The evolution of transposable elements in *Brachypodium distachyon* is governed by purifying selection, while neutral and adaptive processes play a minor role

Robert Horvath[1*], Nikolaos Minadakis[1], Yann Bourgeois[2,3], Anne C. Roulin[1*]

[1] Department of Plant and Microbial Biology, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland

[2] DIADE, University of Montpellier, CIRAD, IRD, Montpellier, France

[3] University House, Winston Churchill Ave, Portsmouth PO1 2UP, United Kingdom

[*] **Corresponding author**: robert.horvath@bluewin.ch; anne.roulin@botinst.uzh.ch

Abstract

Background: Understanding how plants adapt to changing environments and the potential contribution of transposable elements (TEs) to this process is a key question in evolutionary genomics. While TEs have recently been put forward as active players in the context of adaptation, few studies have thoroughly investigated their precise role in plant evolution. Here we used the wild Mediterranean grass *Brachypodium distachyon* as a model species to identify and quantify the forces acting on TEs during the adaptation of this species to various conditions, across its entire geographic range.

Results: Using sequencing data from more than 320 natural *B. distachyon* accessions and a suite of population genomics approaches, we reveal that putatively adaptive TE polymorphisms are rare in wild *B. distachyon* populations. After accounting for changes in past TE activity, we show that only a small proportion of TE polymorphisms evolved neutrally (< 10%), while the vast majority of them are under moderate purifying selection regardless of their distance to genes.

Conclusions: TE polymorphisms should not be ignored when conducting evolutionary studies, as they can be linked to adaptation. However, our study clearly shows that while they have a large potential to cause phenotypic variation in *B. distachyon*, they are not favored during evolution and adaptation over other types of mutations (such as point mutations) in this species.

Key words: Transposable elements, adaptation, *Brachypodium distachyon*, natural selection, age-adjusted SFS

Background

Transposable elements (TEs) are an intrinsic part of eukaryotic genomes and their evolution [1-12]. In addition to modulating genome size, the ability of TEs to create genetic diversity through insertion and excision events can lead to new phenotypes on which selection can act. TEs can alter phenotypes through various mechanisms, including the functional disruption of genes [1, 2], large-scale changes in the regulatory apparatus [3, 4], alteration of epigenetic landscapes [5, 6], ectopic recombination and structural rearrangements [7, 8]. In plants, the dynamics of TE loss and proliferation play a major role in genome evolution [e.g., 9-12]. TEs therefore constitute potentially important drivers of plant evolution, both in nature and during domestication [13].

Beyond their influence on genome structure, and given that their transpositional activity can be stress-inducible [for review 14], TEs are often regarded as more likely than classical point mutations to produce the diversity needed for individuals to respond quickly to challenging environments [15-17]. For instance, punctual TE polymorphisms can lead to gains of fitness and evolve under positive selection [2, 20-25]. TE polymorphisms can even induce more extreme changes in gene expression than single nucleotide polymorphisms (SNPs) in plants [18, 19].

Despite such evidence, whether TE polymorphisms are major contributors to adaptation to changing environments is still debated. Indeed, TE transposition can be disruptive, and purifying selection has been shown to play an important role in TE evolution [e.g., 30, 32]. Based on simulations, it has been suggested that the persistence of TE polymorphisms within a genome without an uncontrolled accumulation, can only be achieved if weak purifying selection is the main force governing TE evolution [33-36]. The uncertainty surrounding this important question in evolutionary genomics results from the limited number of studies that comprehensively tested

55 the extent to which selection shapes TE allele frequencies, both in plants [25, 26] and animals
56 [27-31] and characterized the distribution of fitness effects of new TE insertions. To clarify this
57 question, we used the plant model system *Brachypodium distachyon* [37] to disentangle the
58 effects of purifying and positive selection on TE polymorphisms in natural populations.

59 *B. distachyon* is a wild annual grass endemic to the Mediterranean basin and Middle East.
60 Recent genetic studies based on more than 320 natural accessions spanning from Spain to Iraq
61 (hereafter referred to as the *B. distachyon* diversity panel) revealed that *B. distachyon* accessions
62 cluster into three main genetic lineages (the A, B and C genetic lineages), which further divide
63 into five main genetic clades that display little evidence for historical gene flow (Fig. 1A; [38, 39]).
64 Niche modeling analyses suggest that the species moved southward during the last glacial period
65 and recolonized Europe and the Middle East within the last five thousand years [39].
66 Consequently, while some *B. distachyon* genetic clades currently occur in the same broad
67 geographical areas (Fig. 1A), natural accessions are adapted to a mosaic of habitats [38, 39].
68 These past and more recent shifts in the species distribution led to clear footprints of positive
69 selection in the genome [39, 40] and make *B. distachyon* an ideal study system to investigate the
70 contribution of TEs to the adaptation of plants in the context of environmental changes.

71 In *B. distachyon*, TEs are exhaustively annotated and account for approximately 30% of
72 the genome [37]. Recent TE activity has been reported for many families, but despite past
73 independent bottlenecks and expansions experienced by the different genetic clades, no lineage-
74 specific TE family activity has been observed [32]. Rather, TE activity tends to be homogeneous
75 throughout the species range and across genetic clades, indicating a high level of conservation of
76 the TE regulatory apparatus [32]. While purifying selection shapes the accumulation patterns of

77 TEs in this species [32], some TE polymorphisms have been observed in the vicinity of genes [32],
78 potentially affecting gene expression [41]. These early studies, based on a relatively small number
79 of accessions originating exclusively from Spain and Turkey, suggested that TE polymorphisms
80 could contribute to functional divergence and local adaptation in *B. distachyon* [32].

81 To test this hypothesis, we used the *B. distachyon* diversity panel to identify TE
82 polymorphisms in a large set of 326 natural accessions spanning the whole species distribution.
83 We combined a set of population genomic analyses to assess the proportion of TE polymorphisms
84 associated with positive or purifying selection as well as neutral evolution. We also quantified
85 the strength of purifying selection through forward simulations. Altogether, our work provides
86 the first quantitative estimate of the adaptive, neutral, and disruptive potential of TEs, while
87 accounting for changes in TE activity, in a plant harboring a relatively small genome. Altogether,
88 our result advocate against an extended role of TEs in recent adaptation.

89

90 Results

91 Genetic variation in *Brachypodium distachyon*

92 Using the *B. distachyon* diversity panel (Fig. 1A), we identified 97,660 TE polymorphisms in our
93 *B. distachyon* dataset, of which 9,172 were retrotransposons, 52,249 were DNA-transposons and
94 36,239 were unclassified. We also identified 9 million SNPs across the 326 samples, including
95 182,801 synonymous SNPs. A Principal Component Analysis (PCA) performed either with SNPs or
96 TE polymorphisms reflects the previously described population structure of *B. distachyon* [38,
97 39], with the first two components of the PCA splitting the data according to the demographic
98 structure (Additional file 1: Fig. S1). Investigating the genetic variation caused by

retrotransposons and DNA-transposons revealed that the observed diversity in retrotransposons strongly correlated with the demographic structure (Mantel test; r = 0.79, *p* value = 0.001), while the observed diversity in DNA-transposons only had a weaker correlation (Mantel test; r = 0.36, *p* value = 0.001) with the demographic structure (Additional file 1: Fig. S2).

