact
461 matrix were expected and have been reported for other relatively large plant genomes such as those

462 from the Triticeae tribe^{87,88} as they correspond to the typical Rabl configuration of Triticeae
463 chromosomes^{89,90}.

464

465 The BUSCO v5.3.2⁴² (-l embryophyta_odb10) score of 98.9% (0.6% fragmented and 0.5% missing
466 BUSCOs; Table S8) at the genome level indicates a high completeness of the *T. timopheevii* assembly.
467 The quality of the *T. timopheevii* assembly was assessed with Merqury⁹¹ based on the PacBio HiFi
468 reads using 31-mers. The QV (consensus quality value) and k-mer completeness scores were 65.5
469 and 97.8%, respectively.

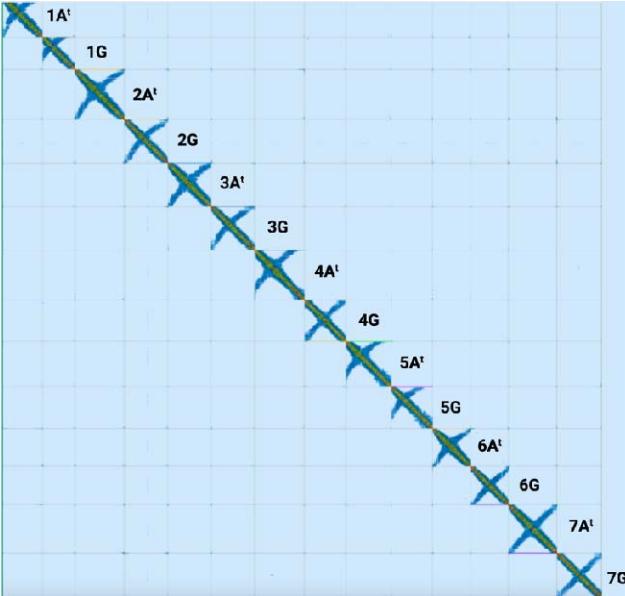
470

471 Completeness of the gene model prediction was also evaluated using BUSCO and produced a score
472 of 99.7% (0.1% fragmented and 0.2% missing BUSCOs; Table S8). The number of HC gene models
473 (70,365) is in the range of a tetraploid Triticeae species (34,000–43,000 high-confidence gene models
474 per haploid genome)⁹².

475

476 Of the total 14 chromosomes, we found telomeric repeats on both ends for 5 chromosomes (1A^t, 2G,
477 3A^t, 6A^t, and 7G) and on one end for 7 chromosomes (1GL, 2A^tS, 3GL, 4GS, 5GL, 6GL and 7A^tL).

478



479

480

481 **Figure 5.** Contact map after the integration of the Hi-C data and manual correction using
482 PretextView.

483

484 **Code availability**

485 All software and pipelines were executed according to the manual and protocol of published tools.
486 No custom code was generated for these analyses.

487

488 **Acknowledgements**

489 This work was supported by the Biotechnology and Biological Sciences Research Council [grant
490 number BB/P016855/1] as part of the Developing Future Wheat (DFW) programme. Part of this work
491 was also delivered via Transformative Genomics the BBSRC funded National Bioscience Research
492 Infrastructure (BBS/E/ER/23NB0006) at Earlham Institute by members of the Genomics Pipelines and
493 Core Bioinformatics Groups. EY and TS were supported by the US. Department of Agriculture,
494 Agricultural Research Service, Project No. 2030-21000-056-00D.

495

496 **Author contributions**

497 SuG, JK and IK designed the study and obtained funding for it. CY, DuS and SA carried out plant
498 maintenance and nucleic acid extraction. SuG, MW and LY generated the genome assembly. SuG and
499 JC carried out manual curation of the assembly. SrG and DaS carried out the genome annotation. EY
500 and TS generated the genome browser. SuG wrote the initial manuscript. All authors have read and
501 approved the final manuscript.

502

503 **Competing interests**

504 The authors declare no competing interests.

505

506 **References**

507

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

510 2. Dvorak, J., Zhang, H.-B. Variation in repeated nucleotide sequences sheds light on the phylogeny of
511 the wheat B and G genomes. *Proceedings of the National Academy of Sciences* **87**, 9640-9644 (1990).

