1 Chromosome-scale genome assembly
2 provides insights into rye biology,
3 evolution, and agronomic potential

4

5 M. Timothy Rabanus-Wallace, Bernd Hackauf, Martin Mascher, Thomas Lux, Thomas Wicker, Heidrun

6 Gundlach, Mariana Báez, Andreas Houben, Klaus F.X. Mayer, Liangliang Guo, Jesse Poland, Curtis J.

7 Pozniak, Sean Walkowiak, Joanna Melonek, Coraline Praz, Mona Schreiber, Hikmet Budak, Matthias

8 Heuberger, Burkhard Steuernagel, Brande Wulff, Andreas Börner, Brook Byrns, Jana Čížková, D. Brian

9 Fowler, Allan Fritz, Axel Himmelbach, Gemy Kaithakottil, Jens Keilwagen, Beat Keller, David Konkin,

10 Jamie Larsen, Qiang Li, Beata Myśków, Sudharsan Padmarasu, Nidhi Rawat, Uğur Sesiz, Biyiklioglu Sezgi,

11 Andy Sharpe, Hana Šimková, Ian Small, David Swarbreck, Helena Toegelová, Natalia Tsvetkova, Anatoly

12 V. Voylokov, Jan Vrána , Eva Bauer, Hanna Bolibok-Bragoszewska, Jaroslav Doležel, Anthony Hall, Jizeng

13 Jia , Viktor Korzun, André Laroche, Xue-Feng Ma, Frank Ordon, Hakan Özkan, Monika Rakoczy-

14 Trojanowska, Uwe Scholz, Alan H. Schulman, Dörthe Siekmann, Stefan Stojałowski, Vijay Tiwari, Manuel

15 Spannagl, Nils Stein

16   ## Abstract [106 / about 100 words]

17   *We present a chromosome-scale annotated assembly of the rye (Secale cereale L. inbred line 'Lo7')*

18   *genome, which we use to explore Triticeae genomic evolution, and rye's superior disease and stress*

19   *tolerance. The rye genome shares chromosome-level organization with other Triticeae cereals, but*

20   *exhibits unique retrotransposon dynamics and structural features. Crop improvement in rye, as well as in*

21   *wheat and triticale, will profit from investigations of rye gene families implicated in pathogen resistance,*

22   *low temperature tolerance, and fertility control systems for hybrid breeding. We show that rye*

23   *introgressions in wheat breeding panels can be characterised in high-throughput to predict the yield*

24   *effects and trade-offs of rye chromatin.*

25 **Main Text [3979/4000 words (excl methods, captions); 7/8 visual**
## 26 **items]**

27 Rye (*Secale cereale* L.) is a member of the grass tribe Triticeae and close relative of wheat (*Triticum*

28 *aestivum* L.) and barley (*Hordeum vulgare* L.), grown primarily for human consumption and animal feed.

29 Rye is uniquely tolerant of biotic and abiotic stresses and thus exhibits high yield potential under

30 marginal conditions. This makes rye an important crop along the northern boreal-hemiboreal belt, a

31 climatic zone predicted to expand considerably in Eurasia and North America with anthropogenic global

32 warming[1]. Rye chromatin introgressions into bread wheat can significantly increase yield by conferring

33 disease resistance and enhanced root biomass[2-5]. Rye also possesses a unique bi-factorial self-

34 incompatibility system[6], and rye genes controlling self-compatibility and male fertility have enabled the

35 establishment of efficient cytoplasmic male sterility (CMS)-based hybrid breeding systems that exploit

36 heterosis at large scales[7]. Implementation of such systems in cereals will be invaluable to meeting future

37 human calorific requirements.

38 Rye is diploid with a large genome (~7—8 Gbp)[8] compared even to the diploid barley genome and the

39 subgenomes of the hexaploid bread wheat[9]. Like barley and wheat, rye entered the genomics era very

40 recently. A virtual gene-order was released in 2013[10], and a shotgun *de novo* genome survey of the same

41 line became available in 2017[11]. Both resources have been rapidly adopted by researchers and

42 breeders[12-14], but cannot offer the same opportunities as the higher quality genome assemblies available

43 for other Triticeae species[9,15-19].

44 We report the assembly of a chromosome-scale genome sequence for rye line 'Lo7', providing insights

45 into rye genome organisation and evolution, and representing a comprehensive resource for genomics-

46 assisted crop improvement.

47 *Results*

48 **An annotated chromosome-scale genome assembly**
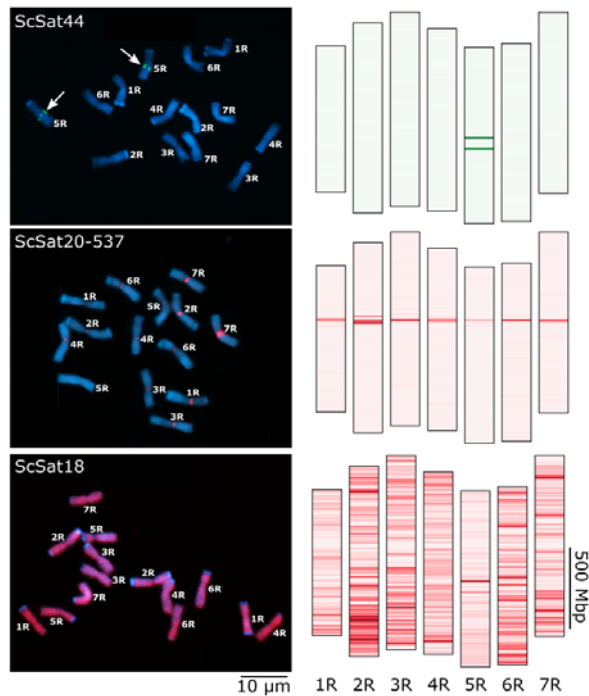
49
50 **Figure M-FISH.** FISH of mitotic rye chromosomes with ScSat44, ScSat18 and ScSat20-537-specific probes (left) and *in silico* predicted repeat
51 distribution (right), showing agreement between real and predicted hybridization sites. Chromosomes are counterstained with DAPI (blue),
52 ScSat44 in green (chromosome 5R is arrowed), ScSat18 and ScSat20-537 in red. Arrows mark chromosome-specific binding of ScSat44 to
53 chromosome 5R. Darkness is scaled evenly between the maximum and minimum densities of each repeat across all assembled chromosomes
54 (Methods).
55

56 We estimated the genome size of 15 rye genotypes by flow-cytometry (Methods, Note S-FLOWCYT) and

57 found 'Lo7' among the smaller of these at 7.9 Gbp. We *de novo* assembled scaffolds representing 6.74

58 Gbp of the 'Lo7' genome (Table 1) from >1.8 Tbp of short read sequence (Methods; Notes S-PSASS, S-

59 ASSDATA). The scaffolds were ordered, oriented and curated using a variety of independent data

60 sources including: (i) chromosome-specific shotgun (CSS) reads[10], (ii) 10X Chromium linked reads, (iii)

61 genetic map markers[11], (iv) 3D chromosome conformation capture sequencing (Hi-C)[20], and (v) a

62 Bionano optical genome map (tbls. S-ASSSTATS—S-OPTSTAT). After intensive manual curation, 83% of

63 this assembled sequence (i.e. ~75.5% of the total genome size) was arranged first into super-scaffolds

64 (N50 >29 Mbp) and then into pseudomolecules. Annotation of various features (Methods) yielded

65 34,441 high confidence genes, which we estimate comprises 97.9% of the entire gene complement (tbl.

66   S-ANNOTSTAT), 19,456 full-length DNA LTR retrotransposons (LTR-RTs) from six transposon families (tbl.

67   S-TEANNOT)[21], 13,238 putative miRNAs in 90 miRNA families (tbls. S-miRNA_sequences—S-miRNA

68   target_table), and 1,382,323 tandem repeat arrays (tbls. S-TANDREPCOMPN-S-SAT_ANNOT).

69   Fluorescence *in situ* hybridisation (FISH) to mitotic 'Lo7' chromosomes using probes targeting tandem

70   repeats showed that scaffolds for which assignment to a chromosome pseudomolecule was difficult are

71   highly enriched in short repeats (Methods; Note S-REP).

Gene collinearity among the Triticeae

73   We used the assembly to closely assess gene-level collinearity between rye, barley and bread wheat

74   (Methods; figs. M-TRACKSa, Note S-COLLIN)[9-11,15,22-24]. As previously reported, Triticeae chromosome

75   groups 1–3 appear essentially collinear across all three species[9,10,15]. Rearrangements such as those

76   between 4R and 7R are observable at high resolution, along with several inversions (e.g. on 1RL and 3RL;

77   fig. M-TRACKSa). Rearrangements affecting subtelomeres were reflected in the absence of hybridisation

78   signals from two subtelomere-specific FISH probes developed in this study (Note S-FISH; tbl. S-FISH).

79   Regions of rye-barley collinearity contrast with distinct low-collinearity 'modules' (henceforth denoted

80   LCMs) that surround the centromeres of chromosomes. Such regions, in which enough gene synteny is

81   conserved to demonstrate identity by descent but the order of orthologs significantly differs among

82   relatives, can now be observed in the sequenced genomes of many species[25,26] (figs. M-TRACKSa; Note

83   S-COLLIN). While centromeres can suffer from assembly difficulties, the LCM boundaries extend well

84   into the pericentromeres, and on several chromosomes occur within large scaffolds validated by

85   multiple sources of data including optical maps. The LCMs of rye, wheat, and barley differ in length, but

86   curiously (i) the sets of genes that fall inside and outside the LCMs are almost the same in all three

87   species, (ii) The LCMs distinctly correlate with regions of low gene density (fig. M-TRACKSb), and (iii)

88   possess a distinct and characteristic repetitive element population (figs M-TRACKSd-g, Note S-REP). We

89   explore these observations in more detail below.

Assembly		Raw scaffolds (after chimera breaking)	In chromosome-scale pseudomolecules	
Scaffolds		109,776	476	
Total length (Mbp)		6,670.03	6,206.74	
N50 length (Mbp)		15.16	29.44	
% length with chromosome assignment		95.3%	100%	
Optical genome map				
Maps		5,601		
Total length (Mbp)		6,660.18		
N50 length (Mbp)		1.671		
Total aligned length (Mbp)		6,248.60		
Uniquely aligned length (Mbp)		6,029.11		
Gene feature annotation		High confidence set	Low confidence set	
Number of genes		34,441	22,781	
Mean gene length		2,892	946	
Mean exons per gene		4.42	1.79	
Proportion of complete BUSCO set		96.4%	5.8%	
LTR-RT annotation		Superfamily	Full-length copies	Mean age (Mya)
RLC_Angela		Copia	11,128	0.53
RLG_Cereba		Gypsy	934	1.24
RLG_Sabrina		Gypsy	3,996	2.10
RLG_WHAM		Gypsy	1,457	2.06
DTC_Clifford		CACTA	1,480	N.A.
DTC_Conan		CACTA	516	N.A.

90 **Table M-STATS.** Genome assembly and annotation statistics. CSS=Chromosome Specific Shotgun. BUSCO=Benchmarking universal single-copy

91 orthologs (v3; https://busco.ezlab.org/).

Evolutionary dynamics of the intergenic space

93 Transposable elements, especially long terminal repeat retrotransposons (LTR-RTs), exert a primary

94 influence on Triticeae genome structure and composition[27-29]. Full-length LTR-RTs represent the same

95 proportion of the total assembly size as exhibited by other major Triticeae reference assemblies (fig. S-

96 RPT_ASSCMP, tbl. S-TE_ASSCMP_ANNOTSTATS), indicating similar assembly completeness[30]. Past LTR-RT

97 activity can be inferred by estimating the insertion ages of individual LTR-RT elements, and the

98 evolutionary relationships among LTR-RT families (Methods; Note S-REP).
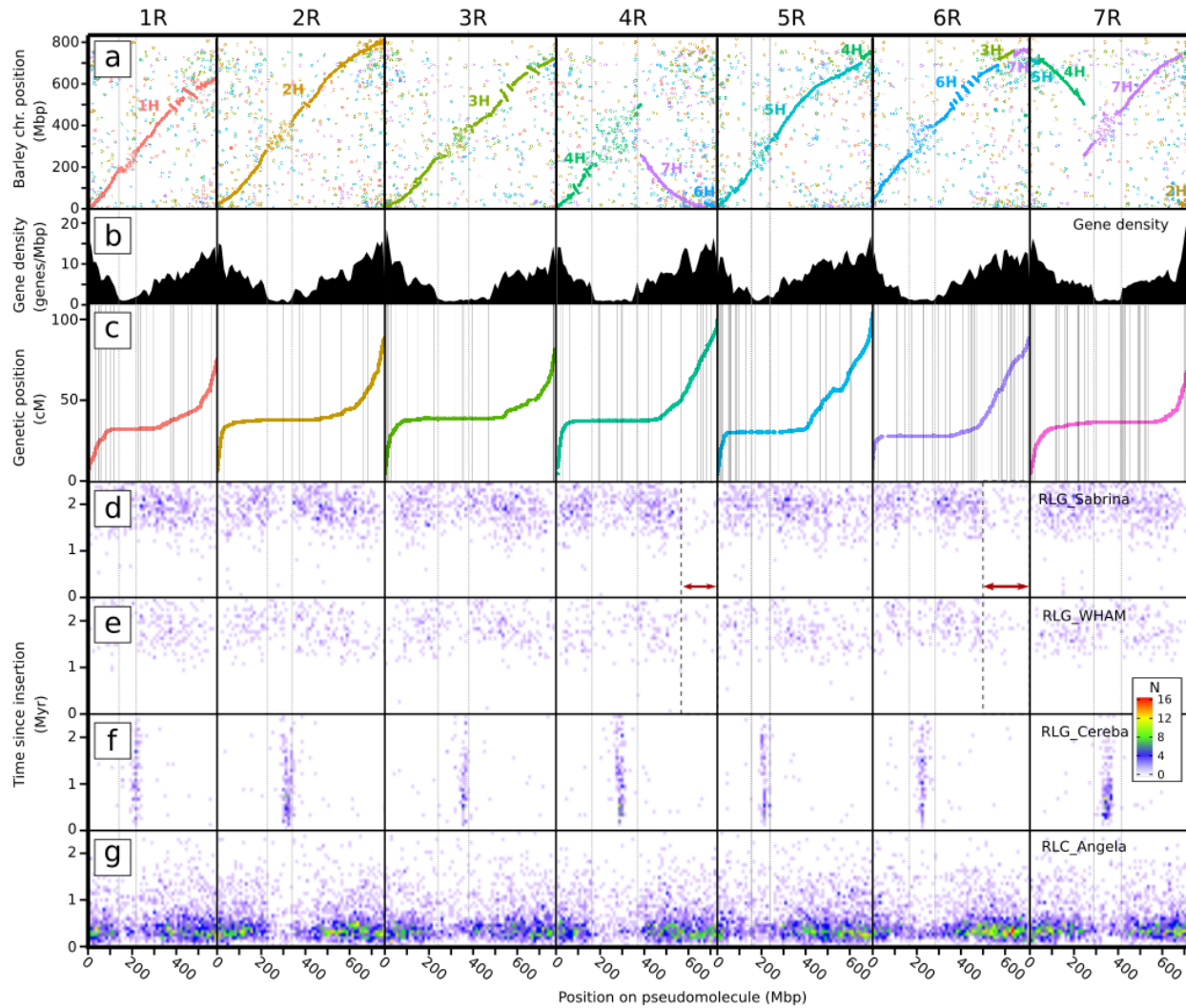
Figure M-TRACKS. Selected information tracks for 'Lo7' chromosomes 1R to 7R (left-to-right). Twin vertical grey lines in each chromosome denote the boundaries of the LCMs for each chromosome. **A)** Gene collinearity with barley (cv. Morex), with the position on the Morex pseudomolecules on the vertical axis. Text and point colours represent barley chromosomes as labelled. **B)** Density of annotated gene models. **C)** Genetic map positions of markers used in assembly. Scaffold boundaries marked by grey vertical lines. **D-G)** Positions and ages of four LTR retrotransposon families in the genome, represented as a heatmap. Binned ages are on the vertical axis (from 0 Mya at the bottom), and bin positions are across the horizontal. Heat represents the number of TEs in each age/position bin (see legend inset). Red arrows mark notable changes in LTR-RT profiles.

As in barley and wheat, rye LTR-RT show clear niche specialisation across genomic compartments [27,28](fig. M-TRACKSc—f; Note S-REP): *RLC_Sabrina*, *RLG_WHAM*, and *RLC_Angela* are depleted in centromeres and pericentromeres, with the depleted region normally corresponding closely to the LCMs

111 (fig. M-TRACKSb-f). *RLC_Cereba* strictly occupies centromeres[31]. The long arm termini of chromosomes

112 4RL and 6RL bear distinct tandem repeat (Note S-REP) and LTR-RT profiles (fig. M-TRACKSc,d; figs. S-

113 TETERMPROF, S-KMERREP): *DTC_Clifford* elements are two to four times more abundant than on the

114 long arm termini of the other chromosomes, while *RLG_Sabrina* and *RLG_WHAM* elements are almost

115 absent. We suspect such changes are most likely the result of ancestral chromosome arm translocations

116 from a close relative. In the case of 4RL the profile changes are particularly clear and we can ascertain

117 that: (i) since the altered TE profile boundaries do not coincide with a collinearity break with wheat or

118 barley (figs. M-TRACKSa, S-KMERREP; Note S-COLLIN), the donor is likely of rye lineage; (ii) since in the

119 donated segments, *DTC_Clifford* is more abundant than *RLG_Sabrina* and *RLG_WHAM*, the donors must

120 have diverged from the 'Lo7' ancestor prior to the expansion of the latter elements in earnest, around

121 3.5 Mya; and (iii) since the recent *RLC_Angela* expansion is recorded across 4R, the introgressions

122 occurred before its beginning around 1.8 Mya.

123

124 The timing of expansions differs markedly between LTR-RT families of the rye genome, demonstrating

125 that older families degrade as younger families expand. Repetitive insertion into the centromere

126 suggests a centromere-outwards chromosome expansion mechanism, as is most apparent for

127 chromosomes 2R, 4R, and 6R, by the distribution of older *Cereba* elements being more distant from the

128 centromere than the younger. Comparing rye with wheat and barley, the variously curved and straight

129 slopes of collinear runs of genes (Note S-COLLIN) suggest physical genome expansion acted quite

130 uniformly across the rye genome since its split from wheat. Conversely, the size changes that separate

131 rye from wheat and barley are pronounced near telomeres, indicating that genome expansion

132 mechanisms alter over million year timescales and likely contribute to both speciation and ancient

133 hybridisation events[32]. In rye, barley and in each individual wheat subgenome, the TE superfamilies

134 *Gypsy* (RLG) and *Copia* (RLC) expanded in the same order[27,28], but not at the same time: The *Gypsy*-to-

135   *Copia* progression was probably set in motion by the LTR-RT composition of a shared ancestral genome,

136   but the *rates* of expansion and suppression of each superfamily would have depended upon functional

137   and selective peculiarities of each genome or sub-genome (arguments expanded in note S-TEEXP).