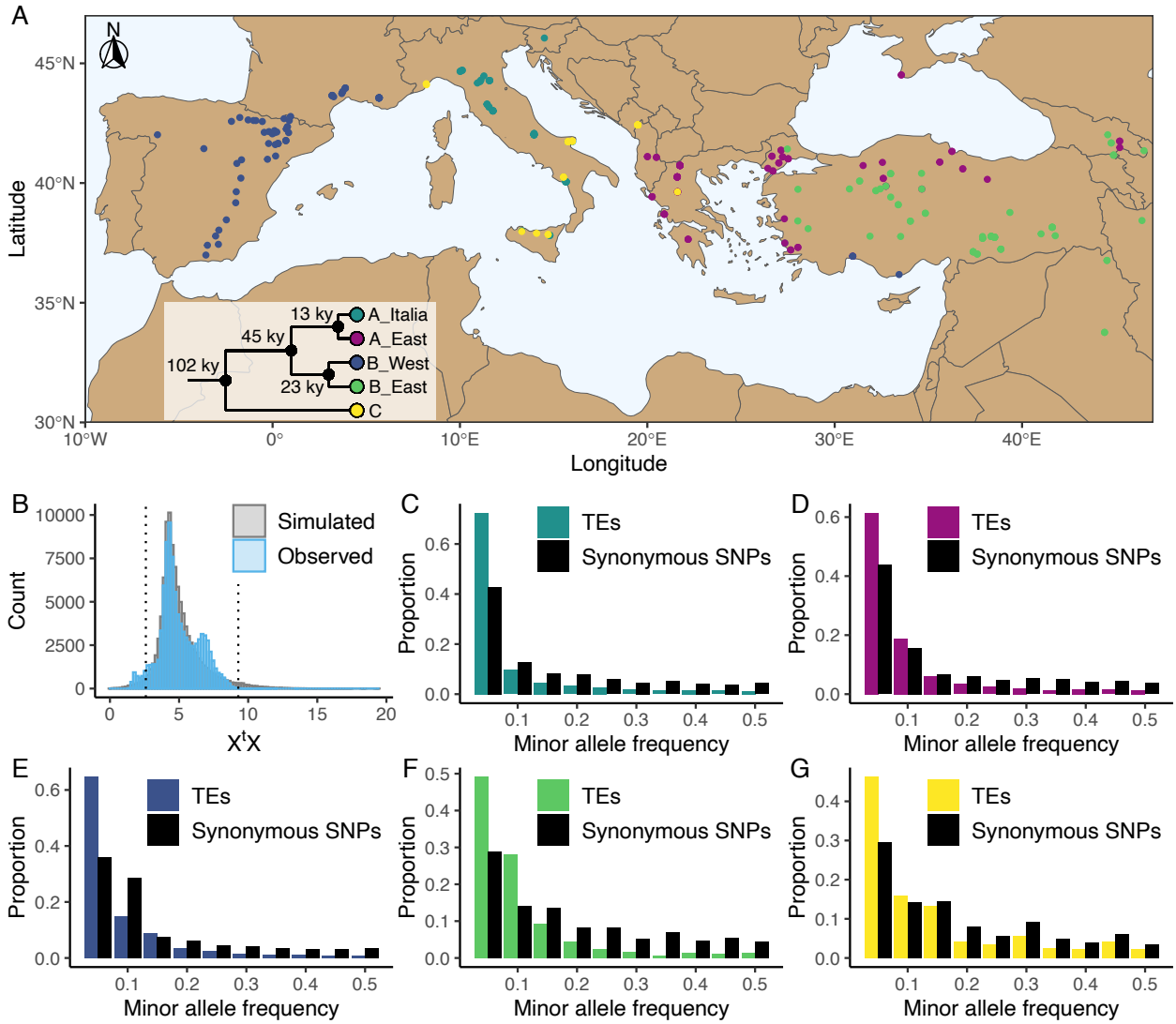

Fig 1. Distribution of the studied accessions and TE polymorphism frequencies. Panel A: Map showing the geographical distribution of the accessions used in the current study. The phylogenetic tree illustrates the phylogeny between the five genetic clades. This panel was made based on the data and results published by Stritt et al. [38] and Minadakis et al. [39]. Panel B: Observed (blue) and simulated (gray) X^tX values of TE polymorphisms in *B. distachyon*. Dotted lines show the 2.5% and 97.5% quantiles of the simulated X^tX values. Panel C-G: Folded site frequency spectrum of TE polymorphisms and synonymous SNPs in all clades. C: A_East; D: A_Italia; E: B_West; F: B_East; G: C.

103 From the initial TE and SNP dataset, we could estimate the time of origin in generations
104 (age) of 50,891 TE polymorphisms and 108,855 synonymous SNPs based on pairwise differences
105 in identity by descent (IBD) regions around the focal mutation (see Materials and methods). The
106 results of the age estimate analysis were checked by contrasting the observed correlation
107 between allele age and frequency of synonymous SNPs to the theoretical predictions of Kimura
108 and Ohta [42] for neutrally evolving mutations. We found that, the observed correlation matched
109 expectations (Additional file 1: Fig. S3), with older alleles found on average at higher frequencies
110 than younger ones. Furthermore, most TE polymorphisms in our dataset were young and only a
111 few were very old (Additional file 1: Fig. S4).

112

113 **The overall contribution of TEs to clade differentiation and adaptation is limited**

114 To examine the overall contribution of TEs to evolution and adaptation in *B. distachyon*, we first
115 identified regions of the genomes that were likely affected by recent selective sweeps. The fast
116 increase in the frequency of a beneficial allele is expected to lead to a longer than average
117 haplotype around the mutation under positive selection. Such events (known as selective
118 sweeps) can be identified by computing the integrated haplotype score (iHS) around focal
119 mutations [43]. We therefore computed iHS along the genome for the four derived genetic
120 clades. Regions of the genomes with significantly higher iHS than average are expected to harbor
121 mutations that were under positive selection during evolution and adaptation. We hypothesized
122 that if TEs constitute an important part of the genetic makeup that led to adaptation in a given
123 genetic clade, then they should be more frequently fixed or at higher frequencies in regions with

124 high iHS than in the corresponding regions that did not experience recent selective sweeps in
125 other clades.

126 First, we tested if more TE polymorphisms were fixed in a specific region of the genome
127 if a genetic clade had a high iHS, and presumably experienced a selective sweep, than in other
128 genetic clades. An analysis of covariance (ANCOVA) revealed that the number of fixed TE
129 polymorphisms per clade did not significantly differ between high iHS regions and the same
130 regions in other clades (Table 1). These results indicate that there is no correlation between the
131 overall number of fixed TE polymorphisms per clade in a region and recent selective sweeps.
132 However, the number of fixed TEs in genomic regions along the genome was significantly affected
133 by the total number of TEs in the region, the TE superfamily, the TE age, the genetic clade and
134 the overall genetic features of the region (e.g., recombination rate, see Materials and methods)
135 but not by the iHS itself (Table 1). Similarly, we tested if the allele frequency of TE polymorphisms
136 was significantly higher in a specific region of the genome if a genetic clade had a high iHS than
137 in other genetic clades. A second ANCOVA revealed that the allele frequency of TE
138 polymorphisms was significantly influenced by the TE superfamily, TE age, clade and overall
139 genetic features of the region but not by the iHS (Table 2). These results indicate that TEs in high
140 iHS regions did not experience a significant increase in their frequency and that TEs in high iHS
141 regions are experiencing the same selective constraints as other TEs.

142

143

144

145 **Table 1** ANCOVA predicting the number of fixed TE polymorphisms per clade in candidate regions
146 under positive selection.

Variable	Sum of squares	degrees of freedom	*F* value	*P* value
Total number of TEs in the region	28969.6	1	35405.64	< 0.001
TE superfamily	887.5	14	77.48	< 0.001
Clade	587	3	239.13	< 0.001
Genomic region	136.7	80	2.09	< 0.001
TE age	45.5	2	27.81	< 0.001
High iHS	0	1	0.03	0.869

147

148

149 **Table 2** ANCOVA predicting the allele frequency of TE polymorphisms per clade in candidate
150 regions under positive selection.