512 3. Ahmed, H. I., et al. Einkorn genomics sheds light on history of the oldest domesticated wheat.
513 *Nature* **620**, 830-838 (2023).

514 4. Rodriguez, S., Maestra, B., Perera, E., Diez, M., Naranjo, T. Pairing affinities of the B-and G-genome
515 chromosomes of polyploid wheats with those of *Aegilops speltoides*. *Genome* **43**, 814-819 (2000).

516 5. Li, L. F., et al. Genome sequences of five *Sitopsis* species of *Aegilops* and the origin of polyploid
517 wheat B subgenome. *Molecular plant* **15**, 488-503 (2022).

518 6. Dvořák, J. *Triticum Species (Wheat)*. *Encyclopedia of Genetics*, 2060–2068 (2001).

519 7. Jiang, J., Gill, B. S. Different species-specific chromosome translocations in *Triticum timopheevii* and
520 *T. turgidum* support the diphylectic origin of polyploid wheats. *Chromosome Research* **2**, 59-64 (1994).

521 8. Maestra, B., Naranjo, T. Structural chromosome differentiation between *Triticum timopheevii* and
522 *T. turgidum* and *T. aestivum*. *Theoretical and Applied Genetics* **98**, 744-750 (1999).

523 9. Rodriguez, S., Perera, E., Maestra, B., Díez, M., Naranjo, T. Chromosome structure of *Triticum*
524 *timopheevii* relative to *T. turgidum*. *Genome* **43**, 923-930 (2000).

525 10. Devi, U., et al. Development and characterisation of interspecific hybrid lines with genome-wide
526 introgressions from *Triticum timopheevii* in a hexaploid wheat background. *BMC Plant Biol* **19**, 183
527 (2019).

528 11. Brown-Guedira, G. L., Singh, S., Fritz, A. K. Performance and Mapping of Leaf Rust Resistance
529 Transferred to Wheat from *Triticum timopheevii* subsp. *armeniacum*. *Phytopathology* **93**, 784-789
530 (2003).

531 12. Singh, A. K., et al. Genetics and mapping of a new leaf rust resistance gene in *Triticum aestivum* L.
532 × *Triticum timopheevii* Zhuk. derivative 'Selection G12'. *J Genet* **96**, 291-297 (2017).

533 13. Leonova, I. N., et al. Microsatellite mapping of a leaf rust resistance gene transferred to common
534 wheat from *Triticum timopheevii*. *Cereal Research Communications* **38**, 211-219 (2010).

535 14. McIntosh, R., Gyarfas, J. *Triticum timopheevii* as a source of resistance to wheat stem rust.
536 *Zeitschrift fur Pflanzenzuchtung* **66**, 240-248 (1971).

537 15. Wu, S., Pumphrey, M., Bai, G. Molecular Mapping of Stem-Rust-Resistance Gene Sr40 in Wheat.
538 *Crop Science* **49**, 1681-1686 (2009).

539 16. Allard, R., Shands, R. Inheritance of resistance to stem rust and powdery mildew in cytologically
540 stable spring wheats derived from *Triticum timopheevii*. *Phytopathology* **44**, 266-274 (1954).

541 17. Perugini, L. D., Murphy, J. P., Marshall, D., Brown-Guedira, G. Pm37, a new broadly effective
542 powdery mildew resistance gene from *Triticum timopheevii*. *Theoretical and Applied Genetics* **116**,
543 417-425 (2008).

544 18. Qin, B., et al. Collinearity-based marker mining for the fine mapping of Pm6, a powdery mildew
545 resistance gene in wheat. *Theoretical and Applied Genetics* **123**, 207-218 (2011).

546 19. Steed, A., et al. Identification of Fusarium Head Blight Resistance in *Triticum timopheevii*
547 Accessions and Characterization of Wheat-T. *timopheevii* Introgression Lines for Enhanced
548 Resistance. *Frontiers in Plant Science* **13**, (2022).