138   **Structural variation and *Secale* genome evolution**

139
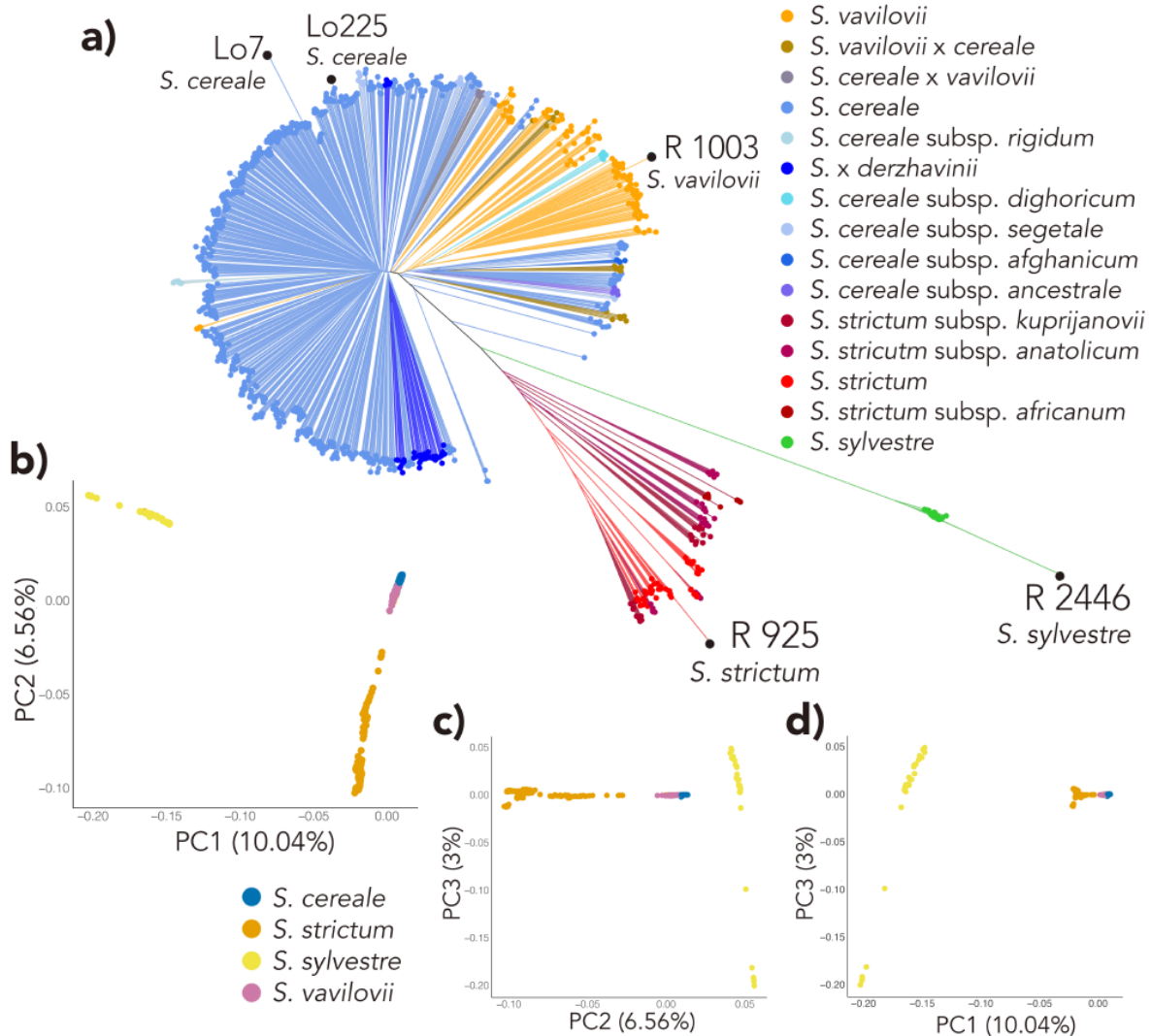
Figure M-PHYLO. Diversity and relationships among *Secale* taxa. The population structure corresponds to the structure of three taxa as presented in Schreiber et al. 2019 but gives a clearer grouping due to the additional wild accessions, especially with regard to S. *vavilovii*, the wild progenitor, which was previously indistinguishable from domesticated rye but is now forming a subgroup within *S. cereale*. **a)** Neighbour-joining tree, with taxonomic assignments to subspecies level, according to genebank passport data. **b—d)** The first three prinicipal components of genetic variance within the dataset, with samples coloured according to species.

147   The many Triticeae gene-collinearity disruptions observable as inversions and pericentromeric LCMs

148   suggest rapid accumulation of structural variations (SVs) that might segregate in rye populations causing

149   undesired linkage in breeding and mapping efforts. To investigate further, we used Hi-C data from single

150   individuals of four rye species to identify candidate SVs among *S. cereale* and three other *Secale* species.

151   We included a second *S. cereale* genotype, 'Lo225', an inbred line from which the mapping population

152   used for assembly was derived. To provide phylogenetic context, we extended the *Secale* phylogeny of

153   Schreiber et al. (2019)[33], adding 347 genotypes, and calling variants against the new genome assembly

154   (Methods; fig. M-PHYLO). Many inversions (>10) were observed to segregate among non-'Lo7' *Secale*

155   genotypes, making assembly artefacts a highly unlikely source of error (Note S-SV). One such 'Lo7'—

156   'Lo225' inversion on 5RL corresponds to a distinct local plateau in the genetic map (fig. M-SV),

157   representing complete linkage between the 382 annotated high confidence genes in this region. Rye

158   pericentromeres are especially prone to large-scale SVs ($p<0.001$; Note S-SV), in agreement with

159   previous findings[29,34]. This confirms SV as one possible mechanism for the formation of LCMs, and helps

160   to explain the lack of genes in these regions, since recombination-suppressed genes are evolutionarily

161   disadvantaged by Muller's ratchet[35]. Such SVs likely contribute to phenotypic diversity (and potentially

162   heterosis, as suspected for maize[36,37]), and influence *Secale* evolution by creating postpollination

163   reproductive barriers that enable allopatric speciation[38].
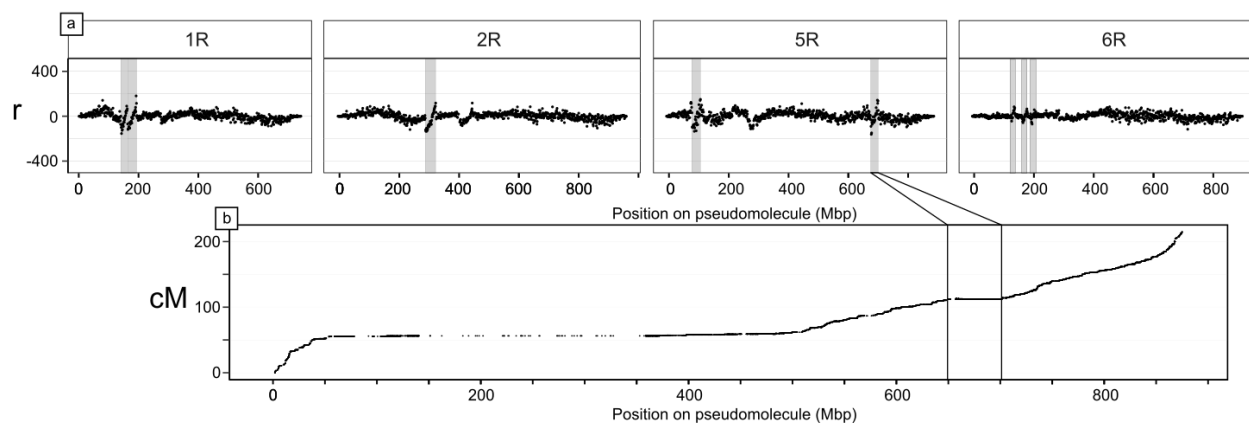
164

165 **Figure M-SV**. Hi-C asymmetry detects SVs between the reference genotype 'Lo7' and *S. cereale* 'Lo225' on four chromosomes. **a)** SVs result in

166 discontinuities in *r*, the ratio of Hi-C links mapping left:right relative to 'Lo7'. Large inversions (marked) typically produce clean, diagonal lines.

167 Visually-identified candidate SVs are shaded, but shading is omitted from some *r* anomalies around centromeres where missing sequence

168 causes artefacts. **b)** The rightmost inversion marked on 5R corresponds to a region of reduced recombination on chromosome 5R.

169 **Revised hypotheses on ancient translocations and the origin of the rye genome**

170 It has been proposed that the cereal rye genome is a mosaic of Triticeae genomes resulting from

171 reticulate evolution because variations in the degree of gene sequence divergence between various

172 regions of the genome and their Triticeae orthologs indicate a number of distinct translocation donors[10].

173 We have presented evidence that the LTR-RT profile (figs. M-TRACKSd—e, Note S-REP) is a result of such

174 reticulation within the rye lineage. It remains to be established whether significant chromatin

175 introgressions occurred involving genera besides *Secale*. We exploited the new assembly to more closely

176 investigate the cause of differential sequence divergence rates by estimating the divergence rate of

177 synonymous coding sequence sites between rye and the wheat D genome (Methods, Note S-REP). The D

178 genome was selected because it (i) contains no large chromosomal translocations relative to ancestral

179 Triticeae karyotype (Note S-COLIN), and (ii) diverged from the ancestral rye genome only after the split

180 from barley, meaning R-D divergence places a coarse lower bound on how much divergence it is possible

181 to accumulate since the R-H split. The rates we recorded (~0.06—0.14 subs/synonymous site/year) can

182 account for the ~5—15% identity spread of divergences that Martis et al. (2013) measured between rye

183 and barley, without recourse to introgressions from beyond the R-D split. No cleanly-delimited

184 divergence-level blocks are immediately evident to support extra-*Secale* introgressions. While some of

185 the variation in divergence levels might yet be caused by such ancient translocations, inferring to what

186 degree is confounded by other sources of random variation, probably including segregating

187 recombination-suppressing SVs as observed in this study. We conclude that the mosaic hypothesis is

188 indeed necessary to explain rye evolution, and currently most parsimonious when limited to

189 introgressive hybridisations primarily between divergent *Secale* populations.

190

191 **Enhancing the benefits of rye germplasm in wheat breeding lines**
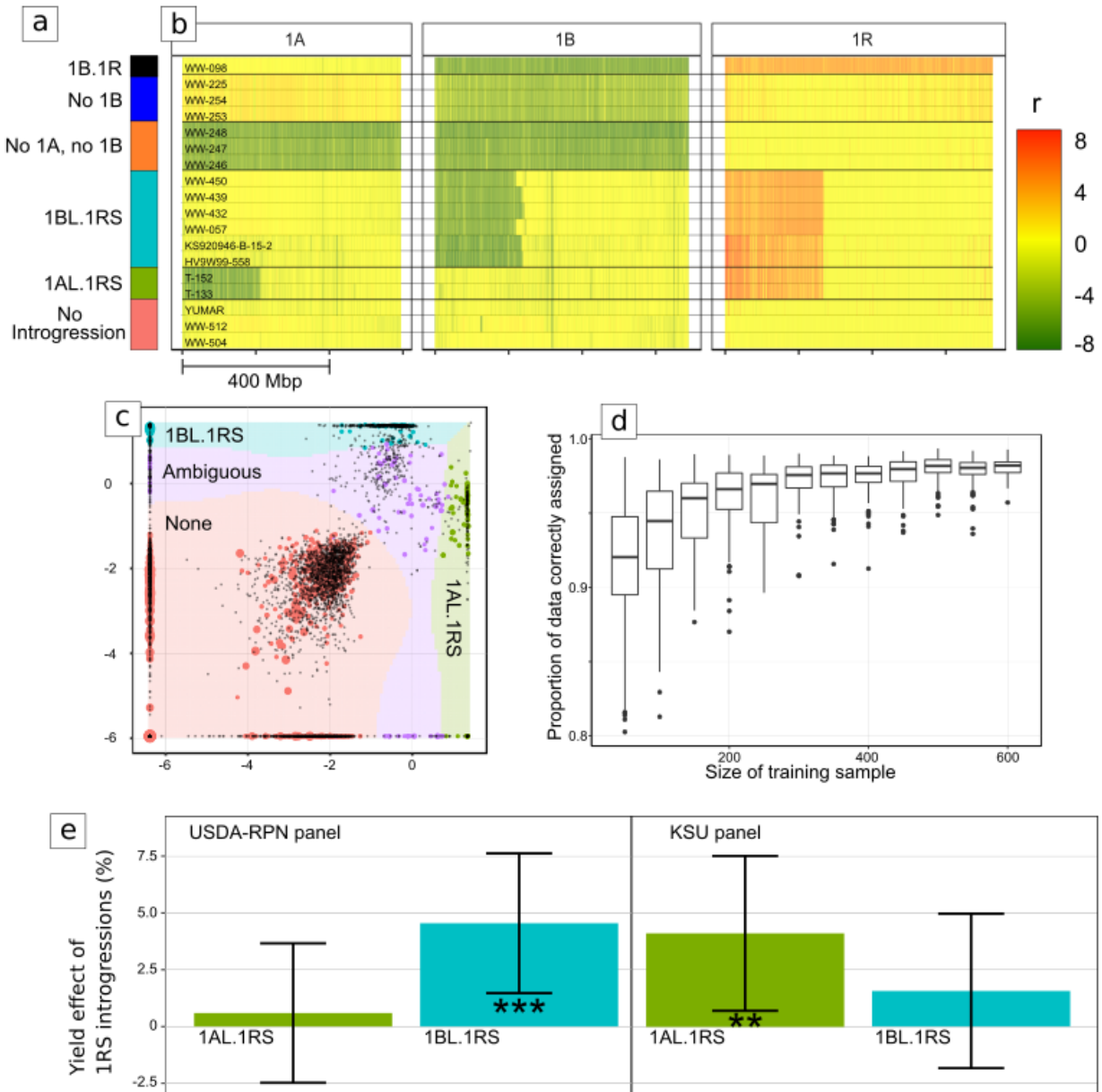
192

193 **Figure M-INTROG.** Combined reference mapping as a means to classify wheat and wheat-rye introgression karyotypes. **a)** Colour key for

194 subfigures b, c, e. **b)** Normalised read mapping depths for 1 Mbp bins of chromosomes 1A, 1B, and 1R, for a selection of wheat lines (including

195 also some *Aegillops tauschii* accessions which contain no A or B subgenome) with various chromosome complements and introgressions (rows).

196 The value *r* denotes the difference between the $\log_2$ reads per million mapping to a bin, compared to *T. aestivum* cv. Chinese Spring. **c)** Visual

197 representation of an SVM classifier, with the two selection features shown on the x and y axes. Points represent training samples, with colour

198 corresponding to human-designated classification, and size proportional to the total number of mapped reads for the sample. Background

199 colours represent the hypothetical classification that would be given to a sample at that position. **d)** Results of cross-validation testing the

200 accuracy of the classifier and its relationship to the size of the training set. **e)** Comparison of yields between non-ambiguous predicted

201 karyotypes, modelled using an MLM with testing year and location as random effects and rye introgressions as fixed effects. Results are shown

202 for panels maintained by two institutions, USDA-RPN (left) and KSU (right). Bar height = Predicted yield effect of introgressions, +/- 1SD.

203 Significant differences (Student's t) given as: '**' p<.01; '***' p<.001.

204

205 The transfer of rye chromatin into bread wheat can provide substantive yield benefits and tolerance to

206 biotic and abiotic stressors[39], though at the expense of bread making quality[40]. These transfers are

207 thought to have involved a single 1BL.1RS Robertsonian translocation originating from cv. Kavkaz and a

208 single 1AL.1RS translocation from cv. Amigo (fig. M-INTROGa)[3,4]. Breeding efforts face a trade-off

209 between yield and quality. Breeders must screen breeding panels for rye introgressions, an effort

210 hitherto dependent upon arduous cytogenetics or marker genotyping, which has limited resolution and

211 is sensitive to genetic variation among lines. With a full reference genome, inexpensive low-density high

212 throughput sequencing (HTS) of a wheat panel proved sufficient to identify the positions of rye

213 introgressions[41]. We implemented an HTS approach on four expansive wheat germplasm panels (KSU,

214 USDA-RPN, CIMMYT, WHEALBI; Methods) segregating for both 1RS.1AL and 1RS.1BL. Translocations into

215 wheat can be observed as obvious changes in normalised read depth across both the translocated and

216 replaced chromosomal regions (fig. M-INTROGb; Note S-INTROG). A range of translocation junctions and

217 karyotypes can be distinguished.

218 The power of this sequence-based approach over previous markers was validated by confirming the

219 karyotype of the novel 1AL.1RS—1BL.1RS recombination line KS090616K-1 (KSU panel; Note S-INTROG)

220 that produces high yields, without sacrificing bread making quality. We confirmed that the KS090616K-1

221 breeding line carried a 1R translocation on group 1A, and after re-sequencing the wheat parents that

222 carry donors of 1A.1R and 1B.1R, used high-density polymorphisms in the translocated 1R arm to

223 precisely identify the recombination breakpoints, which fall at around 6 Mb from the tip of 1RS (Kavkaz-

224 derived) onto the 1AL.1RS (Amigo-derived) line (Note S-INTROG). Moreover, this analysis conclusively

225   confirmed the universal common origins of the Kavkaz- and Amigo-derived translocations respectively

226   (Methods; Note S-INTROG).

227   Visual classification of a whole panel of karyotypes is still time-costly, so we developed an automated

228   support vector machine classifier to alleviate this bottleneck (Methods; figs. M-INTROGc). Automatic

229   classification consistently replicated human assignment with over 97% accuracy (fig. M-INTROGd). We

230   then proved that the automated classifications predict yield. A mixed-effects linear model applied to

231   yield data available for the USDA-RPN and KSU panels showed that 1R introgressions could produce ~3—

232   5% better yields on average (Methods; fig. M-INTROGe; tbls. S-INTROGPHENO—S-GLMRES). The 1A.1R

233   karyotype outyielded 1B.1R in the KSU panel, but the reverse was true of the USDA panel. This likely

234   owes to the diversity of wheat genotypes and environmental conditions used in the trials; the effects of

235   foreign chromatin are highly non-uniform and influenced by diverse factors, in particular the wheat

236   genetic background[40,42]. Only one multi-site study has, to our knowledge, studied yield in 1RS-

237   introgressed wheats on a large scale (Note S-1RS_PUBLIC), in which the best overall yield was achieved

238   by a 1RS.1AL introgression line, both with and without the application of fungicidal treatments and

239   during a drought year, while a 1RS.1BL line in the same panel performed less well, similarly suggesting

240   significant variability in the pathogen resistance and root morphology traits that 1RS can confer to

241   improve yield. Improved knowledge of the individual rye genes that confer these benefits is required to

242   help untangle these factors.