Variable	Sum of squares	degrees of freedom	*F* value	*P* value
TE superfamily	453.2	14	247.3	< 0.001
Clade	17.7	3	45.18	< 0.001
Genomic region	147	80	14	< 0.001
TE age	2	2	7.7	< 0.001
High iHS	0.1	1	0.79	0.374

151

152 A complementary approach to explore the impact of positive selection on TEs consists in
153 investigating their genetic differentiation among populations. Using the five genetic clades as
154 focal populations, we computed X^tX values, a standardized measure of genetic differentiation
155 corrected for the neutral covariance structure across populations [44, 45], for each TE
156 polymorphism. Mutations affected by positive selection are expected to be over-differentiated
157 between clades and display significantly higher X^tX values than other mutations [45]. In contrast,
158 a low X^tX value implies that the mutation is less differentiated than other mutations and
159 potentially evolves under balancing selection, whereas purifying selection and a neutral
160 evolution are not expected to impact the differentiation of a mutation among populations [44].
161 We contrasted the observed X^tX values computed for each TE polymorphism to a simulated
162 pseudo-observed dataset (simulated observations under the demographic model inferred from
163 the covariance matrix of the SNP dataset, for more details see [45]) and found that only a small
164 fraction of the TE polymorphisms (0.06%) displayed X^tX values higher than the 97.5% quantile of
165 the simulated values (Fig. 1B). This indicates that only a few TE polymorphisms are over-
166 differentiated among genetic clades and might have been affected by positive selection.
167 However, a relatively larger portion of the TE polymorphisms (4.3%) displayed X^tX values smaller
168 than the 2.5% quantile of the simulated values (Fig. 1B), indicating that balancing selection might
169 also shape TE frequency in *B. distachyon.*

170 To further examine the contribution of TEs to adaptation, we tested whether and how
171 many TE polymorphisms were significantly associated with environmental factors. If the presence
172 of a TE provides an advantage in a certain environment and contributes to adaptation, we
173 expected a correlation between the environment and the presence/absence of this TE. In this

174 context, we performed genome-environment association analyses (GEA) using all TEs and SNPs
175 identified across the 326 samples and 32 environmental factors associated with precipitation,
176 solar radiation, temperature, elevation and aridity (see in Materials and methods for the full list).
177 The GEA revealed that only nine of the 97,660 TE polymorphisms were significantly associated
178 with some environmental factors (Additional file 2: Table S1), confirming that TEs only had a
179 limited contribution to adaptation in *B. distachyon.* Importantly, two of these nine TEs were
180 found in a gene, and three were in the vicinity of genes (less than 2 kilobase (kb) away, Additional
181 file 2: Table S1).

182

183 **Purifying selection dominates the evolution of TE polymorphisms in *B. distachyon***

184 To further characterize the forces governing the evolution of TE polymorphisms in *B. distachyon*,
185 we examined the genome-wide frequency distribution of TEs. We first computed the folded site
186 frequency spectrum (SFS) and found that the folded SFS of TE polymorphisms was shifted toward
187 a higher proportion of rare minor alleles compared to neutral sites in all genetic clades (Fig. 1C-
188 G). Splitting the TE data into DNA-transposons and retrotransposons resulted in similar folded
189 SFS and shifts in both TE classes (Additional file 1: Fig. S5 and S6).

190 These shifts could be the result of purifying selection as the analyses presented above
191 indicate that positive selection has a negligible effect on TE polymorphism frequencies in
192 *B. distachyon*. However, in contrast to SNPs, TEs do not evolve in a clock-like manner, as their
193 transposition rate is known to vary between generations [46, 47]. Changes in transposition rate
194 and purifying selection can lead to similar shifts in the SFS but can be disentangled using age-
195 adjusted SFS [48]. In brief, if TE polymorphisms are evolving neutrally, they are expected to

196 accumulate on average at the same rate in a population as neutral SNPs of the same age. Hence,
197 Δ frequency, the difference between the average frequency of TE polymorphisms and neutral
198 sites in a specific age bin, will remain close to 0 regardless of the polymorphisms' age. In contrast,
199 if TE polymorphisms evolve under purifying selection, they will tend to occur at lower frequencies
200 than neutral SNPs of the same age, as selection will prevent them from accumulating in the
201 population. Consequently, the Δ frequency will reach negative values for older TE polymorphisms
202 [48].

203 Because this model does not allow for back mutations, as typically observed for DNA-
204 transposons that can excise from the genome, we primarily investigated the age-adjusted SFS of
205 retrotransposons in the four derived clades. This analysis revealed that retrotransposons are
206 indeed prevented by natural selection from randomly accumulating, as older retrotransposons
207 are significantly less frequent than neutral SNPs of the same age (Fig. 2; one-sided Wilcoxon test,
208 Bonferroni corrected p value < 0.01).

209 As previous studies showed that the distance between TE polymorphisms and the next
210 gene can impact the strength of selection affecting TEs [6, 49, 50], we further split our
211 retrotransposon polymorphisms into three categories based on their distance to the next gene:
212 retrotransposons (i) in and up to 1 kb away from genes, (ii) between 1 kb and 5 kb away and (iii)
213 more than 5 kb away. The age-adjusted SFS of all three categories displayed the same pattern as
214 that observed for the whole retrotransposon polymorphism dataset: older retrotransposon
215 polymorphisms were significantly less frequent than neutral sites of the same age regardless of
216 their distance to genes (one-sided Wilcoxon test, Bonferroni corrected p value < 0.01), indicating

217 that retrotransposons more than 5 kb away from genes are also affected by purifying selection

218 (Fig. 2).

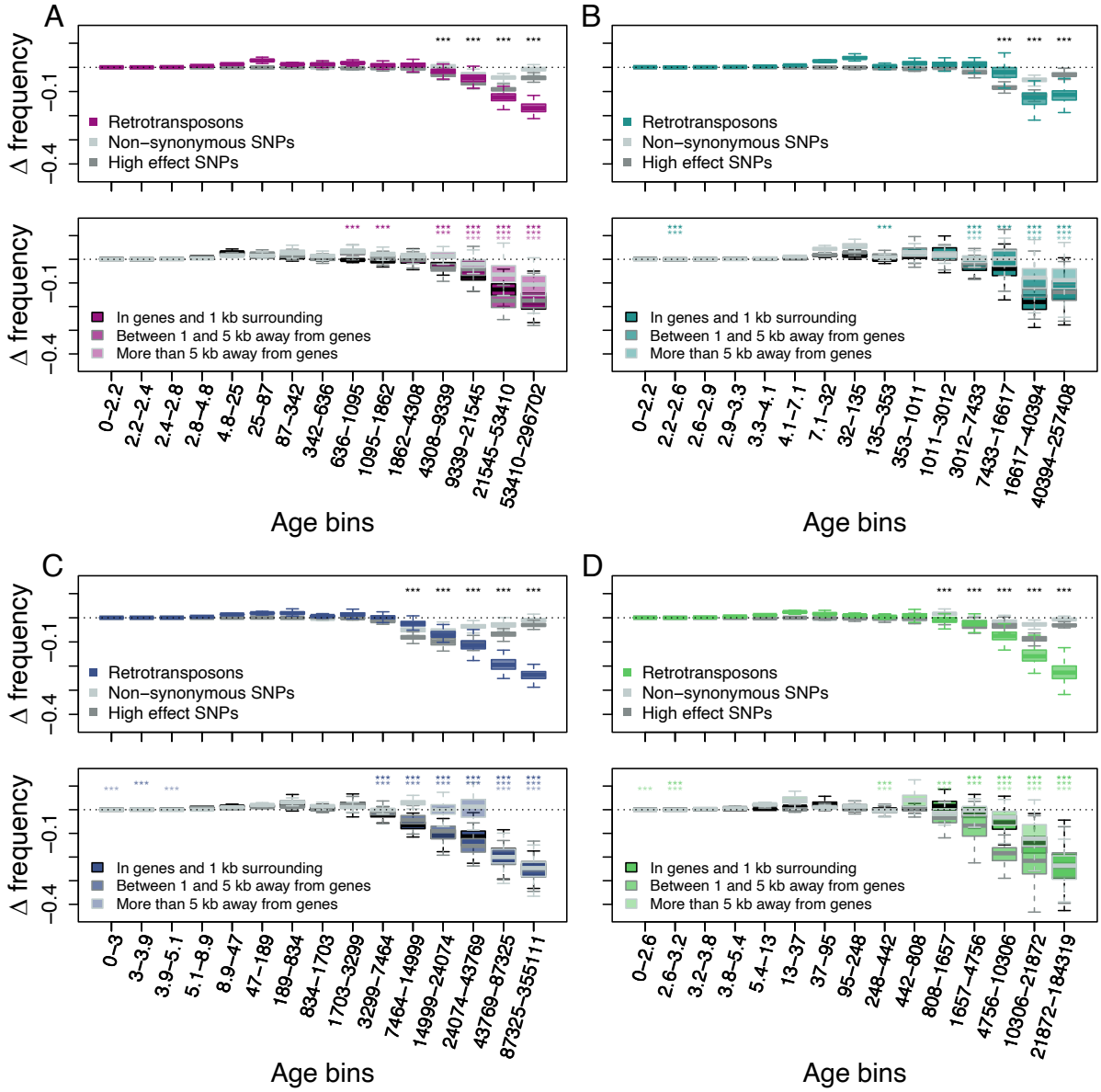

Fig. 2 Age-adjusted SFS of retrotransposons. The top row shows the age-adjusted SFS of all retrotransposons (colored), non-synonymous SNPs (light gray) and high effect SNPs (dark gray) in the four derived clades. The bottom row shows the age-adjusted SFS of retrotransposons based on their distance to the next gene in the four derived clades. The X axes show the age range of the mutations in each bin, and the age range of each bin was chosen so that each bin represents the same number of retrotransposon observations in the top row. The different columns show the four derived clades: A: A_East; B: A_Italia; C: B_West; D: B_East. Boxplots are based on 100 estimations of Δ frequency. Significant deviations of Δ frequency estimates from 0 in the age-adjusted SFS of retrotransposons are shown with asterisks (one-side Wilcoxon tests, Bonferroni corrected *p* value < 0.01: ***).