549 20. Malihipour, A., Gilbert, J., Fedak, G., Brûlé-Babel, A., Cao, W. Characterization of agronomic traits
550 in a population of wheat derived from *Triticum timopheevii* and their association with Fusarium head
551 blight. *European Journal of Plant Pathology* **144**, 31-43 (2016).

552 21. Brown-Guedira, G., et al. Evaluation of a collection of wild *timopheevi* wheat for resistance to
553 disease and arthropod pests. *Plant disease* **80**, 928-933 (1996).

554 22. Badridze, G., Weidner, A., Asch, F., Börner, A. Variation in salt tolerance within a Georgian wheat
555 germplasm collection. *Genetic resources and crop evolution* **56**, 1125-1130 (2009).

556 23. Yudina, R., Leonova, I., Salina, E., Khlestkina, E. Change in salt tolerance of bread wheat as a
557 result of the introgression of the genetic material of *Aegilops speltoides* and *Triticum timopheevii*.
558 *Russian Journal of Genetics: Applied Research* **6**, 244-248 (2016).

559 24. Lehmensiek, A., Bovill, W., Banks, P., Sutherland, M. Molecular characterization of a *Triticum*
560 *timopheevii* introgression in a Wentworth/Lang population. (2008).

561 25. Hu, X., et al. Zn and Fe concentration variations of grain and flag leaf and the relationship with
562 *NAM-G1* gene in *Triticum timopheevii* (Zhuk.) Zhuk. ssp. *timopheevii*. *Cereal Research
563 Communications* **45**, 421-431 (2017).

564 26. Walkowiak, S., et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature*
565 **588**, 277-283 (2020).

566 27. Keilwagen, J., et al. Detecting major introgressions in wheat and their putative origins using
567 coverage analysis. *Scientific Reports* **12**, 1908 (2022).

568 28. Keilwagen, J., et al. Finding needles in a haystack: identification of inter-specific introgressions in
569 wheat genebank collections using low-coverage sequencing data. *Frontiers in Plant Science* **14**,
570 (2023).

571 29. King, J., et al. Introgression of the *Triticum timopheevii* Genome Into Wheat Detected by
572 Chromosome-Specific Kompetitive Allele Specific PCR Markers. *Frontiers in Plant Science* **13**, (2022).

573 30. Grewal, S., et al. Rapid identification of homozygosity and site of wild relative introgressions in
574 wheat through chromosome-specific KASP genotyping assays. *Plant Biotechnol J* **18**, 743-755 (2020).

575 31. Belton, J. M., et al. Hi-C: a comprehensive technique to capture the conformation of genomes.
576 *Methods* **58**, 268-276 (2012).

577 32. Wenger, A. M., et al. Accurate circular consensus long-read sequencing improves variant
578 detection and assembly of a human genome. *Nature Biotechnology* **37**, 1155-1162 (2019).

579 33. Drieguez, P., et al. LeafGo: Leaf to Genome, a quick workflow to produce high-quality de novo
580 plant genomes using long-read sequencing technology. *Genome biology* **22**, 256 (2021).

581 34. Dong, L., et al. Single-molecule real-time transcript sequencing facilitates common wheat
582 genome annotation and grain transcriptome research. *BMC Genomics* **16**, 1039 (2015).

583 35. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011*
584 **17**, 3 (2011).

585 36. Bolger, A. M., Lohse, M., Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data.
586 *Bioinformatics* **30**, 2114-2120 (2014).

587 37. Marçais, G., Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences
588 of k-mers. *Bioinformatics* **27**, 764-770 (2011).

589 38. Wang, H., et al. Estimation of genome size using k-mer frequencies from corrected long reads.
590 *arXiv:200311817 [q-bioGN]*, (2020).

591 39. Ranallo-Benavidez, T. R., Jaron, K. S., Schatz, M. C. GenomeScope 2.0 and Smudgeplot for
592 reference-free profiling of polyploid genomes. *Nature Communications* **11**, 1432 (2020).

593 40. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., Li, H. Haplotype-resolved de novo assembly
594 using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170-175 (2021).

595 41. Formenti, G., et al. Gfastats: conversion, evaluation and manipulation of genome sequences
596 using assembly graphs. *Bioinformatics* **38**, 4214-4216 (2022).