243 **Rye genes for enhanced breeding and productivity**
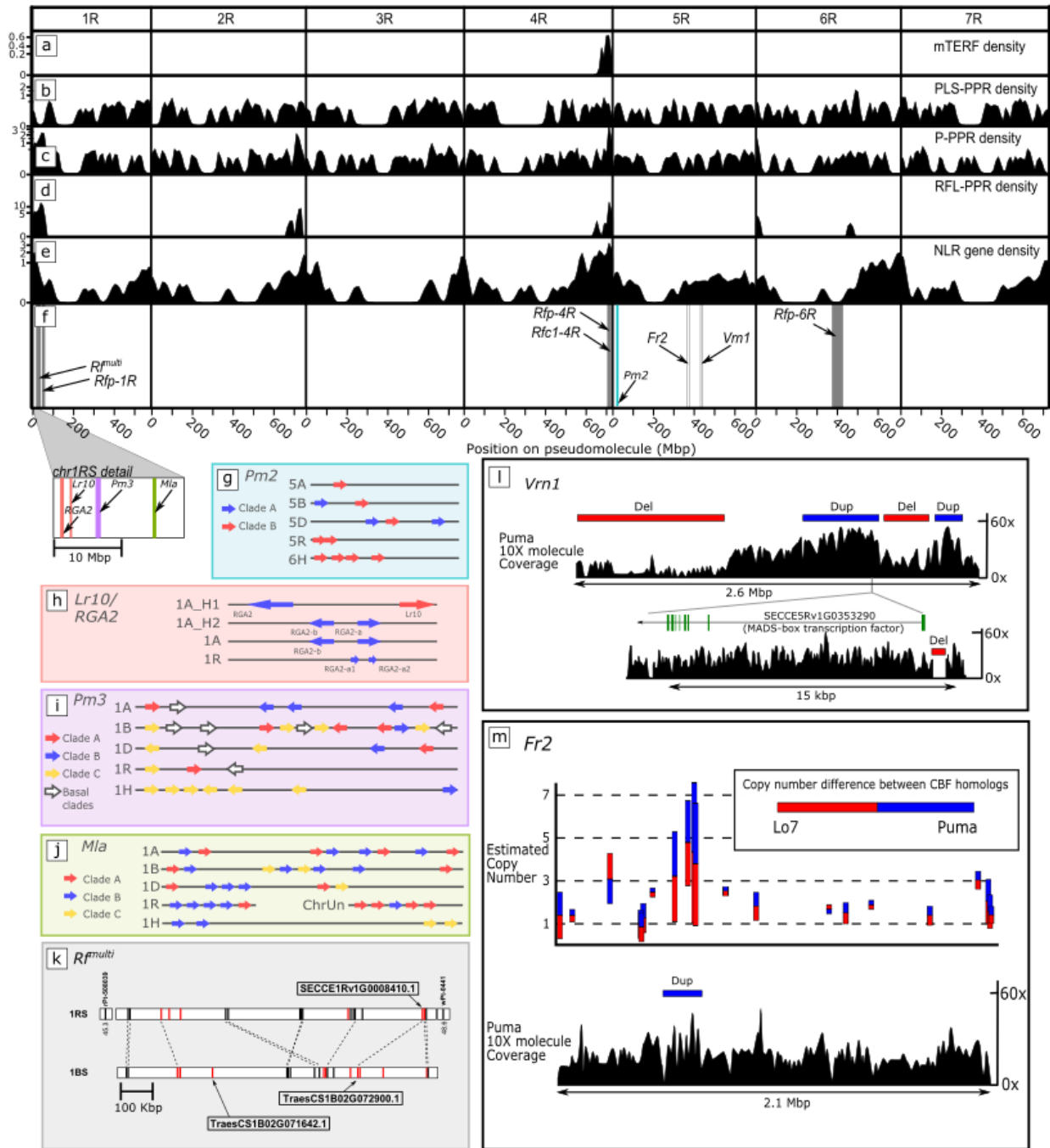
244

245 **Figure M-GENES.** Comparative genomics of rye genes with agricultural significance. **a—e)** Density (instances per Mbp) of mTERFs, PPRs, and

246 NLRs across the pseudomolecules (see also tables S-NLR to S-MTERF). For visualisation, the y-axis is transformed using $x \to x^{1/3}$. **f)** Genes and loci

247 discussed in the text (see also table S-QTLs). Colours correspond to box outlines in panels g—m; **g—j)** Physical organisation of selected NLR

248 gene clusters compared across cultivated Triticeae genomes. **k)** Organisation of RFL genes at the 'Lo7' Rf^{multi} locus compared to its wheat

249 (Chinese Spring) counterpart. Flanking markers are shown on either end of the rye sequence. Two full-length wheat RFLs and a putative rye

250 ortholog are labelled. PPR genes are coloured red. **l—m)** CNV between 'PUMA-SK' and 'Lo7' revealed by 10X Genomics linked read sequencing.

251 (Dup)lications and (Del)etions flagged by the Loupe analysis software are marked. The estimated copy number differences between 'Lo7' and

252 'Puma' are shown for *Cbf* genes within the *Fr2* interval.

253

Enhanced fertilisation control: Rye as a model for hybrid breeding systems in Triticeae

255 Efficient hybrid plant breeding requires lines exhibiting either self-incompatibility (SI), switchable fertility

256 control mechanisms, or gynoecy. Unlike wheat and barley, rye naturally enables both pollen guidance

257 via SI, and switchable fertility via CMS and restorer-of-fertility (*Rf*) genes.

258 Rye's SI is controlled by a two-locus system typical in Poaceae species. Pollen tube germination is

259 suppressed when both stigma and pollen possess identical alleles at two SI loci, termed the *S*- and the *Z*-

260 locus[6], previously mapped to chromosomes 1R and 2R[43-45] respectively. The breakdown of SI is poorly

261 understood, yet essential for the development of inbred lines, which is in turn indispensable for

262 producing heterotic seed and pollen parent lines in hybrid breeding. A DOMAIN OF UNKNOWN

263 FUNCTION gene, designated *DUF247*, is a prime candidate for the *S*-locus in the related ryegrass (*Lolium*

264 *perenne*, Poaceae, Tribe Poeae)[46]. We mapped the rye *S*-locus-controlled SI phenotype to an interval on

265 1R, which falls about 3 Mbp from the rye ortholog of *L. perenne*'s *DUF247* (*SECCE1Rv1G0014240*;

266 Methods; tbls. S-QTLS—S-1RSTS). Similarly, the *Z*-locus-linked marker *TC116908*[45] mapped within about

267 0.2 Mbp of two other *DUF247* homologs (*SECCE2Rv1G0130770*; *SECCE2Rv1G0130780*) on 2R. This

268 proximity suggests that *DUF247* might have been involved in SI since at least the time of the Triticeae—

269 Poaceae split, making it a candidate for investigation relevant to barley and wheat[47,48].

270 Turning to fertility control, mitochondrial genes that selfishly evolve to cause CMS prompt the evolution

271 of nuclear *Rf* genes to suppress their expression or effects. Known *Rf* genes belong to a distinct clade

272 within the family of pentatricopeptide repeat (PPR) RNA-binding factors, whose encoded proteins are

273 referred to as *Rf*-like (RFL)[49,50]. Members of the mitochondrial transcription TERmination Factor (mTERF)

274 family are likely also involved in fertility restoration in cereals[51]. The repertoire of restorer genes is

275 predicted to expand in outcrossing species[35,52]. We investigated this hypothesis by comparing RFL and

276   mTERF gene counts between rye and several closely and distantly allied species including barley and the

277   subgenomes of various wheat species. The numbers of rye RFLs (n=82) and mTERFs (n=131) place it

278   clearly within the range occupied almost exclusively by outcrossers (i.e. $n_{mTERF} > 120$ and $n_{RFL} > 65$; tbls. S-

279   PPR_BREEDINGSYS; Note S-OUTIN), an indicator that rye's younger RFL/mTERF genes evolved under

280   selection to suppress CMS. The 'Lo7' sequence assembly reveals strong overlap in the distribution of

281   PPR-RFLs and mTERF gene clusters, and strong correlation of these clusters with known *Rf* loci

282   (Methods; fig. M-GENESa—f; tbls. S-QTL, S-PPR, S-MTERF). A PPR-RFL/mTERF hotspot on 4RL coincides

283   with known *Rf* loci for two rye CMS systems known as CMS-P (the commercially predominant 'Pampa'-

284   type) and CMS-C[7,14,53,54] (fig. M_GENESb,e,f; tbls. S-PPR, S-QTL). We determined, as previously

285   hypothesised, that these two loci, *Rfp* and *Rfc*, are indeed closely linked but physically distinct[55] (tbl. S-

286   QTLS). Two members of the PPR-RFL clade reside within 0.186 Mbp of the *Rfc1* locus (tbls. S-PPR, S-QTL).

287   The *Rfp* locus is, in contrast, neighboured by four mTERF genes (tbls. S-MTERF, S-QTL), in agreement

288   with previous reports that an mTERF protein represents the *Rfp1* candidate gene in rye[56].

289   While the most commonly used restorer cytoplasm in wheat hybrid breeding is derived from *Triticum*

290   *timopheevii* Zhuk. (CMS-T)[57], alternative sterility-conferring cytoplasms acquired from *Aegilops kotschyi*

291   Bois., *Ae. uniaristata* Vis. and *Ae. mutica* Bois.[58] can be efficiently restored by the wheat locus *Rf*^{multi}

292   (*Restoration-of-fertility in multiple CMS systems*) on chromosome 1BS. Replacement of the *Rf*^{multi} locus

293   by its rye ortholog produces the male-sterile phenotype[59,60]. Characterising this pair of sterility-switching

294   genes could expedite flexible future solutions for the development of exchangeable wheat restorer

295   lines. At the syntenic position of *Rf*^{multi}, the wheat B subgenome and rye share a PPR-RFL gene cluster—

296   with almost twice the number of genes in wheat[9] (fig. M-GENESm; tbl. S-QTLs; Notes S-OUTIN—S-

297   RFMULTI). Only two wheat RFL-PPR genes in the cluster, *TraesCS1B02G071642.1* and

298   *TraesCS1B02G072900.1*, encode full length proteins with only the latter corresponding to a putative rye

299   ortholog (*SECCE1Rv1G0008410.1*). Thus, the absence of a *TraesCS1B02G071642.1* ortholog in the non-

300   restorer rye suggests it as an attractive Rf^{multi} candidate. The only current implementations of a wheat-

301   rye Rf^{multi} CMS system involve 1RS.1BL translocations[5,58,61], which are typically linked to reduced baking

302   quality[40]. Breaking this linkage may now benefit from marker development and/or genome editing

303   approaches targeting *TraesCS1B02G071642.1*.

304

New allelic variety in NLR genes and opportunities for pathogen resistance

306   Nucleotide-binding-site and leucine-rich repeat (NLR)-motif containing genes commonly associate with

307   pest and pathogen resistance[62]. We annotated 792 full-length rye NLR genes (tbls. S-NLR—S-RLOC),

308   finding them enriched in distal chromosomal regions, similar to what has been seen recently in the

309   bread wheat genome[9,63] (fig. M-GENESa; Note S-NLR). Distal parts of chromosomes 4RL and 6RL, which

310   bear a distinct TE composition, are also particularly rich in NLR genes, further corroborating a unique,

311   evolutionary-distinct origin for these segments.

312   We compared the genomic regions in rye that are orthologous to resistance gene loci *Pm2*, *Pm3*, *Mla*,

313   *Lr10* from wheat and barley (tbl. S-RLOC; fig M_GENESg-j; Note S-NLR). Besides the *Lr10* locus, all loci

314   contained complex gene families with several subfamilies that were present or absent in some genomes,

315   indicating either functional redundancy, or the evolution of distinct resistance pathways or targets. For

316   example, the wheat *Pm3* and rye *Pm8/Pm17* genes are orthologs and belong to a subfamily (clade A, fig.

317   M_GENESi) which is absent in barley, whereas a different distinct subfamily (clade B, fig. M-GENESi) of

318   the *Pm3* genes is present in wheat and barley but absent in rye (fig. M-GENESi, Note S-NLR). A similar

319   case occurs in the *Mla* family: One of two main identified clades (clade B, fig. M_GENESj) contains

320   known wheat resistance genes *TmMla1*[64], *Sr33* and *Sr50*[65] and yet is absent in barley, while a second

321   *Mla* subfamily (clade C, fig. M_GENESj) contains all known barley *Mla* resistance alleles[66], yet the clade is

322   absent from rye (Note S-NLR). Rye inbred line 'Lo7', therefore appears to have lost whole subclades of

323   pathogen resistance genes since its split from wheat.

324   The genetic basis of cold tolerance in rye, and its applications to wheat

325
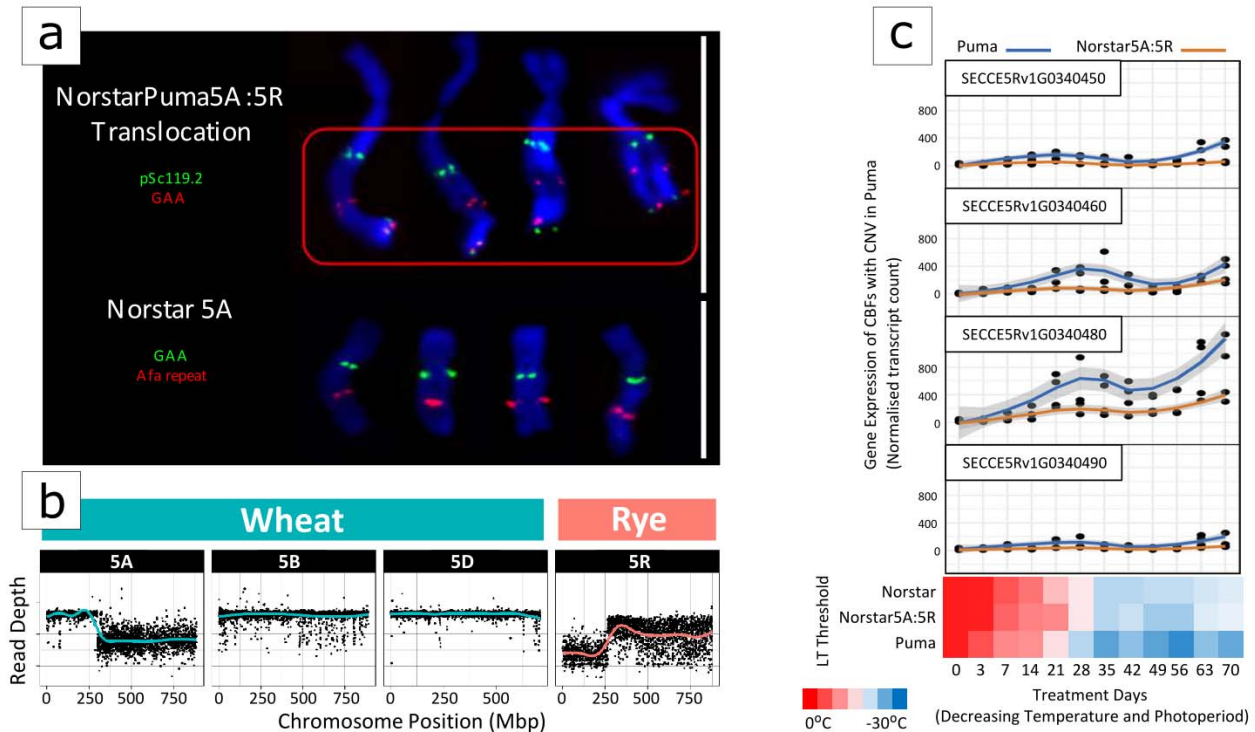
Figure M-COLD. Cold tolerance region *Fr2* in 'Puma' and 'NorstarPuma5A:5R' translocation line. **a)** Chromosome labelling (top) using wheat and rye specific probes for chromosome 5A in 'Norstar' and 5A:5R in the 'Puma'/'Norstar' translocation line confirms the presence of a rye translocation (red box). Read depth (bottom) of group 5 chromosomes confirms the balanced translocation event, gain of a large region of chromosome 5R from 'Puma' (rye - light read line) and loss of a large region on chromosome 5A of 'Norstar' (wheat - light blue line) in 'NorstarPuma5A:5R'. White bars = 10 μm. **b)** Confirmation of the 5A.5R translocation into 'Norstar' using the combined reference mapping method. Read depth is given in $\log_2$ reads per million vs Chinese Spring. **c)** Gene expression analysis of rye *Cbf* genes with copy number variation in 'Puma' (blue line) and 'NorstarPuma5A:5R' (orange line). Plants were grown in a time series with decreasing day length and temperature over a 70 day period and the temperatures at which fifty percent lethality was observed (LT_{50}) were recorded (heatmap).

As the most frost tolerant crop among the Triticeae[67], rye is an ideal model to investigate the genetic architecture of low temperature tolerance (LTT) in cereals. Genetic mapping has revealed a locus *Fr2* on the group 5 chromosomes controlling LTT[68] in rye[69], *T. monococcum*[70], bread wheat[71,72], and barley[73]. In cold-tolerant varieties, the *Fr2* locus up-regulates LTT-implicated *Cbf* genes during seedling development under cold conditions[74]. *Cbf* genes are highly conserved in the Triticeae[75]. We identified the *Fr2* locus as

340 a cluster of 21 *Cbf*-related genes at 614.3—616.5 Mbp on 5R (tbl. S-FR2). The region also contained 12

341 other genes that have been implicated in plant development, such as MYB transcription factors and a

342 FAR1-related gene (tbl. S-FR2). A comparison of annotated Triticeae protein sequences within *Fr2*

343 suggest the *Cbf* gene family expanded in rye, a mechanism for rye's LTT, consistent with findings from

344 other Triticeae[76] (Note S-COLD).

345 To identify variation that may be important for cold acclimation we used recurrent selection to develop

346 an *Fr2* homozygous line of the self-incompatible rye variety 'Puma', which exhibits exceptional LTT. We

347 sequenced 10X Genomics Chromium libraries of this line (designated 'Puma-SK') and performed a

348 comparison to the 'Lo7' reference sequence as a control since 'Lo7' has comparatively poor LTT.