219 Retrotransposon polymorphisms tended to be more deleterious than SNPs predicted to
220 have a high impact on fitness. Indeed, the age-adjusted SFS of retrotransposons resulted in a
221 larger deviation of Δ frequency from 0 than for non-synonymous SNPs and high effect SNPs (Fig.
222 2). In addition, Δ frequency in the oldest (last) age bin was significantly smaller than in all other
223 age bins in the A_East, B_East and B_west clades (one-sided Wilcoxon test, Bonferroni corrected
224 *p* value < 0.01). In the A_Italia clades the oldest age bin was not significantly different form the
225 second oldest age bin (two-sided Wilcoxon test, Bonferroni corrected *p* value N.S.). While older
226 non-synonymous SNPs and high effect SNPs were generally less frequent than neutrally evolving
227 SNPs at the same age, the negative Δ frequency trend was reversed for the oldest non-
228 synonymous SNPs and high effect SNPs (Fig. 2). In all clades, Δ frequency in the oldest age bin
229 was significantly higher than at least the lowest Δ frequency observed in the other age bins for
230 non-synonymous SNPs, as well as high effect SNPs (one-sided Wilcoxon test, Bonferroni
231 corrected *p* value < 0.01). This might be because not all predicted non-synonymous SNPs and
232 high effect SNPs might result in fitness effects. Those SNPs can therefore evolve neutrally or
233 nearly neutrally and persist as polymorphic SNPs much longer in a population than those
234 affecting fitness negatively. Hence, the last age bin of the non-synonymous and high effect SNP
235 age-adjusted SFS likely harbors mainly neutrally and nearly neutrally evolving mutations, and
236 consequently, Δ frequency in not the smallest in the last age bin in these age-adjusted SFS.

237 To assess whether similar forces may drive retrotransposon and DNA-transposon
238 evolution, we repeated the analysis for DNA-transposons. The age-adjusted SFS of DNA-
239 transposons revealed very similar patterns, with Δ frequency showing significant deviations from
240 0 in older age bins (one-sided Wilcoxon test, Bonferroni corrected *p* value < 0.01) and DNA-

241 transposon polymorphisms being more deleterious than non-synonymous SNPs and high effect
242 SNPs (Additional file 1: Fig. S9).

243

244 **Forward simulations allow us to quantify the strength of purifying selection**

245 To evaluate to what extent the proportion of neutrally evolving mutations in the focal group of
246 mutations affects the shape of the age-adjusted SFS, we ran forward simulation with mutations
247 under multiple selective constraints, and we tested what ratio of neutral to selected mutations
248 can lead to an age-adjusted SFS similar to that observed for retrotransposons in *B. distachyon*.
249 Specifically, we investigated the conditions under which we observed a Δ frequency in the oldest
250 age bin significantly smaller than Δ frequency in all other age bins. Our simulations revealed that
251 the shape of the age-adjusted SFS of retrotransposons could only be reproduced if less than 10%
252 of the mutations were neutrally evolving for most of the selective constraint investigated
253 (Additional file 1: Fig. S7 and Additional file 2: Table S2).

254 Finally, we used the results from our simulations to narrow down the selection strength
255 affecting retrotransposons in *B. distachyon* by investigating the age of the oldest
256 retrotransposons in our dataset. The main difference between the age-adjusted SFS of mutations
257 evolving under weak and strong purifying selection is that the oldest mutations are much older
258 in the simulation with weak purifying selection than in the simulation with strong purifying
259 selection. This age difference arises because mutations under strong purifying selection are
260 removed from the population more effectively and, therefore, cannot persist as long in the
261 population. Examining the age of the last retrotransposon bins in the age-adjusted SFS revealed
262 that the ages of the oldest retrotransposons were the most similar to the expected ages of the

oldest mutations in our simulations, with a scaled selection coefficient (S) of -5 and -8 (Fig. 3), indicating that retrotransposons in *B. distachyon* are under moderate purifying selection. In simulations with a nearly neutral selection coefficient (S = -1), the simulated mutations were much older than the oldest observed retrotransposons (Fig. 3). Conversely, in simulations with a strong purifying selection coefficient (S < -10), they were much younger than the oldest observed retrotransposons (Fig. 3).

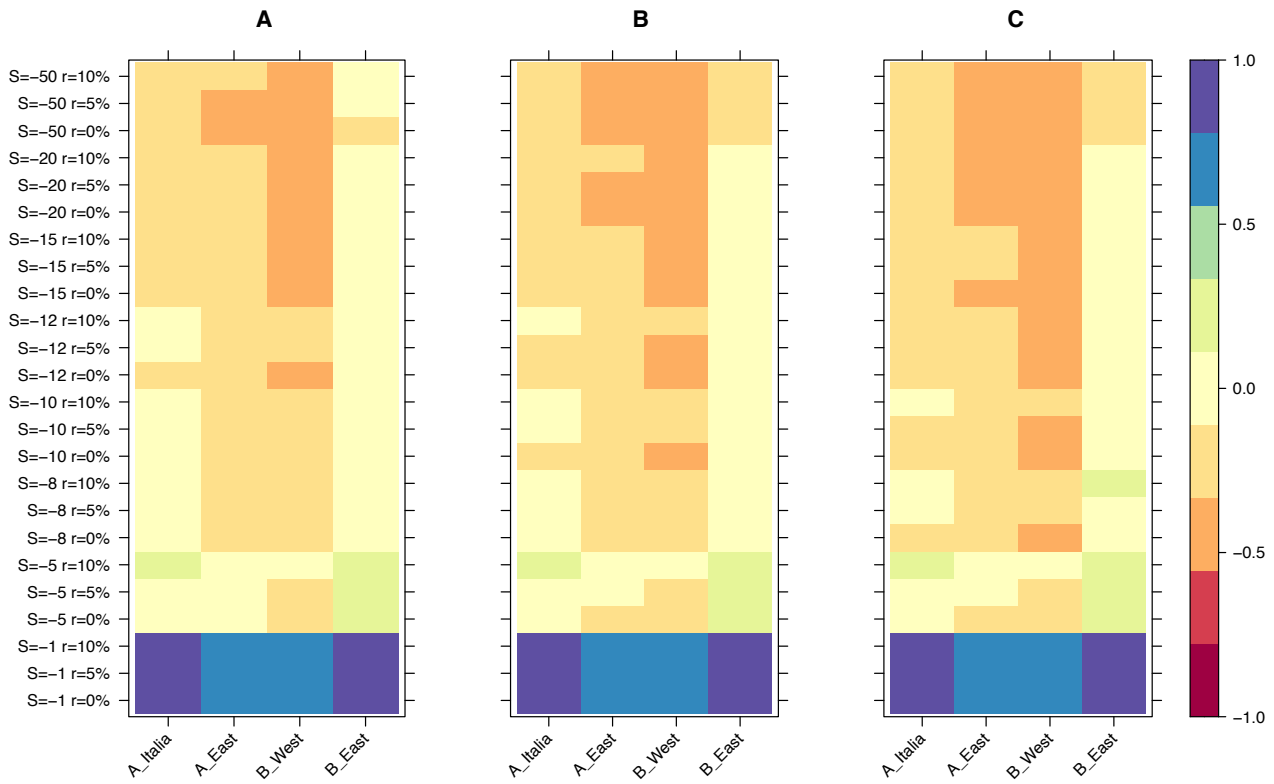

Fig. 3 Relative age difference ((mutation age in simulations - observed mutation age)/maximum absolute age difference) between simulated and observed data in the last bin of the age-adjusted SFS. A: 25% quantile; B: 50% quantile; C: 75% quantile.