597 42. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., Zdobnov, E. M. BUSCO: assessing
598 genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-
599 3212 (2015).

600 43. Laetsch, D., Blaxter, M. BlobTools: Interrogation of genome assemblies. *F1000Research* **6**, (2017).

601 44. Korbel, J. O., Lee, C. Genome assembly and haplotyping with Hi-C. *Nature Biotechnology* **31**,
602 1099-1101 (2013).

603 45. Ghurye, J., et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly.
604 *PLOS Computational Biology* **15**, e1007273 (2019).

605 46. Howe, K., et al. Significantly improving the quality of genome assemblies through curation.
606 *GigaScience* **10**, (2021).

607 47. Zhu, T., et al. Optical maps refine the bread wheat *Triticum aestivum* cv. Chinese Spring genome
608 assembly. *The Plant Journal* **107**, 303-314 (2021).

609 48. Kurtz, S., et al. Versatile and open software for comparing large genomes. *Genome biology* **5**, R12
610 (2004).

611 49. Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L., Swarbreck, D. Leveraging multiple
612 transcriptome assembly methods for improved gene structure annotation. *GigaScience* **7**, (2018).

613 50. Boden, S. A., et al. Updated guidelines for gene nomenclature in wheat. *Theoretical and Applied
614 Genetics* **136**, 72 (2023).

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

617 52. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100
618 (2018).

619 53. Mapleson, D., Venturini, L., Kaithakottil, G., Swarbreck, D. Efficient and accurate detection of
620 splice junctions from RNA-seq with Portcullis. *GigaScience* **7**, (2018).

621 54. Kovaka, S., et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2.
622 *Genome biology* **20**, 278 (2019).

623 55. Shao, M., Kingsford, C. Accurate assembly of transcripts through phase-preserving graph
624 decomposition. *Nature Biotechnology* **35**, 1167-1169 (2017).

625 56. Gotoh, O. A space-efficient and accurate method for mapping and aligning cDNA sequences onto
626 genomic sequence. *Nucleic Acids Research* **36**, 2630-2638 (2008).

627 57. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, (2023).

628 58. Stanke, M., Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that
629 allows user-defined constraints. *Nucleic Acids Research* **33**, W465-W467 (2005).

630 59. Haas, B. J., et al. Automated eukaryotic gene structure annotation using EVidenceModeler and
631 the Program to Assemble Spliced Alignments. *Genome biology* **9**, R7 (2008).

632 60. IWGSC, et al. Shifting the limits in wheat research and breeding using a fully annotated reference
633 genome. *Science* **361**, (2018).

634 61. Shumate, A., Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**,
635 1639-1643 (2021).

636 62. Seppey, M., Manni, M., Zdobnov, E. M. in *Gene Prediction: Methods and Protocols* (ed. Kollmar
637 M) BUSCO: Assessing Genome Assembly and Annotation Completeness (Springer New York, 2019).

638 63. Kong, L., et al. CPC: assess the protein-coding potential of transcripts using sequence features
639 and support vector machine. *Nucleic Acids Research* **35**, W345-W349 (2007).

640 64. Bray, N. L., Pimentel, H., Melsted, P., Pachter, L. Near-optimal probabilistic RNA-seq
641 quantification. *Nature Biotechnology* **34**, 525-527 (2016).

642 65. Consortium, U. UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204-212 (2015).

643 66. Jones, P., et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**,
644 1236-1240 (2014).

645 67. Kourelis, J., Van Der Hoorn, R. A. Defended to the nines: 25 years of resistance gene cloning
646 identifies nine mechanisms for R protein function. *The Plant cell* **30**, 285-299 (2018).

647 68. Chen, R., Gajendiran, K., Wulff, B. B. H. R we there yet? Advances in cloning resistance genes for
648 engineering immunity in crop plants. *Current opinion in plant biology* **77**, 102489 (2024).

649 69. Ni, P., et al. DNA 5-methylcytosine detection and methylation phasing using PacBio circular
650 consensus sequencing. *Nature communications* **14**, 4054 (2023).

651 70. Li, H., Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.
652 *bioinformatics* **25**, 1754-1760 (2009).