349 Mapping depth analysis detected copy number variation (CNV) patterns in four *Fr2 Cbf* genes

350 (*SECCE5Rv1G030450*, *SECCE5Rv1G030460*, *SECCE5Rv1G030480*, and *SECCE5Rv1G030490*; fig. M-COLDm;

351 tbl. S-CNV; Note S-COLD). Encouragingly, all four are members of the *Cbf* subfamily ('group IV', see fig. S-

352 CBFPHYLO) for which CNV has been previously implicated in LTT in wheat[76]. Interestingly, we also

353 detected a 597 bp deletion in the promoter of 'Puma''s *Vrn1* (SECCE5Rv1G0353290) allele. Although the

354 effect of this deletion on LTT is not yet established, *Vrn1* is known to progressively down-regulate the

355 expression of LTT genes during the vegetative/reproductive transition, impairing the plant's ability to

356 acclimate to cold stress[77,78].

357

358 We also assessed LTT-implicated genes' potential for transfer to other members of the Triticeae, mainly

359 wheat. 'Norstar' winter wheat is an important Canadian line with LTT sufficient to allow experiments in

360 the Canadian winter—but weaker than 'Puma''s LTT, making it suitable for a comparison of LTT between

361 wheat and rye[77]. A locus influencing 'Norstar''s superior LTT occurs on chromosome 5A[71] and, like 'Lo7',

362 contains tandemly repeated *Cbfs*[79]. We thus developed a 5A.5RL translocation line in the 'Norstar'

363 winter wheat genetic background using 'Puma' as the 5R donor, which we confirmed using cytogenetics

364   and combined reference mapping (Methods; fig. M-COLDa,b). As a result of the translocation, the wheat

365   *Cbf* and *Vrn1* cluster is replaced completely by the orthologous rye locus (fig. M-COLDb; tbl. S-CNV).

366   However, the LTT of 'Norstar' was not significantly altered by the translocation (fig. M-COLDc),

367   suggesting that the rye *Cbf* gene cluster is activated in wheat, but it is differentially regulated in the

368   wheat background, as previously suggested by Campoli et al. (2009)[74]. We used RNAseq to confirm that

369   expression of 'Puma' *Vrn1* and those *Cbf*s with CNV were indeed attenuated during treatments of cold

370   stress in 'Norstar5A:5R' (fig. M-COLDc; Note S-COLD). Characterisation of these important regulatory

371   factors is an ongoing effort, necessary to facilitate improvement of wheat temperature tolerance using

372   rye cytoplasm introgressions.

373   *Discussion*

374   The high-quality chromosome-scale assembly of rye inbred line 'Lo7' constitutes an important step

375   forward in genome analysis of the Triticeae crop species, and complements the resources recently made

376   available for different wheat species[16,26,80-82] and barley[15,83]. This resource will help reveal the genomic

377   basis of differences in major life-history traits between the self-incompatible, cross-pollinating rye and

378   its selfing and inbreeding relatives barley and wheat. Our comparative genomic exploration

379   demonstrates how LTR-RT movement histories influence genome expansion and record ancient

380   translocations. The precise nature and origin of the LCMs remains an opportunity for future research,

381   requiring harmonisation of knowledge about the mechanics of pericentromeric structural variation, and

382   the evolutionary effects of gene order disruption. The joint utilisation of the rye and wheat genomes to

383   characterise the effects of rye chromatin introgressions may provide a short-term opportunity to

384   breeders as they continue to better separate confounding variables from the genetic combinations that

385   best improve yield in various environments; but these benefits will ultimately be limited by negative

386   linkage so long as whole chromosome arm translocations are involved. Discoveries at the single-gene

387   level—such as the contributions offered here to pathogen resistance, LTT, the root system (tbl. S-QTLs),

388   SI, and male fertility restoration control—will be best tested and exploited by finer-scale manipulation in

389   dedicated experiments[14]. This is an indispensable pre-requisite for the development of gene-based

390   strategies that exploit untapped genetic diversity in breeding materials and *ex situ* gene banks to

391   improve small grain cereals and meet the changing demands of global environments, farmers and

392   society.

393

394

Methods

'Lo7' genome assembly

Descriptions of the assembly methods are given in notes S-PSASS—S-ASSDATA, and figures S-ASSOVER—S-HICSV.

Gene annotation

We performed de novo gene annotation of the rye genome relying on a previously established automated gene prediction pipeline[15,82]. The annotation pipeline involved merging three independent annotation approaches, the first based on expression data, the second an *ab initio* prediction for structural gene annotation in plants and the third on protein homology. To aid the structural annotation, RNAseq data was derived from five different tissues/developmental stages, and IsoSeq data from three (Supplementary Note 3).

IsoSeq nucleotide sequences were aligned to the rye pseudomolecules using GMAP[84] (default parameters), whereas RNASeq datasets were first mapped using Hisat2[85] (arguments --dta) and subsequently assembled into transcript sequences by Stringtie[86] (arguments -m 150 -t -f 0.3). All transcripts from IsoSeq and RNASeq were combined using Cuffcompare[87] and subsequently merged with Stringtie (arguments --merge -m 150) to remove fragments and redundant structures. Transdecoder github.com/TransDecoder) was then used to find potential open reading frames (ORFs) and to predict protein sequences. BLASTp[88] (ncbi-blast-2.3.0+, arguments -max_target_seqs 1 -evalue 1e-05) was used to compare potential protein sequences with a trusted set of reference proteins (Uniprot Magnoliophyta, reviewed/Swiss-Prot) and hmmscan[89] was employed to identify conserved protein family domains for all potential proteins. BLAST and hmmscan results were fed back into Transdecoder-predict to select the best translations per transcript sequence.

417   Homology-based annotation is based on available Triticeae protein sequences, obtained from UniProt

418   (uniprot.org). Protein sequences were mapped to the nucleotide sequence of the pseudomolecules

419   using the splice-aware alignment software GenomeThreader (http://genomethreader.org/; arguments -

420   startcodon -finalstopcodon -species rice -gcmincoverage 70 -prseedlength 7 -prhdist 4). Evidence-based

421   and protein homology based predictions were merged and collapsed into a non-redundant consensus

422   gene set. *Ab initio* annotation using Augustus[90] was carried out to further improve structural gene

423   annotation. To minimise over-prediction, hint files using IsoSeq, RNASeq, protein evidence, and TE

424   predictions were generated. The wheat model was used for prediction.

425   Additionally, an independent, homology-based gene annotation was performed using GeMoMa [91] using

426   eleven plant species: *Arabidopsis thaliana* (n=167), *Brachypodium distachyon* (314), *Glycine max* (275),

427   *Mimulus guttatus* (256_v2.0), *Oryza sativa* (323), *Prunus persica* (298), *Populus trichocarpa* (444),

428   *Sorghum bicolor* (454), *Setaria italica* (312), *Solanum lycopersicum* (390), and *Theobroma cacao* (233).

429   All versions were downloaded from Phytozome (phytozome.jgi.doe.gov/pz). Initial homology search for

430   coding exons was done with mmseqs2[92]. These results were then combined into gene models with

431   GeMoMa using mapped RNAseq data for splice site identification. The resulting eleven gene annotation

432   sets were further combined and filtered using the GeMoMa module GAF. The following filters were

433   applied: a) complete predictions (i.e. predictions starting with Methionine and ending with a stop

434   codon); b) relative GeMoMa score >=0.75; c) evidence>1, (i.e. predictions were perfectly supported by

435   at least two reference organisms), or tpc=1 (i.e., predictions were completely covered by RNA-seq

436   reads), or pAA>=0.7 (i.e., predictions with at least 70% positive scoring amino acid in the alignment with

437   the reference protein).

438   All structural gene annotations were joined with EvidenceModeller[93], and weights were assigned as

439   follows: Expression-based Consensus gene set (RNAseq, and IsoSeq and protein homology-based): 5;

440   homology-based (GeMoMa), 5; ab initio (augustus), 2.

441   In order to differentiate candidates into complete and valid genes, non-coding transcripts, pseudogenes

442   and transposable elements, we applied a confidence classification protocol. Candidate protein

443   sequences were compared against the following three manually curated databases using BLAST: firstly

444   PTREP (botserv2.uzh.ch/kelldata/trep-db), a database of hypothetical proteins that contains deduced

445   amino acid sequences in which, in many cases, frameshifts have been removed, which is useful for the

446   identification of divergent TEs having no significant similarity at the DNA level; secondly UniPoa, a

447   database comprised of annotated Poaceae proteins; thirdly UniMag, a database of validated

448   magnoliophyta proteins. UniPoa and UniMag protein sequences were downloaded from Uniprot

449   (www.uniprot.org/) and further filtered for complete sequences with start and stop codons. Best hits

450   were selected for each predicted protein to each of the three databases. Only hits with an E-value below

451   10e-10 were considered.

452   Furthermore, only hits with subject coverage (for protein references) or query coverage (transposon

453   database) above 75% were considered significant and protein sequences were further classified using

454   the following confidence: a high confidence (HC) protein sequence is has at least one full open reading

455   frame and has a subject and query coverage above the threshold in the UniMag database (HC1) or no

456   BLAST hit in UniMag but in UniPoa and not TREP (HC2); a low confidence (LC) protein sequence is not

457   complete and has a hit in the UniMag or UniPoa database but not in TREP (LC1), or no hit in UniMag and

458   UniPoa and TREP but the protein sequence is complete.

459   The tag REP was assigned for protein sequences not in UniMag and complete but with hits in TREP.

460   Functional annotation of predicted protein sequences was done using the AHRD pipeline

461   (github.com/groupschoof/AHRD). Completeness of the predicted gene space was measured with BUSCO

462   (v3; https://busco.ezlab.org/).

463 **RNA isolation and sequencing**

464 RNA-seq for annotation

465

466 Seeds of 'Lo7' were sown in a Petri dish on moistened filter paper and treated with cold stratification (4

467 °C) for two days during imbibition. After an additional day at room temperature (~20 °C) seedlings were

468 transferred to a 40-well tray containing a peat and sand compost and propagated in a Conviron BDW80

469 cold environment room (CER; Conviron) with set points of 16 h day/8 h night and temperatures of 20/16

470 °C for a further three days. Tissues were sampled at six stages, described in table S-RNAGROWTH. Plants

471 for sampling timepoints 1—3 were transferred to a CER set at 16-hour photoperiod (300 µmol m−2 s−1),

472 temperatures of 20 and 16 °C, respectively, and 60% relative humidity. Plants for sampling timepoints

473 4—6 were transferred to a vernalisation CER running at 6 °C with 8 hours photoperiod for 61 days. After

474 this period the plants were transferred to 1 L pots containing Petersfield Cereal Mix (Petersfield,

475 Leicester, UK) and moved to the CER with settings as described above. Total RNA was extracted from

476 each of the six organ/stages using RNeasy plant mini-kits (Qiagen). For the RNAseq data sets used for

477 the annotation. RNA from 3 biological replicates for each organ/stage was pooled and for the 6 pooled

478 samples, library construction and sequencing on the Illumina NovaSeq platform was performed by

479 Novogene using a standard strand specific protocol (en.novogene.com/next-generation-sequencing-

480 services/gene-regulation/mrna-sequencing-service) and generating >60 M 150 PE reads per sample.

481

482 For the IsoSeq data used in the annotation RNA from root and shoot samples were used (timepoints 1

483 and 2 in table S-RNAGROWTH). The IsoSeq libraries were created starting from 1µg of total RNA per

484 sample and full-length cDNA was then generated using the SMARTer PCR cDNA synthesis kit (Clontech)

485 following PacBio recommendations set out in the IsoSeq method (pacb.com/wp-

486   content/uploads/Procedure-Checklist-Iso-Seq-Template-Preparation-for-Sequel-Systems.pdf). PCR

487   optimisation was carried out on the full-length cDNA using the KAPA HiFi PCR kit (Kapa Biosystems) and

488   10—12 cycles was sufficient to generate the material required for SMRTbell library preparation. The

489   libraries were then completed following PacBio recommendations, without gel-based size-selection

490   (pacb.com/wp-content/uploads/Procedure-Checklist-Iso-Seq-Template-Preparation-for-Sequel-

491   Systems.pdf).

492   The library was quality checked using a Qubit Fluorometer 3.0 (Invitrogen) and sized using the

493   Bioanalyzer HS DNA chip (Agilent Technologies). The loading calculations for sequencing were

494   completed using the PacBio SMRTlink Binding Calculator v5.1.0.26367. The sequencing primer from the

495   SMRTbell Template Prep Kit 1.0-SPv3 was annealed to the adapter sequence of the libraries. Each library

496   was bound to the sequencing polymerase with the Sequel Binding Kit v2.0. Calculations for primer and

497   polymerase binding ratios were kept at default values. Sequencing Control v2.0 was spiked into each

498   library at ~1% prior to sequencing. The libraries were prepared for sequencing using Magbead loading

499   onto the Sequel Sequencing Plate v2.1. The libraries were sequenced on the PacBio Sequel Instrument

500   v1, using 1 SMRTcell v2 per library. All libraries had 600-minute movies, 120 minutes of immobilisation

501   time, and 120 minutes pre-extension time (tbl. S-DATACCESS).

502

503   RNA-seq for expression profiling of 'NorstarPuma5A:5R' and 'Puma'

504   Total RNA was extracted from 48 samples, representing both 'NorstarPuma5A:5R' and 'Puma' lines at

505   each sampling date of the 12 time points during cold acclimation (Note S-COLD), using the Plant RNA

506   Isolation Mini Kit (Agilent Technologies). The yield and RNA purity were determined

507   spectrophotometrically with Nanodrop 1100 (Thermfisher), and the quality of the RNA was verified by

508   Agilent 2100 Bioanalyzer (Agilent Technologies). Purified total RNA was precipitated and re-suspended

509   in RNase-free water to a final concentration of 100 ng/μl. Libraries were constructed using the TruSeq

510   RNA Sample Preparation Kit v2 (Illumina) with two replicates at each time point. Paired-end sequencing

511   was conducted on the Illumina HiSeq2500, generating 101 bp reads (tbl. S-DATACCESS).

512   **Annotation of repetitive elements**

513   For use in the evolutionary analyses presented in the main text (e.g. fig. 4d—g) annotated a high-

514   stringency set of full-length transposon copies belonging to single TE families (tbl. S-TEANNOT) using

515   BLASTn[88] searches (using default parameters) against the 'Lo7' pseudomolecules for long terminal

516   repeats (LTRs) documented in the TREP database

517   (botinst.uzh.ch/en/research/genetics/thomasWicker/trep-db.html) that occur at a user-defined distance

518   range in the same orientation: For RLC_Angela elements, the two LTRs had to be found within a range of

519   7,800—9,300 bp (a consensus RLC_Angela sequence has a length of approximately 8,700 bp), while a

520   range from 6,000—12,000 bp was allowed for RLG_Sabrina and RLG_WHAM elements. For the

521   centromere-specific RLG_Cereba elements, a narrower range of 7,600-7,900 bp was used. Multiple

522   different LTR consensus sequences were used for the searches in order to cover the intra-family

523   diversity. A total of 18 LTR consensus sequences each were used for RLC_Angela, seven for RLG_Sabrina

524   elements, 6 were used for RLG_WHAM elements, and 5 for RLC_Cereba elements.

525   To validate the extracted TE populations, the size range of all isolated copies and the number of copies

526   that flanked by target site duplications (TSDs) were determined. A TSD was accepted if it contained at

527   least 3 matches between 5' and 3' TSD (e.g. ATGCG and ACGAG). This low stringency was applied

528   because TSD generation is error-prone[94], and thus multiple mismatches can be expected. Across all

529   surveys, 80-90% of all isolated full-length elements were flanked by a TSD.

530   The pipeline also extracts so-called "solo-LTRs"—products of intra-element recombination that results in

531   loss of the internal domain and generation of a chimeric solo-LTR sequence—as a metric of how short

532   repetitive sequences are assembled.

533   The two LTRs of each TE copies were aligned with the program Water from the EMBOSS package[95] and

534   nucleotide differences between LTRs were used to estimate the insertion age of each copy based on the

535   estimated intergenic mutation rate of 1.3E-8 substitutions per site per million years[96].

536   Full-length DNA transposons were identified by BLASTn searches of consensus sequences of the terminal

537   inverted repeats (TIRs) of a given family. TIRs were required to be found in opposite orientation in a

538   user-defined distance interval of 7,000—15,000 bp.

539

540   To produce a library of full length LTR-retrotransposons suitable for quantitative assembly completeness

541   comparison (fig. S-RPT_ASSCMP), we required an annotation performed identically to those carried out

542   on other assemblies (tbls. S-TE_ASSCMP_ANNOTSTATS, S-DATAACCESS). We therefore implemented the

543   methods described in Monat et al. (2019)[83] on a selection of genome assemblies given in note S-REP.

544

545   Tandem repeats where annotated with TandemRepeatsFinder[97] under default parameters (tbls. S-

546   SAT_ANNOT, S-TANDREPCOMPN). Overlapping annotations where removed with a priority-based

547   approach assigning higher scoring and longer elements first. Elements which overlapped already

548   assigned elements were either discarded (>90% overlap) or shortened (<=90% overlap) if their

549   remaining length exceeded 49 bp.

550   To obtain a collection of nonredundant tandem repeat units suited for FISH probe development, the

551   consensus sequences of the tandem repeat units (output of TandemRepeatsFinder) where clustered

552   with vmatch dbcluster (vmatch.de) at high stringency with >=98% identity and a mutual overlap >=98%

553   (98 98 -v -identity 98 -exdrop 3 -seedlength 20 -d -p). The 300 largest clusters with member sizes from

554   199 to 343 where each subjected to a multiple sequence alignment with MUSCLE[98] under default

555   parameters. A consensus sequence (>=70% majority) derived per cluster from the MUSCLE score file

556   served as template sequence for the FISH probes (tbl. S-FISH).

557   The distribution of TRs across the genome (main fig. M-FISH) was visualised using R base plotting

558   functions. Colours were selected from the package colourspace palettes 'Reds3' and 'Greens3', e.g.

559   using the command sequential_hcl('Reds3',105)[105:5] to achieve 100 grades of a palette, and then

560   selected to represent relative TR densities by scaling the output of the 'density' function run over the

561   tandem repeats (with automatic bandwidth selection) on each chromosome to between 1 and 100 (for

562   each TR family).