272 # Discussion

273 *B. distachyon* is a widely used model species in evolutionary genomics, molecular ecology,
274 developmental biology and crop functional genomics [for review 51, 52] with past and ongoing
275 TE movements in its genome [32]. In this study, we used a diversity panel containing next-
276 generation sequencing data from over 320 individuals sampled across the whole geographical
277 range of *B. distachyon* to examine the role of TEs during evolution and adaptation. We
278 investigated the frequency with which positive selection led to an increase in the frequency and
279 fixation of TEs and quantified the strength of purifying selection on TE polymorphisms.
280 Accounting for population structure and fluctuant transposition rates, we demonstrate that TEs
281 are rarely part of the genetic makeup that was positively selected during environmental
282 adaptation in *B. distachyon*. Furthermore, we show that the majority of TE polymorphisms found
283 in the natural population of this model species are under weak to moderate purifying selection,
284 with only a small minority of TE polymorphisms evolving neutrally.

285

286 **Rare instances of positive selection on TEs**

287 By combining complementary approaches, we were able to demonstrate that TEs are rarely the
288 target of positive selection in *B. distachyon*. We first probed for footprints of positive selection
289 on TE polymorphisms using the five genetic clades as focal populations. In conducting this
290 analysis, we did not find TE polymorphisms to be at high frequencies or fixed at higher rates than
291 expected, in regions of the genome presumably harboring selective sweeps in at least one of the
292 genetic clades (high iHS regions). This suggested that TEs were rarely the target of positive
293 selection, which we confirmed with a genome-wide scan for overly differentiated TE

294 polymorphisms using X^tX analysis. Indeed, this approach revealed that only a very small
295 proportion of TE polymorphisms are more differentiated than expected under a neutral scenario.

296 Importantly, the X^tX analysis also revealed that a non-negligible fraction of the TE
297 polymorphisms is less differentiated than expected and are shared among genetic clusters. This
298 could be the result of selection favoring the same TE polymorphisms in different accessions to
299 adapt to similar environmental constraints across genetic clades. To test this scenario, we
300 performed GEA with 32 environmental factors, and found only nine TE polymorphisms
301 significantly associated with any of these, and representing a very small proportion (< 0.01%) of
302 all the TE polymorphisms we identified. Interestingly though, these nine TE polymorphisms were
303 associated with environmental variables pertaining to precipitation, temperature and altitude,
304 which are known to drive adaptation in *B. distachyon* [39]. Some insertions were found within or
305 in close proximity of genes, making these polymorphisms very good candidates for future
306 functional validation.

307 Single TE insertions can have a drastic impact on phenotypic variation and be affected by
308 positive selection [for review, see 16, 53, 54]. For instance, TEs have increased in frequency
309 through positive selection in humans [23] or during range expansion in *Arabidopsis* [25] and
310 *D. melanogaster* [20, 22]. Evidently, *B. distachyon* exhibits a different pattern, as causal
311 mutations for adaptation in this grass species are rarely TEs. Only a few studies have thoroughly
312 quantified the extent to which positive selection influences the evolution of TEs [25, 26, 28, 30,
313 31]. But two of these drew similar conclusions to us, in the green anole *Anolis carolinensis* [30]
314 and in the invasive species *Drosophila Suzukii* [31]. In addition, a large number of candidate genes
315 for adaptation were identified with a similar approach focusing on SNPs [39], indicating that

316 population structure or demographic events are not limiting factors for the methods we used.
317 Altogether, these observations call for a closer investigation of which forces, e.g., purifying
318 selection or neutral evolution, are important in shaping TE allele frequency in natural
319 populations.

320

321 **Moderate purifying selection is the dominant force during TE evolution**

322 Our results suggest that purifying selection is an important factor limiting the ability of TE
323 polymorphisms to fix and increase their frequency in *B. distachyon*. Indeed, one of the significant
324 explanatory variables in our ANCOVA models was the genetic clade, a proxy for the effective
325 population size (N_e), which affects the efficiency with which selection can fix beneficial mutations
326 and purge deleterious ones. In *B. distachyon*, the number of fixed TE polymorphisms per clade
327 and the frequency of TE polymorphisms were negatively correlated with N_e, indicating that the
328 accumulation of TEs is significantly lower in genetic clades with a larger N_e, potentially because
329 of a greater efficacy of purifying selection.

330 It is widely accepted that most new TE insertions have a deleterious or no effect on the
331 fitness of the host [22, 26, 28, 30, 33-36, 55]. To properly quantify the effect of purifying selection
332 on TE evolution in *B. distachyon*, we used age-adjusted SFS analyses to evaluate the selective
333 constraint experienced by TE polymorphisms while accounting for previously reported changes
334 in their activity [32]. While this method can only be applied to retrotransposons (because the
335 model does not allow back mutations), it provided a first clue on the importance of purifying
336 selection on TE evolution and revealed that overall, retrotransposons evolved under purifying
337 selection in all four derived genetic clades. Indeed, the Δ frequency was significantly smaller than

338 0, especially for older retrotransposons, meaning that old retrotransposons are less common
339 than neutrally evolving SNPs at the same age. This further demonstrates that even after
340 accounting for the different genetic clades and using a large sample size, retrotransposons evolve
341 under purifying selection in *B. distachyon*.

342 We also revealed that only a minority of retrotransposons evolved neutrally, as the
343 observed shape of the Δ frequency curve could only be reproduced in our simulation if the
344 proportion of neutrally evolving mutations in our focal mutations was below 10%. This estimate
345 gives a first glimpse into the distribution of fitness effects of new TE insertions, a fundamental
346 parameter in genetics that describes the way in which new TE insertions can contribute to
347 evolution and adaptation [56]. Here, we show for the first time that new TE insertions have a less
348 than 10% chance to insert into the genome of *B. distachyon* in a way that will allow them to
349 evolve neutrally, advocating for a large potential of TEs to create, through their movement, new
350 phenotypic variation on which selection can act on. PCAs based on TE polymorphisms allowed us
351 to recover the population structure of *B. distachyon*, implying that demographic history and
352 hence neutral processes may indeed partially explain the differences in the TE distribution we
353 observed between genetic clades, as shown in *Arabidopsis thaliana* and *Arabidopsis lyrata* [26,
354 57], *Drosophila melanogaster* [58], humans [59] and the green anole (*Anolis carolinensis*; [30]).
355 However, and in line with our simulations, the first two axes of the PCA explain less than 7% of
356 the variance, indicating that neutrally evolving TEs contribute only mildly to overall TE diversity
357 in our system.

358 Because TEs can cause phenotypic variation through new insertions [1-8], it is not
359 surprising that most new insertions interfere with the function of the genome, especially in a

360 species with a small genome, such as *B. distachyon* (272 Mb) [37]. The proportion of neutrally
361 evolving TE polymorphisms is expected to be very small in genes, as insertions in genic regions
362 are likely to result in loss-of-function [1, 2]. Similarly, TE insertions in close proximity to genes are
363 expected to be highly disruptive, as regulatory elements such as *cis*-regulatory elements are
364 predominantly located in the proximity of genes. In *A. thaliana*, for instance, TEs located in the
365 vicinity of genes (less than 2 kb) globally result in downregulation [60]. Although only specific
366 families alter gene expression in *B. distachyon* [41], the observed Δ frequency for
367 retrotransposon polymorphisms in genes and in their 1 kb surroundings matched our
368 expectations. The fact that TE polymorphisms located more than 5 kb away from genes are also
369 evolving under purifying selection was more surprising. That said, little is known about the
370 distance between *cis*-regulatory sequences and genes in *B. distachyon.* In plants, TEs are believed
371 to affect gene expression in *trans* through the production of small-interfering RNA [61-65].
372 Hence, the fact that only a small proportion of TEs can accumulate neutrally indicates that, in a
373 gene-dense genome such as that of *B. distachyon* (42.5% of the genome are genes) [65], TE
374 insertions in any genomic compartment may result in some *cis*- or *trans*-regulatory effects visible
375 to selection.