653 71. Krzywinski, M., et al. Circos: An information aesthetic for comparative genomics. *Genome Res* **19**,
654 1639-1645 (2009).

655 72. Avni, R., et al. Wild emmer genome architecture and diversity elucidate wheat evolution and
656 domestication. *Science* **357**, 93-97 (2017).

657 73. Zhu, T., et al. Improved Genome Sequence of Wild Emmer Wheat Zavitan with the Aid of Optical
658 Maps. *G3 (Bethesda)* **9**, 619-624 (2019).

659 74. Martin, F. J., et al. Ensembl 2023. *Nucleic Acids Research* **51**, D933-D941 (2022).

660 75. Wang, Y., et al. MCSanX: a toolkit for detection and evolutionary analysis of gene synteny and
661 collinearity. *Nucleic acids research* **40**, e49-e49 (2012).

662 76. Bandi, V., Gutwin, C. Interactive Exploration of Genomic Conservation. In *Proceedings of 46th*
663 *Graphics Interface Conference* (Canadian Human-Computer Communications Society, 2020).

664 77. Bandi, V., et al. in *Plant Bioinformatics: Methods and Protocols* (ed. Edwards D) Visualization Tools
665 for Genomic Conservation (Springer US, 2022).

666 78. Devos, K. M., Dubcovsky, J., Dvorak, J., Chinoy, C. N., Gale, M. D. Structural evolution of wheat
667 chromosomes 4A, 5A, and 7B and its impact on recombination. *Theor Appl Genet* **91**, 282-288
668 (1995).

669 79. Dvorak, J., et al. Reassessment of the evolution of wheat chromosomes 4A, 5A, and 7B. *Theor
670 Appl Genet* **131**, 2451-2462 (2018).

671 80. King, I. P., et al. Detection of interchromosomal translocations within the Triticeae by RFLP
672 analysis. *Genome* **37**, 882-887 (1994).

673 81. Emms, D., Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics.
674 bioRxiv 466201, 2019.

675 82. Avni, R., et al. Genome sequences of three Aegilops species of the section Sitopsis reveal
676 phylogenetic relationships and provide resources for wheat improvement. *The Plant Journal* **110**,
677 179-192 (2022).

678 83. Maccaferri, M., et al. Durum wheat genome highlights past domestication signatures and future
679 improvement targets. *Nature Genetics* **51**, 885-895 (2019).

680 84. Huerta-Cepas, J., Serra, F., Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of
681 Phylogenomic Data. *Molecular Biology and Evolution* **33**, 1635-1638 (2016).

682 85. Yao, E., et al. GrainGenes: a data-rich repository for small grains genetics and genomics. *Database*
683 **2022**, (2022).

684 86. Grewal, S., et al. 2024. *Triticum timopheevii* genome assembly files [Dataset]. Dryad.
685 <https://doi.org/10.5061/dryad.mpg4f4r6p>

686 87. Dong, P., et al. 3D chromatin architecture of large plant genomes determined by local A/B
687 compartments. *Molecular plant* **10**, 1497-1509 (2017).

688 88. Mascher, M., et al. A chromosome conformation capture ordered sequence of the barley
689 genome. *Nature* **544**, 427-433 (2017).

690 89. Anamthawat-Jónsson, K., Heslop-Harrison, J. Centromeres, telomeres and chromatin in the
691 interphase nucleus of cereals. *Caryologia* **43**, 205-213 (1990).

692 90. Cowan, C. R., Carlton, P. M., Cande, W. Z. The polar arrangement of telomeres in interphase and
693 meiosis. Rabl organization and the bouquet. *Plant Physiology* **125**, 532-538 (2001).

694 91. Rhie, A., Walenz, B. P., Koren, S., Phillippy, A. M. Merqury: reference-free quality, completeness,
695 and phasing assessment for genome assemblies. *Genome biology* **21**, 1-27 (2020).

696 92. Poretti, M., Praz, C. R., Sotiropoulos, A. G., Wicker, T. A survey of lineage-specific genes in
697 Triticeae reveals de novo gene evolution from genomic raw material. *Plant Direct* **7**, e484 (2023).

698