563   **Annotation of miRNAs**

564   MicroRNA identification was performed by following a two-step homology-based pipeline. The 'Lo7'

565   pseudomolecules were compared with all known mature plant miRNA sequences retrieved from

566   miRBase[99] (v21; www.mirbase.org). This step was performed using SUmirFind (https://github.com/

567   hikmetbudak/miRNA-annotation/blob/master/ SUmirFind.pl), an in-house script, and the matches with

568   no mismatch or only one base mismatch between a mature miRNA sequence and the pseudomolecule

569   sequence were accepted. A second in-house script, SUmirFold

570   (https://github.com/hikmetbudak/miRNA-annotation/blob/master/ SUmirFold.pl), was used to obtain

571   precursor sequences of the candidate mature miRNAs from the pseudomolecules and assess their

572   secondary structure-forming abilities with UNAFold[100] (tbls. S-miRNA1—S-miRNAX) together with the

573   following criteria: 1) No mismatches are allowed at Dicer cut sites; 2) No multi-branched loops are

574   allowed in the hairpin containing the mature miRNA sequence; 3) Mature miRNA sequence cannot be

575   located at the head portion of the hairpin; 4) No more than 4 and 6 mismatches are allowed in the

576   miRNA and its hairpin complement (miRNA*), respectively[101,102]. The final set of identified miRNAs from

577   the pseudomolecules was obtained by SUmirScreen script (https://github.com/hikmetbudak/miRNA-

578   annotation/blob/master/ SUmirScreen.py). The resulting miRNAs were mapped back to the

579   pseudomolecules and the genomic distribution statistics were recorded with SUmirLocate script

580   (https://github.com/hikmetbudak/miRNA-annotation/blob/master/ SUmirLocate.py).

581   Coding targets of the identified miRNAs were predicted by the web-tool psRNAtarget, using *S.*

582   *cereale* coding sequences retrieved from NCBI[103,104]. Potential target sequences were compared with the

583   viridiplantae proteins by using BLASTx[88] (arguments -evalue 1E-6 –outfmt 5). Functional annotations of

584   the potential targets were performed using Blast2GO software[105]. Finally, repeat contents of the pre-

585   miRNAs were assessed with RepeatMasker (http://www.repeatmasker.org/).

586   **Fluorescence *in situ* hybridisation (FISH)**

587

588   Three days old roots of the rye accession WR 'Lo7' were pre-treated in 0.002 M 8-hydroxyquinoline at

589   7°C for 24 h and fixed in ethanol:acetic acid (3:1 v/v). Chromosome preparation and FISH were

590   performed according to the methods described by Aliyeva-Schnorr et al. (2015)[106]. The hybridization

591   mixture contained 50% deionized formamide, 2× SSC, 20% dextran sulfate, and 5 ng/µl of each probe.

592   Slides were denatured at 75°C for 3 min, and the final stringency of hybridization was 76%. Thirty-four to

593   forty-five nt long 5'-labelled oligo probes designed for the *in silico* identified repeats and the published

594   probes sequence pSc119.2.1[107] were used as probes (tbl. S-FISH). Images were captured using an

595   epifluorescence microscope BX61 (Olympus) equipped with a cooled CCD camera (Orca ER,

596   Hamamatsu). Chromosomes were identified visually based primarily on morphology, heterochromatic

597   DAPI+ bands, and the localisation of pSc119.2.1[107].

598 **Rye Gene level synteny with other Triticeae species**

599 High confidence gene sequences from the 'Lo7' gene annotation were aligned to the annotated

600 transcriptomes of bread wheat[9] (*Triticum aestivum* cv. Chinese Spring) and barley[15] (*Hordeum vulgare*

601 cv. Morex) using BLASTn[88] with default parameters. The lowest E-value alignment for each gene against

602 the transcriptome associated with each subject genome (or subgenome) was selected, with the longest

603 alignment chosen in the case of a tie. Only reciprocal best matches per (sub/)genome were accepted.

604 BLAST hit filtering and subsequent visualisation were performed in the R statistical environment

605 exploiting the packages 'data.table' and 'ggplot'.

606 **Wheat (D subgenome)—rye substitution rate variation across the genome**

607

608 Probable orthologs shared by the wheat D subgenome[9] and rye line 'Lo7' were identified by aligning

609 BLASTp[88] (default parameters) the predicted proteins of either each genome against the other and

610 applying the reciprocal best match criterion. The identified homologs were first aligned at the protein

611 level and, based on the protein alignment, a codon-by-codon DNA alignment was generated. For

612 comparison of substitution rates, only fourfold degenerate third codon positions were used, namely

613 those of the codons for Ala, Gly, Leu, Pro, Arg, Ser, Thr and Val. From the alignments of fourfold

614 degenerate sites, the ratio of synonymous substitutions per synonymous site was calculated for each

615 gene pair, if at least 100 fourfold degenerate sites could be aligned. Substitution rates along

616 chromosomes were calculated as a 100 genes running average. Because even bi-directional closest

617 homologs may still include "deep paralogs" (i.e. genes that were duplicated in the ancestor of which one

618 copy was deleted in one species while the other copy is deleted in the other species), we performed the

619 same analysis using exclusively single-copy genes. Single-copy genes were identified as follows: all

620 individual rye coding DNA sequences (CDSs) were used in BLASTn searches against all other predicted

621   rye CDSs. A gene was considered single-copy if it had no homologs with E-values below 10e-20.

622   Substitution rates were then calculated as the rates of synonymous substitutions per synonymous site in

623   fourfold degenerate codon sites in coding regions of genes.

624   **Phylogenetic analysis**

625   The genotyping-by-sequencing (GBS) data set of 603 samples from Schreiber et al. (2019)[33] was

626   extended by a 347 further GBS samples from the IPK gene bank (mainly wild *Secale* taxa), and the five

627   samples used in the Hi-C SV-detection study ('Lo7', 'Lo225', 'R1003', 'R925', 'R2446'). The resulting

628   sample set (n=955) and passport data are listed in table S-DIVERSITYPSPT. DNA isolated from the five Hi-

629   C samples was sent to Novogene (en.novogene.com/) for Illumina library construction and sequencing in

630   multiplex on the NovaSeq platform (paired end 150 bp reads, approximately 140 Gbp per sample, S2

631   flow cell). Demultiplexing, adapter trimming, read mapping and variant calling correspond to the

632   approach described in Schreiber et al. (2019)[33], using the new reference for read mapping. The data set

633   was filtered for a maximum of 30% missing data and a minor allele frequency of 1% resulting in 72,465

634   SNPs used for the phylogenic analyses. A neighbor joining tree was constructed with the R package 'ape'

635   version 5.3[108], based on genetic distances computed with the R package SNPRelate[109]. PCA was

636   performed with smartPCA from the EIGENSOFT package (github.com/DReichLab/EIG) using least square

637   projection without outlier removal.

638   **Wheat-rye introgression haplotype identification and classification**

639

640   We assayed for the presence of 1R germplasm in wheat genotypes *in silico* by mapping various wheat

641   sequence data to a combined reference genome made up of the pseudomolecules of rye line 'Lo7' (this

642   study) and wheat cv. Chinese Spring[9]. Publicly available data was obtained from the Wheat and barley

643 Legacy for Breeding Improvement (WHEALBI) project resources[110] (n=506), the International Maize and

644 Wheat Improvement Centre (CIMMYT; n=903), and Kansas State University (KSU; n=4277). GBS libraries

645 were constructed and sequenced for samples from the United States Department of Agriculture

646 Regional Performance Nursery (USDA-RPN; n=875; tbl. S-DATAACCESS) as described in Rife et al.

647 (2018)[111]. Based upon the approach described by Keilwagen et al. (2019)[91], reads were demultiplexed

648 with a custom C script (github.com/umngao/splitgbs) and aligned to the combined reference using

649 bwa[112] mem (arguments -M) after trimming adapters with cutadapt[113]. The aligned reads from all panels

650 were filtered for quality using samtools[114] (arguments flags -F3332 -q20). The numbers of reads aligned

651 to 1 Mbp non-overlapping bins on each pseudomolecule were tabulated. The counts were expressed as

652 *rpmm* ⬚ *log$_2$(reads mapped to bin per million reads mapped)*. To control for mappability biases over the

653 genome, the *rpmm* for each bin was normalised by subtracting the *rpmm* attained by the Chinese Spring

654 sample for the same bin to give the normalised *rpmm*, *r*.

655 To investigate the possibility of classifying the samples automatically, visual representations of *r* across

656 the combined reference genome were inspected, and obvious cases of 1R.1A and 1R.1B introgression

657 were distinguished from several other karyotypes including non-introgressed samples, and ambiguous

658 samples showing a slight overabundance of 1RS reads, but less discernible signals of depletion in 1A or

659 1B (see Note S-INTROG). We defined the following feature vectors: *featureA = -log[(mean(r1A$_I$) -*

660 *mean(r1A$_N$)) x (mean(r1R$_I$) - mean(r1R$_N$))]* and *featureB = -log[(mean(r1B$_I$) - mean(r1B$_N$)) x (*

661 *mean(r1R$_I$) - mean(r1R$_N$))]*. Whenever the term inside the log was negative (and would thus give an

662 undefined result), the value of the feature was set to the minimum of the defined values for that

663 feature. The quantity *mean(r1R$_I$)* refers to the average value of r for all bins within the terminal 200 Mbp

664 of the normally (*I*)ntrogressed end of 1R (an $_N$ in the subscript denotes the terminal 300 Mbp of the

665 normally (*N*)on-introgressed arm), and so forth for other chromosomes. This choice of feature definition

666 meant that, wherever little difference in *r* occurred between 1RS and 1RL, suggesting no presence of

667   rye, the factor $mean(r1R_i)$ - $mean(r1R_N)$ would pull the feature values close to the origin, and differences

668   between r on the long and short arms of 1A or 1B would pull the values of A or B respectively away from

669   the origin, depending upon which introgressions are present. A classifier was developed by training a

670   support vector machine to distinguish non-introgressed, 1A.1R-introgressed, 1B.1R-introgressed, and

671   ambiguously-introgressed samples, using the function ksvm (arguments type="C-svc", kernel='rbfdot',

672   C=1) from the R package kernlab. Classification results are given in table S-INTROG_PREDICTED. Testing

673   was performed by generating sets of between 50 and 600 random samples from the dataset and using

674   these to train a model, then using the kernlab::predict to test the model's accuracy of prediction on the

675   remaining data not used in training. This was repeated 100 times for each training data set size.

676

677   To investigate the 1R-recombinant genotype KS090616K, raw reads of genotypes Larry, TAM112 and

678   KS090616K (NCBI SRA project id: PRJNA566411) were mapped to the combined wheat/rye reference,

679   and mapping results processed with samtools[114]. The bcftools[114] mpileup and call functions were used to

680   detect and genotype single-nucleotide polymorphisms (SNPs) between the two samples. SNP positions

681   at which Larry and TAM112 carried different alleles were used to partition chromosome 1RS in

682   KS090616K into parental haplotypes (Note S-INTROG).

683

684   To confirm the common origin of the 1AL.RS and 1BL.1RS introgressions, predicted 1RS carriers were

685   selected to form a combined 1RS panel (over twelve hundred lines) to call SNPs. A total of over 3 million

686   SNPs were called with samtools/bcftools (mpileup -q 20, -r chr1R:1-300000000; call -mv). SNPs were

687   filtered based on combined minimum read depth of 25, minor allele frequency of 0.01. A total of over

688   900 thousand SNPs were obtained. All pair-wise identity by state (IBS) percentages were calculated and

689   the square root values of percent different calls were used to derive a heatmap for all pair-wise

690   comparisons.

691 **Identification and analysis of gene families**

692 Resistance gene homologs

693

694 To investigate rye homologs of the wheat and barley genes *Pm2*, *Pm3*, *Mla*, *Lr10* and *RGA2* (GeneBank

695 IDs in tbl. S-NLRSEARCH), homology searches were performed against the rye 'Lo7', bread wheat[9] (cv.

696 Chinese Spring), and barley[15] (cv. Morex) genome sequences, using BLASTn[88] (default parameters). Hits

697 with at least 80% sequence identity were visualised using dotter[115] for manual assessment and

698 annotation. The obtained coding sequences were converted to protein sequences, allowing comparison

699 with the EMBOSS program WATER (emboss.sourceforge.net), ClustalW[116], or MUSCLE[98], with reference

700 sequences and other obtained sequences to aid distinction between potentially functional full-length

701 genes, and pseudogenes with truncations or premature stop codons.

702 Annotated genes were aligned using MUSCLE (default parameters), and the phylogenetic relationships

703 among them were inferred using MisterBayes[117] (GTR substitution model with gamma distributed rate,

704 variation across sites, and a proportion of invariable sites).

705 Manually-annotated positions of the genes *Pm2*, *Pm3*, *Mla*, *Lr10* and *RGA2* on the 'Lo7'

706 pseudomolecules were compared with the annotated NLR genes identified by the gene feature

707 annotation pipeline (described above) in order to link the genome-wide NLR analysis with the detailed

708 analysis of the four R loci. Pairwise distances between NLRs were calculated based on the resultant tree

709 using the cophenetic.phylo function in the R package 'ape'[108], and multidimensional scaling on the

710 pairwise distances was conducted with the core R function 'cmdscale'.

711

712   PPR and mTERF genes

713

714   The 'Lo7' pseudomolecules were scanned for ORFs with the getorf program of the EMBOSS package[95].

715   ORFs longer than 89 codons were searched for the presence of PPR motifs using hmmsearch from the

716   HMMER[118] package (http://hmmer.org) and the profile hidden Markov models (HMMs) as defined in

717   Cheng et al. (2016)[119] for the PPR family PF02536 from the Pfam 32.0 database (http://pfam.xfam.org)

718   and for the mTERF motif[120]. Downstream processing of the hmmsearch results for the PPR proteins

719   followed the pipeline described in Cheng et al. (2016)[119]. A score was attributed to each PPR sequence

720   (the sum of hmmsearch scores for all PPR motifs in the protein). In parallel, the HC and LC protein

721   models from the gene feature annotation (described above) were screened to identify the annotated

722   proteins containing PPR motifs. Five-hundred and twenty-six PPR models were identified in the HC and

723   seventy-six in the LC protein datasets respectively, and scored using the same approach as with the

724   hmmsearch results. Where putative exons identified from the six-frame translations of the genome

725   sequence overlapped with gene models in the 'Lo7' annotation, only the highest scoring of the

726   overlapping models were retained. P- and PLS-class genes with scores below 100 and 240, respectively,

727   were removed from the annotation, as they are unlikely to represent functional PPR genes. Only genes

728   encoding mTERF proteins longer than 100 amino acids were included in the final annotation.

729

730   Mapping genes governing the reproduction biology in rye

731

732   Molecular markers previously mapped in relation to Rf and SI genes were integrated in the 'Lo7'

733   assembly (S-QTL) based on BLASTn sequence similarity searches as described by Hackauf et al. (2009)[121].

734   The *S* locus genomic region in rye was identified using orthologous gene models from *Brachypodium*

735 *dystachion* including *Bradi2g35750*, that is predicted to encode a protein of unknown function

736 *DUF247*[46]. Furthermore, we included the marker *SCM1* from Hackauf and Wehling (2002)[122] in our

737 analyses, that represents the rye ortholog of a thioredoxin-like protein linked to the *S* locus in the grass

738 *Phalaris coerulescens*[123,124]. Likewise, the isozyme marker *Prx7* linked to the *S* locus in rye was

739 investigated as described by Wricke and Wehling (1985)[43]. The *S* locus was mapped in a F2 population

740 (n= 96), produced by crossing the self-incompatible variety 'Volhova' with the self-fertile line No. 5

741 ('l.5'), the latter of which carrying the mutation for self-fertility at the *S* locus on chromosome 1R

742 (Voylokov et al. 1993, Fuong et al. 1993). Progeny from this cross are heterozygous for the self-fertility

743 mutation. The gametic selection caused by self-incompatibility in such crosses was used for the mapping

744 of *S* relative to markers (S-1RSTS) according to previously described protocols[45,125]. The SI mechanism

745 prevents fertilization of all pollen grains except those carrying the *Sf* allele. As a consequence, only those

746 50% of the pollen grains carrying the mutation will be able to grow and fertilize upon self-pollination of a

747 F1 hybrid from the cross. Therefore, the functional *S* allele results in distorted segregation of marker loci

748 linked to the self-fertility mutation in the F2. The degree of segregation distortion depends on the

749 recombination frequency *r* between the segregation distortion locus (SDL) and analyzed marker loci. For

750 example, after selfing a F1 with the constitution *SM1/SfM2*, where *S* and *Sf* are active (wild type) and

751 inactive (mutant) alleles of the self-incompatibility locus *S*, respectively, and M1 and M2 are alleles of a

752 marker locus linked in coupling phase, the expected segregation ratio for the marker will be as

753 follows[126]:

female gamete	male gamete	
	r M1	$(1-r)$ M2
0.5 M1	$r/2$ M1M1	$(1-r)/2$ M1M2
0.5 M2	$r/2$ M1M2	$(1-r)/2$ M2M2

754

755   In case of $r= 0$ the frequency of heterozygous genotypes for the marker locus is equal to 0.5, and a

756   significant excess of homozygous genotypes for the allele that originated from the self-fertile line (M22)

757   is observed. Distorted segregation of marker loci were statistically analysed for mapping the S locus as

758   outlined by Voylokov et al. (1998)[127].

759

760   Genes affecting low temperature tolerance

761

762   The line 'Puma-SK' was produced by subjecting 'Puma' by recurrent selection under extreme cold winter

763   conditions (-30 °C) to purify for the alleles contributing to increased cold tolerance. 'Puma-SK' was used

764   in an intergeneric cross with the Canadian winter wheat cultivar 'Norstar', which generated a winter

765   wheat introgression line (containing a segment of 5RL from 'Puma' (designated herein as 'Norstar-

766   5A5R') that contained $Fr2$[128].

767   To characterize the $Fr2$ region in 'Puma-SK' and the introgression in 'NorstarPuma5A:5R', whole genome

768   sequencing was performed using the Chromium 10X Genomics platform. Nuclei were isolated from 30

769   seedlings, and high molecular-weight genomic DNA was extracted from nuclei using phenol chloroform

770   according to the protocol of Zheng et al. (2012)[129]. Genomic DNA was quantified by fluorometry using

771   Qubit 2.0 Broad Range (Thermofisher) and size selection was performed to remove fragments smaller

772   than 40 kbp using pulsed field electrophoresis on a Blue Pippin (Sage Science) according to the

773   manufacturer's specifications. Integrity and size of the size selected DNA were determined using a

774   Tapestation 2200 (Agilent), and Qubit 2.0 Broad Range (Thermofisher), respectively. Library preparation

775   was performed as per the 10X Genomics Genome Library protocol

776   (https://support.10xgenomics.com/genome-exome/library-prep/doc/user-guide-chromium-genome-

777   reagent-kit-v2-chemistry) and uniquely barcoded libraries were prepared and multiplexed for

778   sequencing by Illumina HiSeq. De-multiplexing and the generation of fastq files was performed using

779   LongRanger mkfastq (https://support.10xgenomics.com/genome-

780   exome/software/pipelines/latest/using/mkfastq; default parameters).