376 To further ascertain the strength of purifying selection, we used forward simulation and
377 showed that simulations assuming a moderately weak selection pressure (S = -5 or S = -8) against
378 TE polymorphisms best fitted our observed data. In theory, no TE polymorphisms under strong
379 purifying selection should be present in a natural population, as such mutations are expected to
380 be quickly lost, especially in a predominantly selfing species where most loci are expected to be
381 homozygous. Therefore, it is not surprising that TE polymorphisms in *B. distachyon* are under

382 weak to moderate selection, as also shown, for example, for the L1 retrotransposons in humans
383 [27] or the BS retrotransposon family in *Drosophila melanogaster* [58].

384 While some of the parameters we chose for our simulations, such as the dominance or
385 selfing rate, can affect the efficiency of TE purging, it is unlikely that discrepancies in the true and
386 assumed values for these parameters would have led to drastically different results. For example,
387 we assumed codominance for all mutations, which might not hold true for each TE
388 polymorphism. However, because of the high selfing rate observed in *B. distachyon* [38],
389 heterozygous loci are expected to be rare, and dominance is unlikely to have a strong impact on
390 our observations. Similarly, with a higher selfing rate, deleterious TE polymorphisms should be
391 removed more efficiently by purifying selection. To check whether a lower selfing rate could
392 allow a higher proportion of TE polymorphisms to evolve neutrally, we reran the simulations
393 assuming fully outcrossing individuals. This also resulted in simulation with weak to moderate
394 selection strength on TE polymorphisms best fitting the observed data, further strengthening our
395 results.

396 While the analyses of positive selection and GEA were based on both DNA-transposons
397 and retrotransposons, we only used retrotransposons to assess the strength of selection on TE
398 polymorphisms, as the age-adjusted SFS was developed with the assumption of no back
399 mutations [48]. Yet, DNA-transposons do not solely transpose through cut and paste mechanisms
400 as they would otherwise not be so abundant in Eukaryotic genomes. DNA-transposons can also
401 create extra copies of themselves by transposing during chromosome replication or repair from
402 a position that has already been replicated, or repaired [66]. We therefore repeated the age-
403 adjusted SFS analyses using DNA-transposons to evaluate whether DNA-transposons were

404 affected by similar selective constraints. The folded SFS of DNA-transposons and
405 retrotransposons display similar shifts toward high proportions of rare alleles and Δ frequency
406 deviations from 0 in the age-adjusted SFS of DNA-transposons and retrotransposons are
407 comparable. Hence, we argue that the conclusion drawn for retrotransposons also holds for DNA-
408 transposons, and that purifying selection affect TEs broadly.

409

410 **Conclusion**

411 Adaptation to different environmental conditions is a complex process that involves various
412 mutation types. Here, we show that the vast majority of TE polymorphisms are under purifying
413 selection in the small genome of *B. distachyon*. Conversely, only a very small proportion of TEs
414 seem to have contributed to adaptation. The observed lack of neutrally evolving TE
415 polymorphisms in *B. distachyon* advocates for a large potential of TE polymorphisms to
416 contribute to the genetic diversity and phenotypic variation on which selection can act and
417 highlights the need to consider TE polymorphisms during evolutionary studies. Finally, our work
418 shows that the ability of TEs to cause phenotypic variation does not necessarily translate into
419 being favored during evolution and adaptation over other mutations with more subtle effects,
420 such as SNPs.

421

422

423

424

425 # Materials and methods

426 **Whole-genome resequencing data**

427 In this study, we analyzed a total of 326 publicly available whole-genome sequencing data from
428 *Brachypodium distachyon* accessions sampled around the Mediterranean Basin (Fig. 1A;
429 Additional file 2: Table S3). Our *B. distachyon* dataset consisted of 47 samples published by
430 Gordon et al. [8], 57 samples published by Skalska et al. [67], 65 samples published by Gordon et
431 al. [68], 86 samples published by Stritt et al. [38] and 71 samples published by Minadakis et al.
432 [39], covering all five genetic clades previously described in this species [38, 39]. Each sample was
433 assigned to a genetic clade based on previously published results [39].

434

435 **Data processing**

436 Raw reads were trimmed using Trimmomatic 0.36 [69] and mapped to the *B. distachyon*
437 reference genome version 3.0 [37] using bowtie2 [70] and yaha [71], and TE polymorphisms were
438 identified using the TEPID pipeline [72] and the recently updated TE annotation by Stritt et al.
439 [73] and Wyler et al. [65]. TE polymorphisms include both TE insertion polymorphisms (TIPs;
440 insertions absent from the reference genome but present in at least one natural accession) and
441 TE absence polymorphisms (TAPs; insertions present in the reference genome but absent from
442 at least one natural accession). The class, superfamily and family of each TE call were assigned
443 based on the TEPID results and the TE annotation from the reference genome. TIPs that were
444 less than 100 base pairs (bp) apart in different samples and assigned to the same TE family were
445 merged.

446 Single nucleotide polymorphisms (SNPs) were called using GATK v.4.0.2.1 [74] using
447 HaplotypeCaller [75] following Minadakis et al. [39]. The SNP calls were hard filtered using the
448 following conditions: QD < 5.0; FS > 20.0; SOR > 3.0; MQ < 50.0; MQRankSum < 2.5; MQRankSum
449 > -2.5; ReadPosRankSum < 2.0; ReadPosRankSum > -2.0. Because *B. distachyon* displays a high
450 selfing rate [33], most genetic variants are expected to be homozygous within an individual.
451 Hence, all TE calls were treated as homozygous, and heterozygous SNP calls were removed from
452 our dataset to reduce false variant calls. Additionally, all sites with multiallelic TE and SNP calls
453 were removed. SNPs were classified as synonymous, non-synonymous and of high fitness effect
454 using SnpEff [76]. SNPs and TE polymorphisms were merged into a single vcf file using custom
455 scripts provided in github (see section Availability of data and materials).

456 To estimate the age of each SNP and TE polymorphism, the SNPs and TEs found in the
457 A_East, A_Italia, B_East and B_West clades were polarized using the C clade, which was identified
458 as the most ancestral *B. distachyon* clade [38] and used as the outgroup throughout this study.
459 An estimate for the time of origin of all SNPs and TE polymorphisms was calculated with GEVA, a
460 nonparametric approach that relies on pairwise differences in identity by descent (IBD) regions
461 around the focal mutation to estimate the time of origin [77]. GEVA was run separately for each
462 clade using the genetic map produced by Huo et al. [78] and a mutation rate of $7x10^{-9}$
463 substitutions/generation. The theoretical prediction of the correlation between allele age and
464 allele frequency of neutrally evolving mutations based on N_e [42] was compared to the observed
465 correlation between allele age and frequency of synonymous SNPs to check the sanity of the age
466 estimates.

467 The observed SNP and TE diversity was first examined using a principal component
468 analysis (PCA), and correlations between TE diversity and genetic clades were tested with a
469 mantel test using the ade4 package version 1.7-22 [79] in R version 4.1.2 [80]. The folded site
470 frequency spectrum (SFS) was computed for TE polymorphisms and SNPs using the minor allele
471 frequency in R version 4.1.2 [80]. Finally, the map of the geographical distribution of the used
472 accessions was done in R using the rnaturalearth package 0.3.3 [81].

473

474 **Analyses of positive selection**

475 Regions of the genome affected by positive selection were identified using the integrated
476 haplotype score (iHS), a measure of the amount of extended haplotype homozygosity along the
477 ancestral allele relative to the derived allele for a given polymorphic site [82]. iHS was calculated
478 using the SNP dataset, and regions displaying longer haplotypes and hence high iHS were
479 identified in R using the rehh package [83, 84]. The threshold to distinguish between regions of
480 high iHS and other regions was selected such that less than 5% of the *B. distachyon* genome was
481 classified as high iHS regions in each clade (Additional file 2: Table S4). Candidate regions under
482 positive selection were defined as all regions that were found to have high iHS in each clade
483 separately.

484 A first ANCOVA was used to model the number of fixed TE polymorphisms in each clade
485 found in the candidate region under positive selection based on the following genetic features:
486 total number of TEs, TE superfamily, TE age (split into three categories: young: age < 10,000
487 generations; intermediate: age between 10,000 generations and 60,000 generations; old: age >
488 60,000 generations), clade, genomic region (a unique ID for each candidate region under positive

489 selection) and iHS classification of the regions in each clade (high or average). A second ANCOVA
490 was used to model the allele frequency of TE polymorphisms found in the candidate region under
491 positive selection based on the following genetic features: TE superfamily, TE age, clade, genomic
492 region and iHS classification of the regions in each clade. The TE superfamily was included to
493 account for different evolutionary behaviors of TEs from different superfamilies. Age accounted
494 for differences in the fixation rate and frequency distribution between young and old TEs. The
495 clade was included to account for clade-specific differences such as differences in N_e. Finally, a
496 unique ID for each candidate region under positive selection was included to account for region-
497 specific differences such as differences in the recombination rate and GC content. In the end,
498 regions that were found to have a high iHS in some clades were compared to the same regions
499 in the other clades. All ANCOVAs were run in R using the car package [85].

500 The standardized allele frequency of a mutation across populations (X^tX) values [44] were
501 computed for the combined TE and SNP dataset using Baypass version 2.3 [45, 86]. The X^tX values
502 were used to identify over- and under differentiated TE polymorphisms between clades. A
503 pseudo-observed dataset (POD) of 100,000 SNPs was simulated under the demographic model
504 inferred from the covariance matrix of the SNP dataset. The POD was then used to determine the
505 97.5% (over-differentiated polymorphisms) and 2.5% (under differentiated polymorphisms)
506 quantiles.

507

508 **Genome-environment association analyses**

509 We identified TE polymorphisms significantly associated with environmental factors using
510 genome-environment association analyses (GEA) following Minadakis et al. [39]. GEAs were run

511 with GEMMA 0.98.5 [87] using the combined TE and SNP vcf file against the following 32
512 environmental factors extracted by Minadakis et al. [39]: altitude, aridity from March to June,
513 aridity from November to February, annual mean temperature, mean temperature of warmest
514 quarter, mean temperature of coldest quarter, annual precipitation, precipitation of wettest
515 month, precipitation of driest month, precipitation seasonality, precipitation of wettest quarter,
516 precipitation of driest quarter, precipitation of warmest quarter, precipitation of coldest quarter,
517 mean diurnal Range, isothermality, temperature seasonality, maximum temperature of warmest
518 month, minimum temperature of coldest month, temperature annual range, mean temperature
519 of wettest quarter, mean temperature of driest quarter, precipitation from March to June,
520 precipitation from November to February, solar radiation from March to June, solar radiation
521 from November to February, mean temperature between March and June, mean temperature
522 between November and February, maximum temperature between March and June, maximum
523 temperature between November and February, minimum temperature between March and June
524 and minimum temperature between November and February. We applied a False Discovery Rate
525 (FDR, [88]) threshold of 5% to control for false positive rates.

526

527 **Age-adjusted frequency spectra and analyses of purifying selection**

528 Footprints of purifying selection on TE polymorphisms were first evaluated using folded SFS. An
529 age-adjusted site frequency spectrum (age-adjusted SFS) approach was used to further
530 investigate the impact of purifying selection on retrotransposons while accounting for
531 nonconstant transposition rates. Briefly, the age-adjusted SFS is a summary statistic that
532 describes the difference between the average frequency of TEs at a specific age and the average

533 frequency of neutral sites of the same age [48]. Therefore, the TE dataset was sorted by age and
534 split into equally large bins with respect to the number of observations in each age bin. Neutral
535 sites were then randomly down-sampled to match the number of observations in the TE dataset
536 and its age distribution [48].

537 The difference between the average TE and neutral site frequency, or Δ frequency, was
538 computed for each age bin [48]. This method allows for an unbiased comparison between the
539 allele frequencies of TEs and neutral sites, and is robust to transposition rate changes and
540 demographic changes [48]. However, the theory behind this method was developed assuming
541 no back mutations and is therefore best suited for retrotransposons, as DNA-transposons can
542 exit an insertion site [48]. We used the synonymous SNPs identified with SnpEff as the neutrally
543 evolving sites. However, because estimating the population wide frequency of TEs is more
544 challenging than estimating SNP frequencies, putative biases in frequency estimates need to be
545 assessed before performing age-adjusted SFS analyses. To do so, the SNP dataset was resampled
546 so that the SNP dataset used in the age-adjusted SFS had a frequency distribution that matched
547 the observed TE frequency distribution. The age-adjusted SFS of retrotransposons was
548 contrasted against the age-adjusted SFS of non-synonymous, as well as against high fitness effect
549 SNPs. Therefore, 10,000 non-synonymous and high fitness effect SNPs were randomly selected
550 for each clade to reach approximately the same number of retrotransposon polymorphisms, non-
551 synonymous and high fitness effect SNPs for final comparisons. To estimate the variation in Δ
552 frequency estimates, all age-adjusted SFS were computed 100 times. All Wilcoxon tests and
553 Bonferroni *p* value corrections were done in R version 4.1.2 [80].

554

555 **Forward simulation**

556 We used SLiM 4.0.1 [89, 90] to run forward simulations and assess the proportion of neutrally
557 evolving retrotransposons and the average selection strength affecting them. The simulations
558 were designed to reflect the population size and demographic history of *B. distachyon*. The
559 simulated genomic fragment was 1 megabase (Mb) long and included neutral (synonymous)
560 mutations as well as focal mutations that evolved under different selective constraints. The focal
561 mutations were a mix of neutrally evolving mutations and mutations evolving under a constant
562 selection pressure. Therefore, the ratio (r) of focal mutations that evolved neutrally was either
563 0%, 5%, 10%, 25% or 50%. The scaled selection coefficient (S, defined as $N_e s$, with s the strength
564 of selection and N_e the effective population size) affecting the remaining focal mutations was set
565 at the beginning of the simulation to be either -1, -5, -8, -10, -12, -15, -20 or -50 to cover
566 effectively neutral ($0 > S \geq -1$), intermediate ($-1 > S \geq -10$) and strongly deleterious ($-10 > S$)
567 selective constraints. The selfing rate was set to 70%, as *B. distachyon* is a highly selfing species
568 with occasional outcrossing [38, 39]. In addition, a high recombination rate was chosen to
569 minimize the effects of linked selection in the small genomic fragment simulated. Simulations for
570 each combination of these two parameters were run 20 times to assess the variation in the
571 resulting age-adjusted SFS. The shape of the resulting age-adjusted SFS was used to narrow down
572 the ratio of neutrally evolving TE polymorphisms. Similarly, the age distribution of the mutations
573 in the oldest bin of the age-adjusted SFS was used to narrow down the strength of selection
574 affecting TE polymorphisms.

575

576

577 # Supplementary Information

578 **Additional file 1: Supplemental Figures. Figure S1.** Principal Component Analyses using TE (left
579 panel) and SNP (right panel) polymorphisms. **Figure S2.** Principal Component Analyses using
580 retrotransposon (left panel) and DNA-transposon (right panel) polymorphisms. **Figure S3.**
581 Observed correlation between age in generations and frequency of synonymous SNPs in the four
582 derived genetic clades. The red points show the expected age of a neutrally evolving mutation at
583 a specific frequency based on the predictions of Kimura and Ohta (1973). Panel A: clade A_East;
584 panel B: clade A_Italia; panel C: clade B_West and panel D: clade B_East. **Figure S4.** Distribution
585 of the observed TE age scaled by the effective population size (N_e) in the four derived genetic
586 clades of *B. distachyon*. **Figure S5.** Folded site frequency spectrum of DNA-transposons and
587 synonymous SNPs in all genetic clades. Panel A: A_East; panel B: A_Italia; panel C: B_West; panel
588 D: B_East; panel E: C. **Figure S6.** Folded site frequency spectrum of retrotransposons and
589 synonymous SNPs in all genetic clades. Panel A: A_East; panel B: A_Italia; panel C: B_West; panel
590 D: B_East; panel E: C. **Figure S7.** Age-adjusted SFS of simulated mutations under negative
591 selection in the four derived clades. The four columns show the results for the A_East, A_Italia,
592 B_West and B_East genetic clades, respectively. Each line shows the results for the different
593 scaled selection coefficients (S). The five colored curves in each plot show the shape of the age-
594 adjusted SFS with varying ratios of neutrally evolving mutations, and the gray curves show
595 variation within one standard deviation based on the 20 runs for each simulation. The X axes
596 show the age bin from the youngest to the oldest, with each age bin including the same number
597 of observations for each simulation. **Figure S8.** Relative age difference ((mutation age in
598 simulations - observed mutation age)/maximum absolute age difference) between simulated

599 data assuming fully outcrossing individuals and observed data in the last bin of the age-adjusted
600 SFS. A: 25% quantile; B: 50% quantile; C: 75% quantile. **Figure S9.** Age-adjusted SFS of DNA-
601 transposons (colored), non-synonymous SNPs (light gray) and high effect SNPs (dark gray) in the
602 four derived clades. The X axes show the age range of the mutations in each bin, and the age
603 range of each bin was chosen so that each bin represents the same number of DNA-transposons
604 observations. A: A_East genetic clades; B: A_Italia genetic clades; C: B_West genetic clades; D:
605 B_East genetic clades. Boxplots are based on 100 estimations of Δ frequency.