781   Sequencing reads from 'Puma-SK' and 'NorstarPuma5A:5R' were aligned to the rye line 'Lo7' and bread

782   wheat   cv.   Chinese   Spring[9]   genome   assemblies,   respectively,   using   LongRanger   WGS

783   (https://support.10xgenomics.com/genome-exome/software/pipelines/latest/using/wgs;   arguments   -

784   vcmode 'freebayes'). Large scale structural variants detected by LongRanger were visualized with a

785   combination   of   Loupe   (https://support.10xgenomics.com/genome-

786   exome/software/visualization/latest/what-is-loupe;   tbl.   S-DATAACCESS).   Short   variants   were   called

787   using the Freebayes software (github.com/ekg/freebayes) implemented within the Longranger WGS

788   pipeline. For determining the introgression, 'NorstarPuma5A5R' reads which did not map to the Chinese

789   Spring   reference   were   aligned   to   the   'Lo7'   assembly   using   the   LongRanger   align   pipeline

790   (https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-

791   pipelines). Samtools[114] bedcov was used to calculate the genome-wide read coverage across both

792   references. Copy number variation between 'Puma-SK' and 'Lo7' was detected using a combination of

793   barcode coverage analysis output by the Longranger WGS pipeline, and read depth-of-coverage based

794   analysis using CNVnator[130] and cn.mops[131].

795   To identify differentially expressed genes that may be contributing to the phenotypic differences in cold

796   tolerance, 'Puma-SK' and 'NorstarPuma5A:5R' were grown and crown tissues harvested at different

797   stages of cold acclimation. Both genotypes were grown for 14 days (d) at 20 °C with a 10 hour (h) day

798   length.   Plants   were   then   treated   to   decreasing   temperatures   and   daylengths   over   a   70d   period,

799   designed to mimic field conditions for winter growth habit. After the initial 14 d growth period, the

800   temperature was reduced to 18 °C, then after 3 d (15 °C), 7 d (12 °C), 14 d (9 °C), 21 (6 °C), 28 d (3 °C), 35

801   d (2 °C), 42 d (2 °C), 49 d (2 °C), 56 d (2 °C), 63 d (2 °C), and 70 d (2 °C). In addition to adjusting the

802   temperature, the day length was adjusted incrementally from 13.5 h at 0 d to 9.2 h at 70 d. Day length

803   changes were programmed to occur on day 3 and day 4 of each week. For each change in temperature,

804   crowns were sampled from two independent replicate plants for each genotype, which were used for

805   analysis of gene expression by RNA sequencing. Crown tissue was sampled one hour after the lights

806   came on in the morning to minimize circadian rhythm effects. In addition, at each change in

807   temperature, five plants from each genotype were used to analyze the rate of plant phenological

808   development (dissection of the plant crown to reveal shoot apex development) and cold hardiness

809   during cold acclimation. Cold hardiness was determined using LT50 measurements, the temperature at

810   which 50% of the plants are killed by LT stress, using the procedure outlined by Fowler et al. (2016)[72].

811   Sequencing adapters were removed and low-quality reads were trimmed using Trimmomatic[132]. RNA

812   reads from 'NorstarPuma5A:5R' and 'Puma' were aligned to the 'Lo7' reference using Hisat2[85] (default

813   arguments) and transcripts were quantified with htseq[133]. Differential expression analysis was carried

814   out using DESeq2[134] (default parameters).

815   **Data Availability**

816   Data access information including raw sequence data, selected assembly visualisations, gene

817   annotation, and optical map data, is tabulated in the table S-DATAACCESS.

818

819 **Acknowledgements**

820

Author information

Harrow Research and Development Centre, Agriculture and Agri-Food Canada, 2585 County Road 20,
Harrow, ON, N0R 1G0, Canada

Jamie Larsen

Lethbridge Research and Development Centre, Agriculture and Agri-Food Canada, 5403 1st Avenue
South, Lethbridge AB T1J 4B1, Canada

André Laroche

Chinese Academy of Crop Sciences (CAAS), No.12 Zhongguancun South Street, Haidian District, Beijing
100081, China

Jizeng Jia

Plant Genomics, Earlham Institute, Norwich Research Park, Norwich, Norfolk, NR4 7UG, UK

Anthony Hall, David Swarbreck, Gemy Kaithakottil

Department of Botany, Federal University of Pernambuco, Av. Prof. Moraes Rego, 1235, Cidade
Universitária, Recife PE, 50670-901, Brazil

Mariana Báez

Plant Genome and Systems Biology (PGSB), Helmholtz Zentrum München, Ingolstädter Landstr. 1,
85764 Neuherberg, Germany

Thomas Lux, Heidrun Gundlach, Manuel Spannagl, Klaus F.X. Mayer

National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, No. 1
Shizishan Street, Hongshan District, Wuhan, Hubei Province, China

Qiang Li

HYBRO Saatzucht GmbH & Co. KG, Langlinger Str. 3, 29565 Wriedel, Germany

Dörthe Siekmann

877   Institute of Experimental Botany, Czech Academy of Sciences, Centre of the Region Hana for

878   Biotechnological and Agricultural Research, Šlechtitelů 31, 779 00 Olomouc, Czech Republic

879   Jaroslav Doležel, Jana Čížková, Jan Vrána, Hana Šimková, Helena Toegelová

880   Computational Systems Biology, John Innes Centre, Norwich Research Park, Norwich, NR4 7UH, UK

881   Burkhard Steuernagel

882   Crop Genetics, John Innes Centre, Norwich Research Park, Norwich, NR4 7UH, UK

883   Brande Wulff

884   Institute for Biosafety in Plant Biotechnology, Julius Kühn-Institute, Erwin-Baur-Str. 27, 06484

885   Quedlinburg, Germany

886   Jens Keilwagen

887   Institute for Breeding Research on Agricultural Crops, Julius Kühn-Institute, Rudolf-Schick-Platz 3a,

888   18190 Groß Lüsewitz, Germany

889   Bernd Hackauf

890   Institute for Resistance Research and Stress Tolerance, Julius Kühn-Institute, Erwin-Baur-Str. 27, 06484

891   Quedlinburg, Germany

892   Frank Ordon

893   KWS SAAT SE & Co. KGaA, Grimsehlstr. 31, 37574 Einbeck, Germany

894   ~ and ~

895   Federal State Budgetary Institution of Science Federal Research Center, Kazan Scientific Center of

896   Russian Academy of Sciences, ul. Lobachevskogo, 2/31, Kazan 420111, Tatarstan, Russian Federation

897   Viktor Korzun

898   Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstr. 3, 06466 Stadt Seeland,

899   Germany

900   Andreas Houben, Uwe Scholz, Martin Mascher, Mona Schreiber, Nils Stein, Sudharsan Padmarasu, M.

901   Timothy Rabanus-Wallace, Axel Himmelbach, Andreas Börner

902   **Department of Crop Sciences CiBreed - Center for Integrated Breeding Research, Georg-August**

903   **University Göttingen, Von Siebold Straße 8, D-37075 Göttingen, Germany**

904   Nils Stein

905   **German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany**

906   Martin Mascher

907   **Montana BioAgriculture Inc., Montana, USA**

908   Hikmet Budak

909   **Kansas State University, 4024 Throckmorton Hall, Kansas State University, Manhattan, KS 66506, USA**

910   Jesse Poland, Liangliang Guo, Alan Fritz

911   **Production Systems, Natural Resources Institute Finland (Luke), Latokartanonkaari 9, 00790 Helsinki,**

912   **Finland**

913   Alan H. Schulman

914   **Noble Research Institute, LLC, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA**

915   Xue-Feng Ma

916   **Vavilov Institute of General Genetics, Russian Academy of Sciences, Gubkina 3, 119991 Moscow,**

917   **Russia**

918   Anatoly V. Voylokov

919   **Molecular Biology, Genetics and Bioengineering, Sabanci University, University Cad No 27, Istanbul,**

920   **Turkey**

921   Biyiklioglu Sezgi

922   **Department of Genetics and Biotechnology, Saint Petersburg State University, Universitetskaya emb.**

923   **7/9, 199034, St. Petersburg, Russia**

924 Natalia Tsvetkova

925 **Plant Breeding, Technical University of Munich, Liesel-Beckmann-Str. 2, 80333 München, Germany**

926 Eva Bauer

927 **ARC Centre of Excellence in Plant Energy Biology, School of Molecular Sciences, The University of**

928 **Western Australia, 35 Stirling Highway, Crawley 6009 WA, Australia**

929 Ian Small, Joanna Melonek

930 **Department of Field Crops, University of Cukurova Faculty of Agriculture, Balcalı, Çukurova**

931 **Üniversitesi Rektörlüğü, 01330 Sarıçam/Adana, Turkey**

932 Hakan Ozkan, Uğur Sesiz

933 **Plant Sciences and Landscape Architecture, University of Maryland, College Park, 4291 Fieldhouse**

934 **Drive 2102, Plant Sciences Building, College Park, MD 20742, USA**

935 Vijay Tiwari, Nidhi Rawat

936 **Crop Development Centre, University of Saskatchewan, 51 Campus Drive, Saskatoon, Saskatchewan**

937 **S7N 5A8, Canada**

938 Curtis J. Pozniak

939 **Department of Plant Science, University of Saskatchewan, 51 Campus Drive, Saskatoon, Saskatchewan**

940 **S7N 5A8, Canada**

941 Brook Byrns, Sean Walkowiak, D. Brian Fowler

942 **University of Saskatchewan, Global Institute for Food Security, 110 Gymnasium Place, Saskatoon, SK,**

943 **S7N 0W9, Canada**

944 Andy Sharpe

945 **Aquatic and Crop Resource Development, National Research Council Canada, 110 Gymnasium Pl,**

946 **Saskatoon, SK S7N 0W9, Canada**

947 David Konkin

948   **Department of Plant and Microbial Biology, University of Zürich, Zollikerstrasse 107, 8008 Zürich,**

949   **Switzerland**

950   Beat Keller, Coraline Praz, Thomas Wicker

951   **Department of Biology, ETH Zürich, Wolfgang-Pauli-Strasse 27, 8093 Zürich, Switzerland**

952   Matthias Heuberger

953   **Department of Plant Genetics Breeding and Biotechnology, Warsaw University of Life Sciences -**

954   **SGGW, Nowoursynowska Str 159, 02-776 Warsaw, Poland**

955   Monika Rakoczy-Trojanowska, Hanna Bolibok-Bragoszewska

956   **Department of Genetics, Plant Breeding and Biotechnology, West Pomeranian University of**

957   **Technology Szczecin, Słowackiego 17, 71-434 Szczecin, Poland**

958   Stefan Stojałowski, Beata Myśków

959

960 **Author Contributions**

961 *Project conception and consortium coordination*

962

963 N. S. (leader), K. F. X. M., M. M., V. T., N. R.

964

965 *Manuscript and main figures*

966

967 M. T. R-W. (leader), N. S., B. H., with input from all authors.

968

969 *Genome assembly and data integration*

970

971 M. T. R-W. (leader), M. M.

972

973 *Provision, curation, cultivation, and phenotyping of genetic resources*

974

975 A. B. (*Secale* diversity panel); V. K. ('Lo7'); D. B. F., B. H., Q. L., C. J. P., B. B. ('Norstar', 'Puma'); V. K., B. H.,
976 M. R-T., H. B-B., S. S., B. M. (*Secale* genome size estimation panel).

977

978 *Sequencing data* <mark>*[ADD SECTION FOR FUNDING SEQUENCING?]*</mark>

979

980 J. L., A. L., J. J., A. Hall, D. S., A. Himmelbach, S. P., A. H. S., J. P., B. B.

981

982 *Genome size estimation and chromosome flow sorting*

983

984 J. D., J. Č., J. V.

985

986 *Bionano optical map*

987

988 H. Š., H. T., E. B.

989

990 *FISH*

991

992 M. B., A. Houben.

993

994 *Gene annotation*

995

996 D. S., G. K., T. L., M. S., K. F. X. M., J. K.

997

998 *Repetitive element annotation and analysis*

999

1000 H. G., T. W., M. S., K. F. X. M.

1001

1002 *miRNA annotation*

1003

1004 H. B., B. S.

1005

1006 *Secale diversity analysis*

1007
1008   M. S., M. M., H. S., U. S.
1009
1010   *Hi-C-based SV detection*
1011
1012   M. T. R-W. with input from M. M.
1013
1014   *Resistance gene identification and analysis*
1015
1016   B. S., B. W, B. H., B. K., C. P., T. W.
1017
1018   *SI and CMS gene identification and analysis*
1019
1020   B. H., I. S., J. M.
1021
1022   *Mapping of S- locus*
1023
1024   B. H., A. V. V., N.T.
1025
1026   *Wheat-rye introgression analysis*
1027
1028   J. P., L. G., M. T. R-W., M. M., with input from B. H.
1029
1030   *Low temperature tolerance analysis*
1031
1032   C. J. P., B. B., S. W.
1033
1034
1035

Competing Interests

V. K. is an employee of KWS SAAT SE & Co. KGaA. Dörthe Siekmann is an employee of HYBRO Saatzucht GmbH & Co. KG.

Corresponding Authors

1041

1042   Correspondence to Nils Stein (stein@ipk-gatersleben.de).

1043

References

1. Beck, H.E. *et al.* Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific data* **5**, 180214 (2018).
2. Sharma, S. *et al.* Integrated genetic map and genetic analysis of a region associated with root traits on the short arm of rye chromosome 1 in bread wheat. *Theoretical and Applied Genetics* **119**, 783-793 (2009).
3. Lukaszewski, A.J. Introgressions between wheat and rye. in *Alien introgression in wheat* 163-189 (Springer, 2015).
4. Kim, W., Johnson, J., Baenziger, P., Lukaszewski, A. & Gaines, C. Agronomic effect of wheat-rye translocation carrying rye chromatin (1R) from different sources. *Crop Science* **44**, 1254-1258 (2004).
5. Crespo-Herrera, L.A., Garkava-Gustavsson, L. & Åhman, I. A systematic review of rye (*Secale cereale* L.) as a source of resistance to pathogens and pests in wheat (*Triticum aestivum* L.). *Hereditas* **154**, 14 (2017).
6. Lundqvist, A. Self-incompatibility in rye: I. Genetic control in the diploid. *Hereditas* **42**, 293-348 (1956).
7. Geiger, H. & Schnell, F. Cytoplasmic Male Sterility in Rye (*Secale cereale* L.) 1. *Crop science* **10**, 590-593 (1970).
8. Doležel, J. *et al.* Plant genome size estimation by flow cytometry: inter-laboratory comparison. *Annals of Botany* **82**, 17-26 (1998).
9. IWGSC. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191 (2018).
10. Martis, M.M. *et al.* Reticulate evolution of the rye genome. *The Plant Cell* **25**, 3685-3698 (2013).
11. Bauer, E. *et al.* Towards a whole-genome sequence for rye (*Secale cereale* L.). *The Plant Journal* **89**, 853--869 (2017).
12. Schneider, A., Rakszegi, M., Molnár-Láng, M. & Szakács, É. Production and cytomolecular identification of new wheat-perennial rye (*Secale cereanum*) disomic addition lines with yellow rust resistance (6R) and increased arabinoxylan and protein content (1R, 4R, 6R). *Theoretical and Applied Genetics* **129**, 1045-1059 (2016).
13. Li, J., Zhou, R., Endo, T.R. & Stein, N. High-throughput development of SSR marker candidates and their chromosomal assignment in rye (*Secale cereale* L.). *Plant breeding* **137**, 561-572 (2018).
14. Hackauf, B. *et al.* QTL mapping and comparative genome analysis of agronomic traits including grain yield in winter rye. *Theoretical and applied genetics* **130**, 1801-1817 (2017).
15. Mascher, M. *et al.* A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**, 427--433 (2017).
16. Maccaferri, M. *et al.* Durum wheat genome highlights past domestication signatures and future improvement targets. *Nature genetics* **51**, 885 (2019).
17. Zhu, T. *et al.* Improved genome sequence of wild emmer wheat Zavitan with the aid of optical maps. *G3: Genes, Genomes, Genetics* **9**, 619-624 (2019).
18. Zimin, A.V. *et al.* Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Research* **27**, 787--792 (2017).

1091 19. Braun, E.-M. *et al.* Gene expression profiling and fine mapping identifies a gibberellin 2-
1092 Oxidase gene co-segregating with the dominant dwarfing gene *Ddw1* in rye (*Secale*
1093 *cereale* L.). **10**, 857 (2019).

1094 20. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals
1095 folding principles of the human genome. *Science* **326**, 289-293 (2009).

1096 21. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements.
1097 *Nature Reviews Genetics* **8**, 973 (2007).

1098 22. Naranjo, T., Roca, A., Goicoechea, P. & Giraldez, R. Arm homoeology of wheat and rye
1099 chromosomes. *Genome* **29**, 873-882 (1987).

1100 23. Moore, G., Devos, K., Wang, Z. & Gale, M. Cereal genome evolution: grasses, line up
1101 and form a circle. *Current biology* **5**, 737-739 (1995).

1102 24. Devos, K.M. *et al.* Chromosomal rearrangements in the rye genome relative to that of
1103 wheat. *Theoretical and Applied Genetics* **85**, 673-680 (1993).

1104 25. Dvorak, J. *et al.* Reassessment of the evolution of wheat chromosomes 4A, 5A, and 7B.
1105 *Theoretical and applied genetics* **131**, 2451-2462 (2018).

1106 26. Luo, M. *et al.* Genome sequence of the progenitor of the wheat D genome *Aegilops*
1107 *tauschii. Nature* (2017).

1108 27. Wicker, T. *et al.* Impact of transposable elements on genome structure and evolution in
1109 bread wheat. *Genome biology* **19**, 103 (2018).

1110 28. Wicker, T., Gundlach, H. & Schulman, A.H. The Repetitive Landscape of the Barley
1111 Genome. in *The Barley Genome* 123-138 (Springer, 2018).

1112 29. Dvořák, J. Triticeae genome structure and evolution. in *Genetics and Genomics of the*
1113 *Triticeae* 685-711 (Springer, 2009).

1114 30. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR
1115 Assembly Index (LAI). *Nucleic acids research* **46**, e126-e126 (2018).

1116 31. Presting, G.G., Malysheva, L., Fuchs, J. & Schubert, I. A TY3/GYPSY
1117 retrotransposon-like sequence localizes to the centromeric regions of cereal
1118 chromosomes. *The Plant Journal* **16**, 721-728 (1998).

1119 32. Serrato-Capuchina, A. & Matute, D.R. The role of transposable elements in speciation.
1120 *Genes* **9**, 254 (2018).

1121 33. Schreiber, M., Himmelbach, A., Börner, A. & Mascher, M. Genetic diversity and
1122 relationship between domesticated rye and its wild relatives as revealed through
1123 genotyping-by-sequencing. *Evolutionary applications* **12**, 66-77 (2019).

1124 34. Dvorak, J. *et al.* Structural variation and rates of genome evolution in the grass family
1125 seen through comparison of sequences of genomes greatly differing in size. *The Plant*
1126 *Journal* **95**, 487-503 (2018).

1127 35. Wright, S.I., Ness, R.W., Foxe, J.P. & Barrett, S.C. Genomic consequences of
1128 outcrossing and selfing in plants. *International Journal of Plant Sciences* **169**, 105-118
1129 (2008).

1130 36. Springer, N.M. *et al.* Maize inbreds exhibit high levels of copy number variation (CNV)
1131 and presence/absence variation (PAV) in genome content. *PLoS genetics* **5**, e1000734
1132 (2009).

1133 37. Sun, S. *et al.* Extensive intraspecific gene order and gene structural variations between
1134 *Mo17* and other maize genomes. *Nature genetics* **50**, 1289 (2018).

1135 38. Baack, E., Melo, M.C., Rieseberg, L.H. & Ortiz-Barrientos, D. The origins of reproductive
1136 isolation in plants. *New Phytologist* **207**, 968-984 (2015).

1137 39. Friebe, B., Jiang, J., Raupp, W., McIntosh, R. & Gill, B. Characterization of wheat-alien
1138 translocations conferring resistance to diseases and pests: current status. *Euphytica* **91**,
1139 59-87 (1996).

1140 40. Graybosch, R.A. Mini review: uneasy unions: quality effects of rye chromatin transfers to
1141 wheat. *Journal of Cereal Science* **33**, 3-16 (2001).

41. Keilwagen, J. *et al.* Detecting large chromosomal modifications using short read data from genotyping-by-sequencing. *Frontiers in plant science* **10**, 1133 (2019).

42. Kumlay, A. *et al.* Understanding the effect of rye chromatin in bread wheat. *Crop science* **43**, 1643-1651 (2003).

43. Wricke, G. & Wehling, P. Linkage between an incompatibility locus and a peroxidase isozyme locus (Prx 7) in rye. *Theoretical and applied genetics* **71**, 289-291 (1985).

44. Voylokov, A., Fuong, F. & Smirnov, V. Genetic studies of self-fertility in rye (Secale cereale L.). 1. The identification of genotypes of self-fertile lines for the Sf alleles of self-incompatibility genes. *Theoretical and applied genetics* **87**, 616-618 (1993).

45. Hackauf, B. & Wehling, P. Approaching the self-incompatibility locus Z in rye (Secale cereale L.) via comparative genetics. *Theoretical and Applied Genetics* **110**, 832-845 (2005).

46. Manzanares, C. *et al.* A gene encoding a DUF247 domain protein cosegregates with the S self-incompatibility locus in perennial ryegrass. *Molecular biology and evolution* **33**, 870-884 (2015).

47. Abdel-Ghani, A.H., Parzies, H.K., Ceccarelli, S., Grando, S. & Geiger, H.H. Estimation of quantitative genetic parameters for outcrossing-related traits in barley. *Crop science* **45**, 98-105 (2005).

48. Whitford, R. *et al.* Hybrid breeding in wheat: technologies to improve hybrid wheat seed production. *Journal of experimental botany* **64**, 5411-5428 (2013).

49. Chen, L. & Liu, Y.-G. Male sterility and fertility restoration in crops. *Annual review of plant biology* **65** (2014).

50. Melonek, J., Stone, J.D. & Small, I. Evolutionary plasticity of restorer-of-fertility-like proteins in rice. *Scientific reports* **6**, 35152 (2016).

51. Bernhard, T., Koch, M., Snowdon, R.J., Friedt, W. & Wittkop, B. Undesired fertility restoration in *msm1* barley associates with two *mTERF* genes. *Theoretical and Applied Genetics* **132**, 1335-1350 (2019).

52. Gaborieau, L., Brown, G.G. & Mireau, H. The propensity of pentatricopeptide repeat genes to evolve into restorers of cytoplasmic male sterility. *Frontiers in plant science* **7**, 1816 (2016).

53. Geiger, H., Yuan, Y., Miedaner, T. & Wilde, P. Environmental sensitivity of cytoplasmic genic male sterility (CMS) in Secale cereale L. *Fortschritte der Pflanzenzuechtung* (1995).

54. Geiger, H. Cytoplasmatisch-genische Pollensterilität in Roggenformen iranischer Abstammung. *Naturwissenschaften* **58**, 98-99 (1971).

55. Stojałowski, S., Jaciubek, M.o.a. & Masojć, P. Rye SCAR markers for male fertility restoration in the P cytoplasm are also applicable to marker-assisted selection in the C cytoplasm. *J Appl Genet* **46**, 371-373 (2005).

56. Wilde, P. *et al.* Restorer Plants. (US Patent App. 16/064,304, 2019).

57. Gupta, P.K. *et al.* Hybrid wheat: past, present and future. *Theoretical and Applied Genetics*, 1-21 (2019).

58. Lukaszewski, A.J. Chromosomes 1BS and 1RS for control of male fertility in wheats and triticales with cytoplasms of Aegilops kotschyi, Ae. mutica and Ae. uniaristata. *Theoretical and applied genetics* **130**, 2521-2526 (2017).

59. Tsunewaki, K. Fine mapping of the first multi-fertility-restoring gene, Rf multi, of wheat for three Aegilops plasmons, using 1BS-1RS recombinant lines. *Theoretical and Applied Genetics* **128**, 723-732 (2015).

60. Hohn, C.E. & Lukaszewski, A.J. Engineering the 1BS chromosome arm in wheat to remove the Rf multi locus restoring male fertility in cytoplasms of Aegilops kotschyi, Ae. uniaristata and Ae. mutica. *Theoretical and applied genetics* **129**, 1769-1774 (2016).
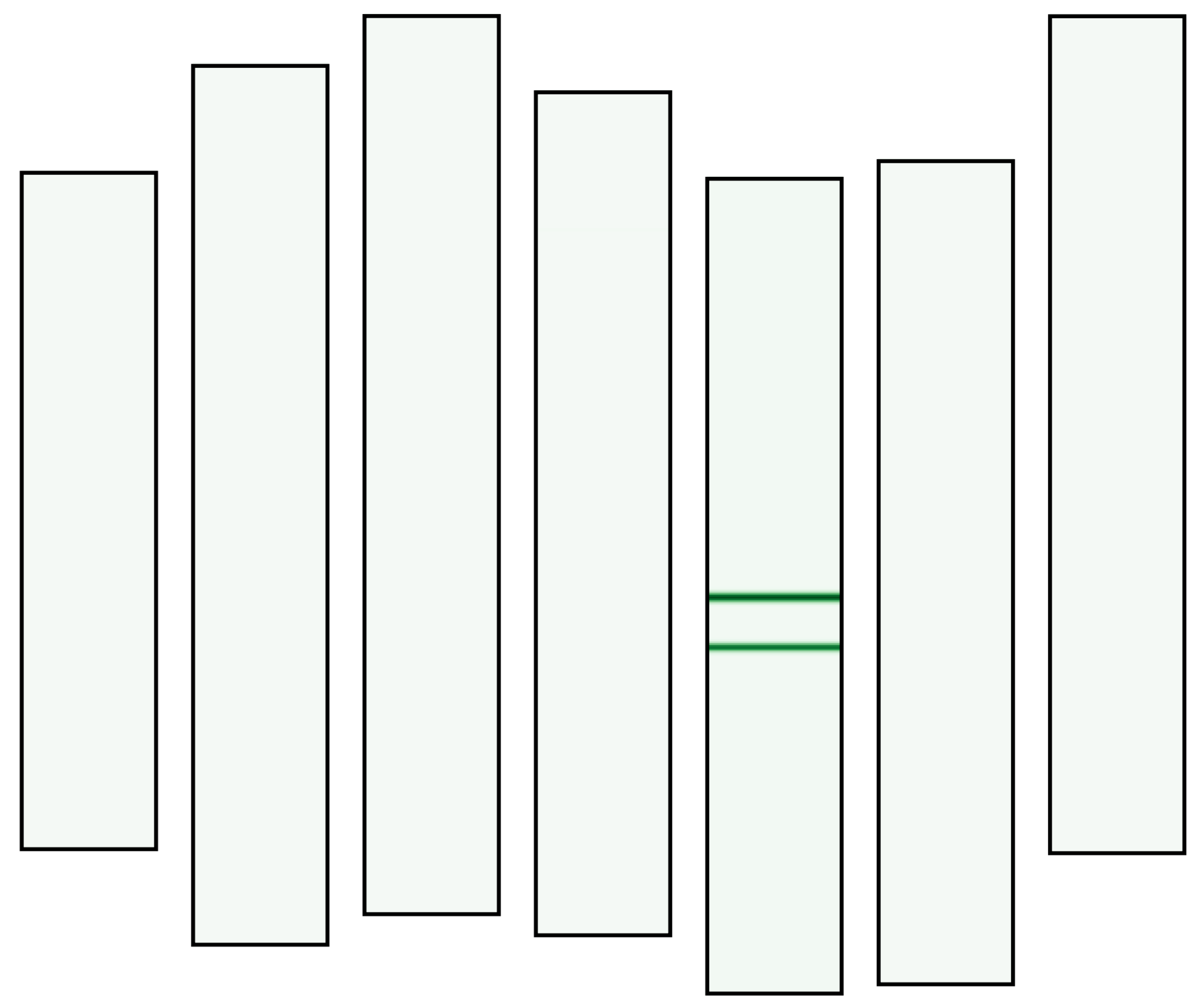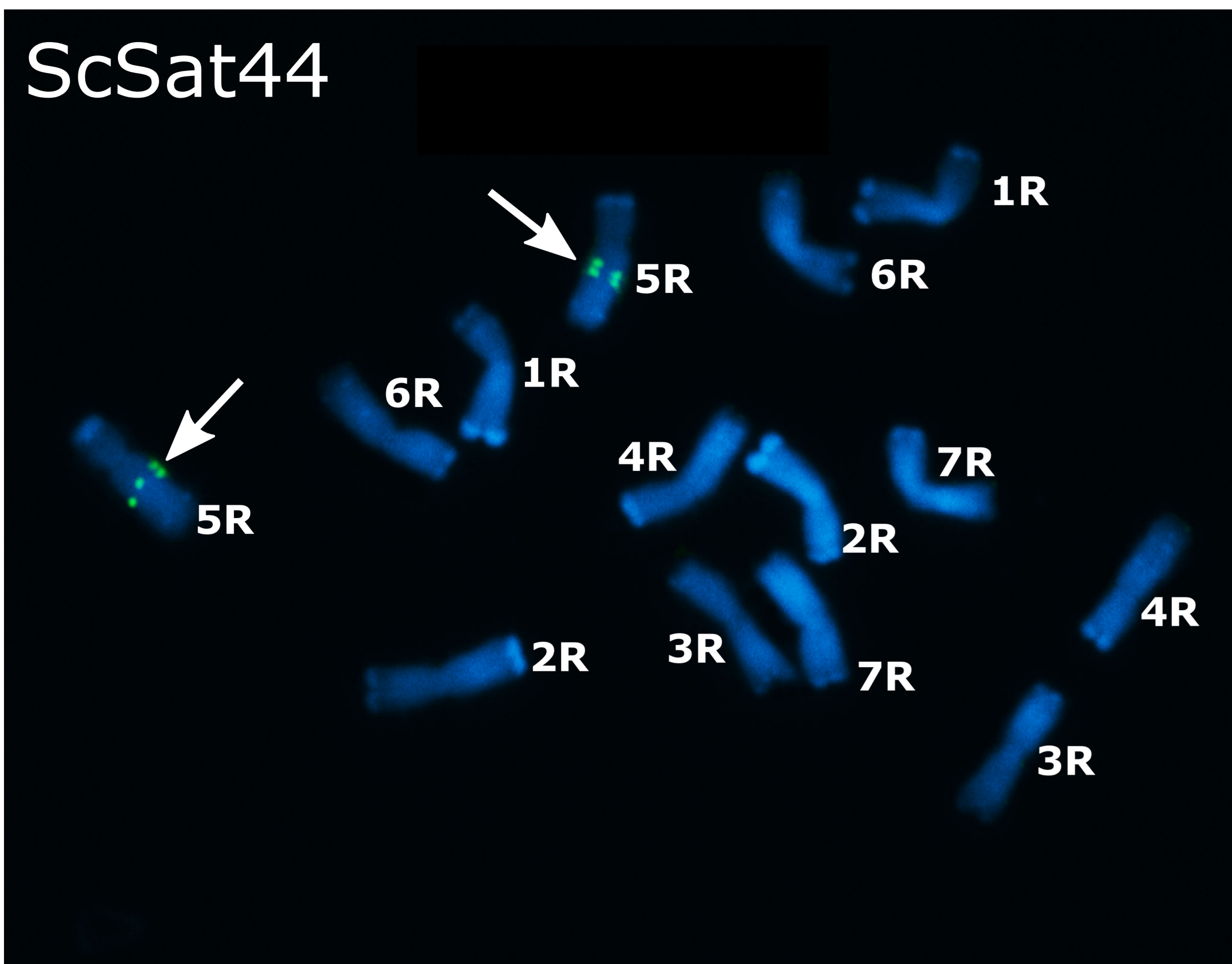
61. Jung, W.J. & Seo, Y.W. Employment of wheat-rye translocation in wheat improvement and broadening its genetic basis. *Journal of Crop Science and Biotechnology* **17**, 305-313 (2014).

62. Kourelis, J. & van der Hoorn, R.A. Defended to the nines: 25 years of resistance gene cloning identifies nine mechanisms for R protein function. *The Plant Cell* **30**, 285-299 (2018).

63. Steuernagel, B. *et al.* Physical and transcriptional organisation of the bread wheat intracellular immune receptor repertoire. (2018).

64. Jordan, T. *et al.* The wheat *Mla* homologue *TmMla1* exhibits an evolutionarily conserved function against powdery mildew in both wheat and barley. *The Plant Journal* **65**, 610-621 (2011).

65. Mago, R. *et al.* The wheat *Sr50* gene reveals rich diversity at a cereal disease resistance locus. *Nature plants* **1**, 15186 (2015).

66. Seeholzer, S. *et al.* Diversity at the *Mla* powdery mildew resistance locus from cultivated barley reveals sites of positive selection. *Molecular plant-microbe interactions* **23**, 497-509 (2010).

67. Dvorak, J. & Fowler, D. Cold Hardiness Potential of Triticale and Teraploid Rye 1. *Crop Science* **18**, 477-478 (1978).

68. Jung, W.J. & Seo, Y.W. Identification of novel C-repeat binding factor (*CBF*) genes in rye (*Secale cereale* L.) and expression studies. *Gene* **684**, 82-94 (2019).

69. Börner, A., Korzun, V., Voylokov, A., Worland, A. & Weber, W. Genetic mapping of quantitative trait loci in rye (*Secale cereale* L.). *Euphytica* **116**, 203-209 (2000).

70. Vágújfalvi, A., Galiba, G., Cattivelli, L. & Dubcovsky, J. The cold-regulated transcriptional activator *Cbf3* is linked to the frost-tolerance locus *Fr-A2* on wheat chromosome 5A. *Molecular Genetics and Genomics* **269**, 60-67 (2003).

71. Båga, M. *et al.* Identification of quantitative trait loci and associated candidate genes for low-temperature tolerance in cold-hardy winter wheat. *Functional & integrative genomics* **7**, 53-68 (2007).

72. Fowler, D., N'Diaye, A., Laudencia-Chingcuanco, D. & Pozniak, C. Quantitative trait loci associated with phenological development, low-temperature tolerance, grain quality, and agronomic characters in wheat (*Triticum aestivum* L.). *PLoS One* **11**, e0152185 (2016).

73. Francia, E. *et al.* Two loci on chromosome 5H determine low-temperature tolerance in a 'Nure'(winter)×'Tremois'(spring) barley map. *Theoretical and Applied Genetics* **108**, 670-680 (2004).

74. Campoli, C., Matus-Cádiz, M.A., Pozniak, C.J., Cattivelli, L. & Fowler, D.B. Comparative expression of *Cbf* genes in the Triticeae under different acclimation induction temperatures. *Molecular Genetics and Genomics* **282**, 141-152 (2009).

75. Akhtar, M. *et al. DREB1/CBF* transcription factors: their structure, function and role in abiotic stress tolerance in plants. *Journal of genetics* **91**, 385-395 (2012).

76. Würschum, T., Longin, C.F.H., Hahn, V., Tucker, M.R. & Leiser, W.L. Copy number variations of *CBF* genes at the *Fr-A2* locus are essential components of winter hardiness in wheat. *The Plant Journal* **89**, 764-773 (2017).

77. Fowler, D., Chauvin, L., Limin, A. & Sarhan, F. The regulatory role of vernalization in the expression of low-temperature-induced genes in wheat and rye. *Theoretical and Applied Genetics* **93**, 554-559 (1996).

78. Galiba, G., Vágújfalvi, A., Li, C., Soltész, A. & Dubcovsky, J. Regulatory genes involved in the determination of frost tolerance in temperate cereals. *Plant Science* **176**, 12-19 (2009).

79. Babben, S. *et al.* Association genetics studies on frost tolerance in wheat (*Triticum aestivum* L.) reveal new highly conserved amino acid substitutions in *CBF-A3, CBF-A15, VRN3* and *PPD1* genes. *BMC genomics* **19**, 409 (2018).

1243 80. Avni, R. *et al.* Wild emmer genome architecture and diversity elucidate wheat evolution
1244   and domestication. *Science* **357**, 93--97 (2017).
1245 81. Ling, H.-Q. *et al.* Draft genome of the wheat A-genome progenitor *Triticum urartu*.
1246   *Nature* **496**, 87--90 (2013).
1247 82. IWGSC. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum
1248   aestivum*) genome. *Science* **345**, 1251788 (2014).
1249 83. Monat, C. *et al.* TRITEX: chromosome-scale sequence assembly of Triticeae genomes
1250   with open-source tools. *BioRxiv*, 631648 (2019).
1251 84. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for
1252   mRNA and EST sequences. *Bioinformatics* **21**, 1859-1875 (2005).
1253 85. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory
1254   requirements. *Nature methods* **12**, 357 (2015).
1255 86. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from
1256   RNA-seq reads. *Nature biotechnology* **33**, 290 (2015).
1257 87. Ghosh, S. & Chan, C.-K.K. Analysis of RNA-Seq data using TopHat and Cufflinks. in
1258   *Plant Bioinformatics* 339-361 (Springer, 2016).
1259 88. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment
1260   search tool. *Journal of molecular biology* **215**, 403-410 (1990).
1261 89. Potter, S.C. *et al.* HMMER web server: 2018 update. *Nucleic Acids Research* **46**, W200-
1262   W204 (2018).
1263 90. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids
1264   research* **34**, W435-W439 (2006).
1265 91. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: Homology-Based Gene Prediction
1266   Utilizing Intron Position Conservation and RNA-seq Data. in *Gene Prediction* 161-177
1267   (Springer, 2019).
1268 92. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for
1269   the analysis of massive data sets. *Nature biotechnology* **35**, 1026 (2017).
1270 93. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using
1271   EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**,
1272   R7 (2008).
1273 94. Wicker, T. *et al.* DNA transposon activity is associated with increased mutation rates in
1274   genes of rice and other grasses. *Nature communications* **7**, 12790 (2016).
1275 95. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular biology open
1276   software suite. (Elsevier current trends, 2000).
1277 96. Ma, J. & Bennetzen, J.L. Rapid recent growth and divergence of rice nuclear genomes.
1278   *Proceedings of the National Academy of Sciences* **101**, 12404-12410 (2004).
1279 97. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids
1280   research* **27**, 573-580 (1999).
1281 98. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high
1282   throughput. *Nucleic acids research* **32**, 1792-1797 (2004).
1283 99. Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and
1284   deep-sequencing data. *Nucleic acids research* **39**, D152-D157 (2010).
1285 100. Markham, N.R. & Zuker, M. UNAFold. in *Bioinformatics* 3-31 (Springer, 2008).
1286 101. Alptekin, B., Akpinar, B.A. & Budak, H. A comprehensive prescription for plant miRNA
1287   identification. *Frontiers in plant science* **7**, 2058 (2017).
1288 102. Akpinar, B.A., Kantar, M. & Budak, H. Root precursors of microRNAs in wild emmer and
1289   modern wheats show major differences in response to drought stress. *Functional &
1290   integrative genomics* **15**, 587-598 (2015).
1291 103. Dai, X., Zhuang, Z. & Zhao, P.X. psRNATarget: a plant small RNA target analysis server
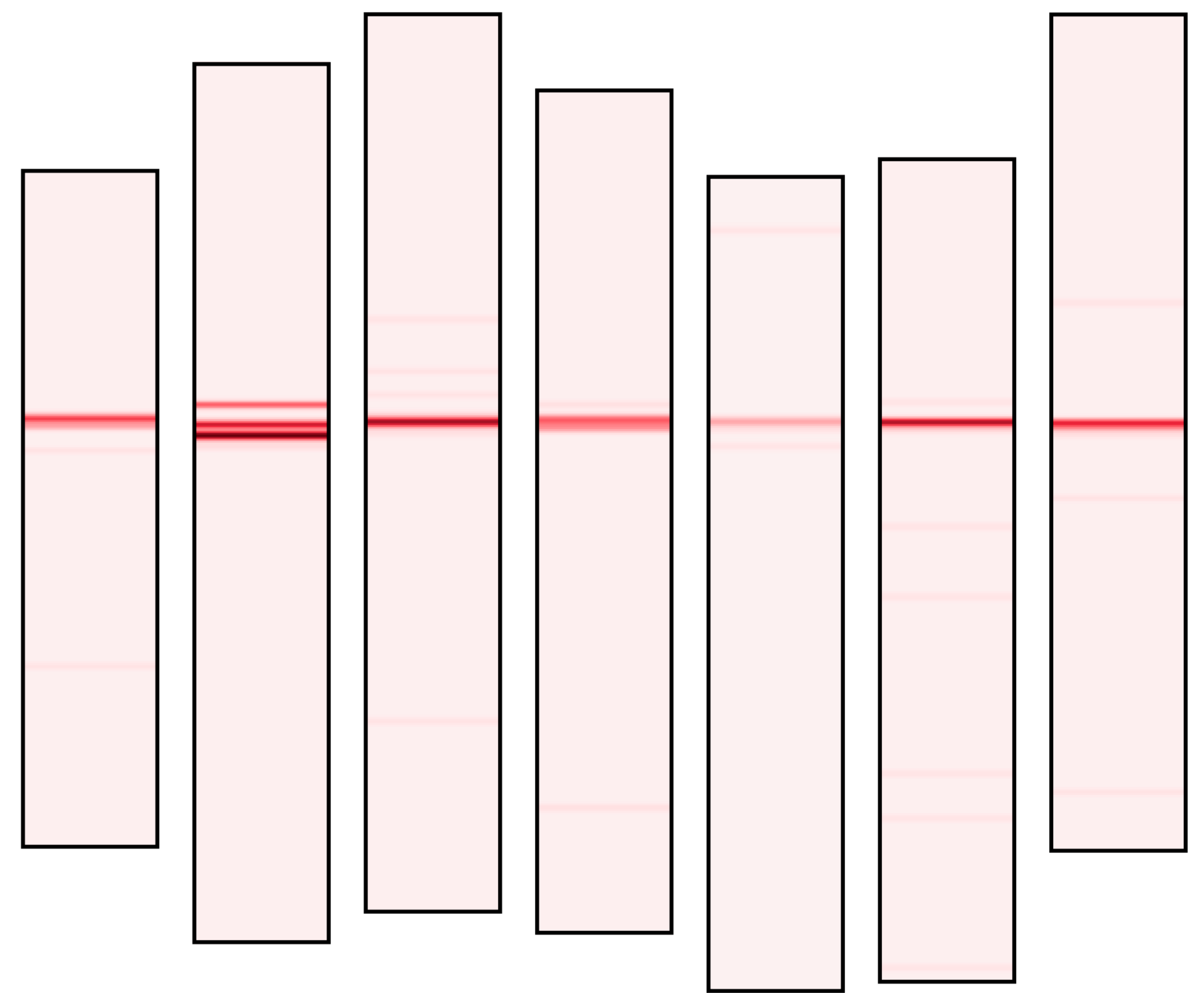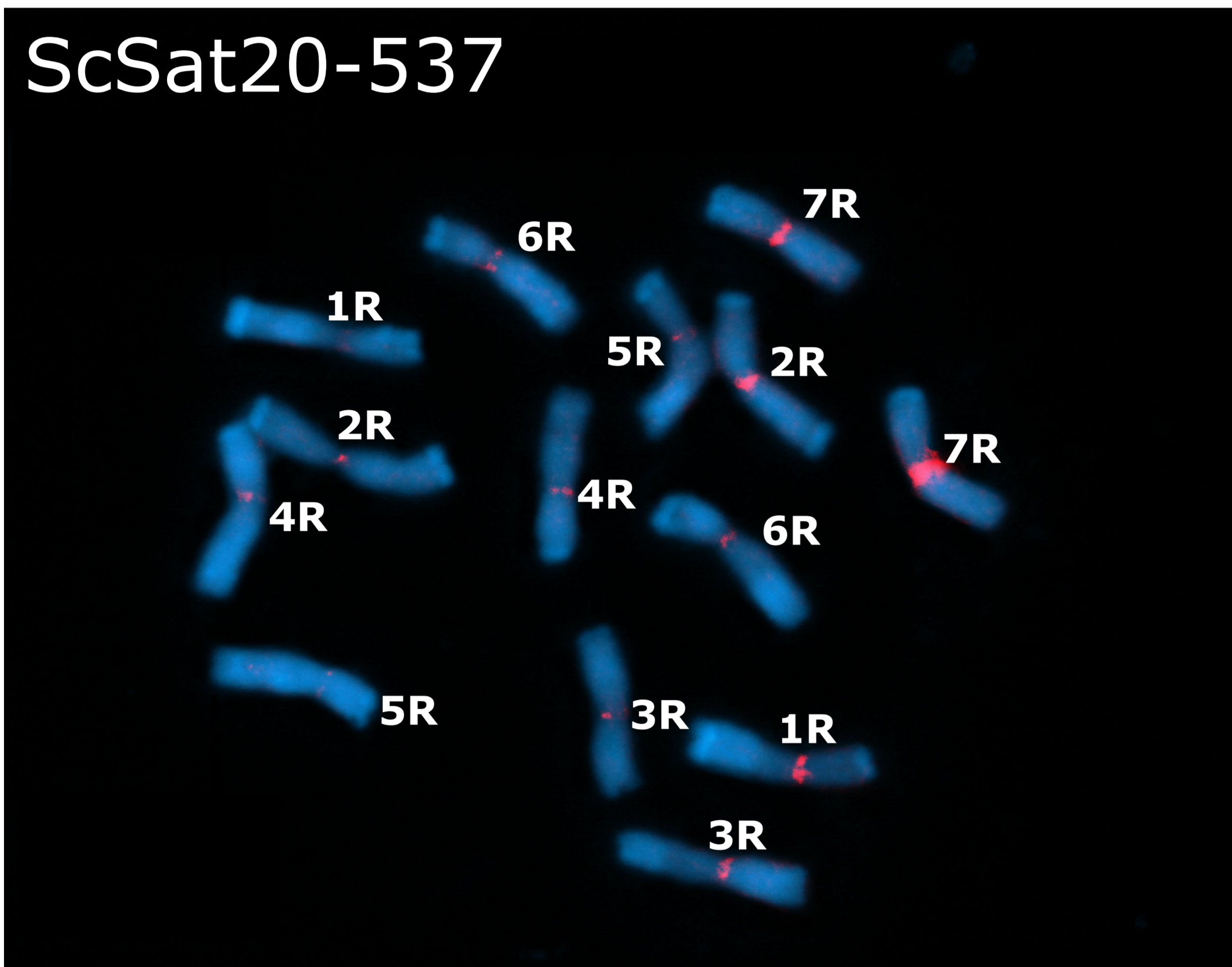1292   (2017 release). *Nucleic acids research* **46**, W49-W54 (2018).

1293 104. Dai, X. & Zhao, P.X. psRNATarget: a plant small RNA target analysis server. *Nucleic*
1294   *acids research* **39**, W155-W159 (2011).
1295 105. Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in
1296   functional genomics research. *Bioinformatics* **21**, 3674-3676 (2005).
1297 106. Aliyeva-Schnorr, L., Ma, L. & Houben, A. A fast air-dry dropping chromosome
1298   preparation method suitable for FISH in plants. *JoVE (Journal of Visualized*
1299   *Experiments)*, e53470 (2015).
1300 107. Cuadrado, A., Jouve, N. & Ceoloni, C. Variation in highly repetitive DNA composition of
1301   heterochromatin in rye studied by fluorescence *in situ* hybridization. *Genome* **38**, 1061-
1302   1069 (1995).
1303 108. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and
1304   evolutionary analyses in R. *Bioinformatics* **35**, 526-528 (2018).
1305 109. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal
1306   component analysis of SNP data. *Bioinformatics* **28**, 3326-3328 (2012).
1307 110. Pont, C. *et al.* Tracing the ancestry of modern bread wheats. *Nature genetics* **51**, 905
1308   (2019).
1309 111. Rife, T.W., Graybosch, R.A. & Poland, J.A. Genomic analysis and prediction within a US
1310   public collaborative winter wheat regional testing nursery. *The plant genome* **11**(2018).
1311 112. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler
1312   transform. *Bioinformatics* **26**, 589-595 (2010).
1313 113. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing
1314   reads. *EMBnet. journal* **17**, 10-12 (2011).
1315 114. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-
1316   2079 (2009).
1317 115. Sonnhammer, E.L. & Durbin, R. A dot-matrix program with dynamic threshold control
1318   suited for genomic DNA and protein sequence analysis. *Gene* **167**, GC1-GC10 (1995).
1319 116. Thompson, J.D., Gibson, T.J. & Higgins, D.G. Multiple sequence alignment using
1320   ClustalW and ClustalX. *Current protocols in bioinformatics*, 2.3. 1-2.3. 22 (2003).
1321 117. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model
1322   choice across a large model space. *Systematic biology* **61**, 539-542 (2012).
1323 118. Finn, R.D., Clements, J. & Eddy, S.R. HMMER web server: interactive sequence
1324   similarity searching. *Nucleic acids research* **39**, W29-W37 (2011).
1325 119. Cheng, S. *et al.* Redefining the structural motifs that determine RNA binding and RNA
1326   editing by pentatricopeptide repeat proteins in land plants. *The Plant Journal* **85**, 532-
1327   547 (2016).
1328 120. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic acids research*
1329   **47**, D427-D432 (2018).
1330 121. Hackauf, B., Rudd, S., Van der Voort, J., Miedaner, T. & Wehling, P. Comparative
1331   mapping of DNA sequences in rye (*Secale cereale* L.) in relation to the rice genome.
1332   *Theoretical and applied genetics* **118**, 371-384 (2009).
1333 122. Hackauf, B. & Wehling, P. Identification of microsatellite polymorphisms in an expressed
1334   portion of the rye genome. *Plant Breeding* **121**, 17-25 (2002).
1335 123. Langridge, P., Baumann, U. & Juttner, J. Revisiting and revising the self-incompatibility
1336   genetics of Phalaris coerulescens. *The Plant Cell* **11**, 1826-1826 (1999).
1337 124. Li, X., Nield, J., Hayman, D. & Langridge, P. Cloning a putative self-incompatibility gene
1338   from the pollen of the grass *Phalaris coerulescens*. *The Plant Cell* **6**, 1923-1932 (1994).
1339 125. Hackauf, B., Korzun, V., Wortmann, H., Wilde, P. & Wehling, P. Development of
1340   conserved ortholog set markers linked to the restorer gene Rfp1 in rye. *Molecular*
1341   *breeding* **30**, 1507-1518 (2012).

1342 126. Wagner, H., Weber, W. & Wricke, G. Estimating linkage relationship of isozyme markers
1343   and morphological markers in sugar beet (*Beta vulgaris* L.) including families with
1344   distorted segregations. *Plant Breeding* **108**, 89-96 (1992).
1345 127. Voylokov, A., Korzun, V. & Börner, A. Mapping of three self-fertility mutations in rye
1346   (*Secale cereale* L.) using RFLP, isozyme and morphological markers. *Theoretical and
1347   Applied Genetics* **97**, 147-153 (1998).
1348 128. Fowler, D.B. Cold acclimation threshold induction temperatures in cereals. *Crop Science*
1349   **48**, 1147-1154 (2008).
1350 129. Zhang, M. *et al.* Preparation of megabase-sized DNA from a variety of organisms using
1351   the nuclei method for advanced genomics research. *Nature Protocols* **7**, 467 (2012).
1352 130. Abyzov, A., Urban, A.E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover,
1353   genotype, and characterize typical and atypical CNVs from family and population
1354   genome sequencing. *Genome research* **21**, 974-984 (2011).
1355 131. Klambauer, G. *et al.* cn. MOPS: mixture of Poissons for discovering copy number
1356   variations in next-generation sequencing data with a low false discovery rate. *Nucleic
1357   acids research* **40**, e69-e69 (2012).
1358 132. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
1359   sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
1360 133. Anders, S., Pyl, P.T. & Huber, W. HTSeq—a Python framework to work with high-
1361   throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).
1362 134. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion
1363   for RNA-seq data with DESeq2. *Genome biology* **15**, 550 (2014).
1364

ScSat44

ScSat20-537

ScSat18

10 μm

500 Mbp

1R   2R   3R   4R   5R   6R   7R

a) Radial phylogenetic tree with tip labels: Lo7 *S. cereale*, Lo225 *S. cereale*, R 1003 *S. vavilovii*, R 925 *S. strictum*, R 2446 *S. sylvestre*

Legend:
- *S. vavilovii*
- *S. vavilovii* x *cereale*
- *S. cereale* x *vavilovii*
- *S. cereale*
- *S. cereale* subsp. *rigidum*
- *S.* x *derzhavinii*
- *S. cereale* subsp. *dighoricum*
- *S. cereale* subsp. *segetale*
- *S. cereale* subsp. *afghanicum*
- *S. cereale* subsp. *ancestrale*
- *S. strictum* subsp. *kuprijanovii*
- *S. stricutm* subsp. *anatolicum*
- *S. strictum*
- *S. strictum* subsp. *africanum*
- *S. sylvestre*

b) PC2 (6.56%) vs PC1 (10.04%)

c) PC3 (3%) vs PC2 (6.56%)

d) PC3 (3%) vs PC1 (10.04%)

Legend:
- *S. cereale*
- *S. strictum*
- *S. sylvestre*
- *S. vavilovii*

a mTERF density

b PLS-PPR density

c P-PPR density

d RFL-PPR density

e NLR gene density

f Rfp-$4R$ · $Rfc1$-$4R$ · $Fr2$ · $Vrn1$ · Rfp-$6R$ · Rf^{multi} · Rfp-$1R$ · $Pm2$

Position on pseudomolecule (Mbp)

chr1RS detail — $Lr10$ $RGA2$ $Pm3$ Mla — 10 Mbp

g $Pm2$
- Clade A
- Clade B
5A, 5B, 5D, 5R, 6H

h $Lr10/RGA2$
1A_H1, 1A_H2, 1A, 1R
RGA2, Lr10, RGA2-a, RGA2-b, RGA2-a1, RGA2-a2

i $Pm3$
- Clade A
- Clade B
- Clade C
- Basal clades
1A, 1B, 1D, 1R, 1H

j Mla
- Clade A
- Clade B
- Clade C
1A, 1B, 1D, 1R, 1H, ChrUn

k Rf^{multi}
1RS, 1BS
SECCE1Rv1G0008410.1
TraesCS1B02G072900.1
TraesCS1B02G071642.1
100 Kbp

l $Vrn1$
Del, Dup, Del, Dup
Puma 10X molecule Coverage
60x / 0x
2.6 Mbp
SECCE5Rv1G0353290 (MADS-box transcription factor)
Del
15 kbp

m $Fr2$
Copy number difference between CBF homologs
Lo7 — Puma
Estimated Copy Number
Dup
Puma 10X molecule Coverage
60x / 0x
2.1 Mbp

a

NorstarPuma5A:5R
Translocation

pSc119.2
GAA

Norstar 5A

GAA
Afa repeat

b

Read Depth

Wheat | Rye

5A | 5B | 5D | 5R

0 250 500 750 | 0 250 500 750 | 0 250 500 | 0 250 500 750

Chromosome Position (Mbp)

c

Puma —— Norstar5A:5R ——

Gene Expression of CBFs with CNV in Puma
(Normalised transcript count)

SECCE5Rv1G0340450

800
400
0

SECCE5Rv1G0340460

800
400
0

SECCE5Rv1G0340480

800
400
0

SECCE5Rv1G0340490

800
400
0

LT Threshold

Norstar
Norstar5A:5R
Puma

0 3 7 14 21 28 35 42 49 56 63 70

Treatment Days
(Decreasing Temperature and Photoperiod)

0°C   -30°C