606 **Additional file 2: Supplemental Tables. Table S1.** List of TEs significantly associated with at least
607 one environmental factor in the GWAS. **Table S2.** Difference in Δ frequency between the oldest
608 and second oldest age bin in the different simulations **Table S3.** List of published samples used in
609 this study. **Table S4.** List of thresholds used and percentage of the genome classified as high iHS
610 regions in the four derived clades.

611

612 Acknowledgments

613 We would like to thank Jeffrey Ross-Ibarra and Mitra Menon as well as Fabrizio Menardo,
614 Michael Thieme, Wenbo Xu, Jigisha, Lars Kaderli and Serafin Schefer for all the discussions and
615 their comments on this project. We thank Emmanuelle Botté for professional editing.

616

617 Declarations

618 **Ethics approval and consent to participate**

619 Not applicable.

620

621 **Consent for publication**

622 All authors have read and approved the submission of this manuscript.

623

624 **Availability of data and materials**

625 The datasets supporting the conclusions of this article are publicly available on the European

626 Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home) and National Center for

627 Biotechnology Information (https://www.ncbi.nlm.nih.gov/sra/), and the archive numbers of

628 the accessions used are listed in the Additional file 1: Table S2. The scripts generated are

629 available on GitHub https://github.com/Roberthorv/TE_in_Brachypodium/tree/main.

630

631 **Competing interests**

632 The authors declare no competing interests.

633

634 **Funding**

635 This project was funded by the Swiss National Science Foundation (project project

636 31003A_182785) and the Research Priority Program Evolution in Action from the University of

637 Zürich. Data analyzed in this paper were generated in collaboration with the Genetic Diversity

638 Center (GDC), ETH Zürich.

639

640

641 **Authors' contributions**

642 A.R. and R.H. conceived the study. R.H. carried out the study, did the TE and SNP calling,
643 performed the age-adjusted SFS analyses, conducted the X^tX analyses and ran the forward
644 simulations. N.M. performed the GEA. Y.B. ran the iHS analyses. A.R. acquired the fundings. R.H.
645 wrote the manuscript and A.R. revised it. All authors discussed the results and commented on
646 the manuscript.

647

648 # References

649 1. Bhattacharyya MK, Smith AM, Ellis TH, Hedley C, Martin C. The wrinkled-seed character
650 of pea described by Mendel is caused by a transposon-like insertion in a gene encoding
651 starch-branching enzyme. Cell. 1990;60(1):115-22.

652 2. Hof AE, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, et al. The industrial melanism
653 mutation in British peppered moths is a transposable element. Nature. 2016;534:102-
654 105.

655 3. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat. Rev.
656 Genet. 2008;9:397-405.

657 4. Qiu Y, Köhler C. Mobility connects: transposable elements wire new transcriptional
658 networks by transferring transcription factor binding motifs. Biochem. Soc. Trans.
659 2020;48(3):1005-1017.

660 5. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the
661 genome. Nat. Rev. Genet. 2007;8(4):272-85.

662 6. Hollister JD, Gaut BS. Epigenetic silencing of transposable elements: a trade-off between
663 reduced transposition and deleterious effects on neighboring gene expression. Genome
664 Res. 2009;19(8):1419-1428.
665 7. Xiao H, Jiang N, Schaffner E, Stockinger EJ, van der Knaap E. A retrotransposon-mediated
666 gene duplication underlies morphological variation of tomato fruit. Science.
667 2008;319(5869):1527-1530.
668 8. Gordon SP, Contreras-Moreira B, Woods DP, Des Marais DL, Burgess D, Shu S, et al.
669 Extensive gene content variation in the Brachypodium distachyon pan-genome correlates
670 with population structure. Nat. Commun. 2017;8:2184.
671 9. Bennetzen JL, Kellogg EA. Do plants have a one-way ticket to genomic obesity? Plant Cell.
672 1997;9(9):1509-1514.
673 10. Vitte C, Panaud O. Formation of solo-LTRs through unequal homologous recombination
674 counterbalances amplifications of LTR retrotransposons in rice Oryza sativa L. Mol. Biol.
675 Evol. 2003;20(4):528–540.
676 11. Piégu B, Guyot R, Picault N, Roulin A, Sanyal A, Kim H, et al. Doubling genome size without
677 polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza
678 australiensis, a wild relative of rice. Genome Res. 2006;16(10):1262-1269.
679 12. Wendel JF, Jackson SA, Meyers BC, Wing RA. Evolution of plant genome architecture.
680 Genome Biol. 2016;17(37):1-14.
681 13. Lisch D. How important are transposons for plant evolution? Nat. Rev. Genet. 2013;14:49-
682 61.

683 14. Negi P, Rai AN, Suprasanna P. Moving through the stressed genome: emerging regulatory
684 roles for transposons in plant stress response. Front. Plant Sci. 2016;7(1448):1-20.
685 15. Rey O, Danchin E, Mirouze M, Loot C, Blanchet S. Adaptation to Global Change: A
686 Transposable Element-Epigenetics Perspective. Trends Ecol. Evol. 2016;31(7):514-526.
687 16. Dubin MJ, Mittelsten Scheid O, Becker C. Transposons: a blessing curse. Curr. Opin. Plant
688 Biol. 2018;42:23-29.
689 17. Quadrana L, Etcheverry M, Gilly A, Caillieux E, Madoui MA, Guy J, et al. Transposition
690 favors the generation of large effect mutations that may facilitate rapid adaption. Nat.
691 Commun. 2019;10(3421):1-10.
692 18. Uzunović J, Josephs EB, Stinchcombe JR, Wright SI. Transposable elements are important
693 contributors to standing variation in gene expression in Capsella grandiflora. Mol. Biol.
694 Evol. 2019;36(8):1734-1745.
695 19. Castanera R, Morales-Díaz N, Gupta S, Purugganan M, Casacuberta JM. Transposons are
696 a major contributor to gene expression variability under selection in rice populationse.
697 Life. 2023;12:RP86324.
698 20. González J, Karasov TL, Messer PW, Petrov DA. Genome-wide patterns of adaptation to
699 temperate environments associated with transposable elements in Drosophila. PLoS
700 Genet. 2010;6(4):e1000905.
701 21. Studer A, Zhao Q, Ross-Ibarra J, Doebley J. Identification of a functional transposon
702 insertion in the maize domestication gene tb1. Nat. Genet. 2011;43(11):1160-1163.
703 22. Barrón MG, Fiston-Lavier AS, Petrov DA, González J. Population genomics of transposable
704 elements in Drosophila. Annu. Rev. Genet. 2014;48:561-581.

705 23. Rishishwar L, Wang L, Wang J, Yi SV, Lachance J, Jordan IK. Evidence for positive selection
706 on recent human transposable element insertions. Gene. 2018;675:69-79.
707 24. Niu XM, Xu YG, Li ZW, Bian YT, Hou XH, Chen JF, et al. Transposable elements drive rapid
708 phenotypic variation in Capsella rubella. Proc. Nati. Acad. Sci. USA. 2019;116(14):6908-
709 6913.
710 25. Jiang J, Xu YC, Zhang ZQ, Chen JF, Niu XM, Hou XH, et al. Forces driving transposable
711 element load variation during Arabidopsis range expansion. bioRxiv. 2022;1-12.
712 https://www.biorxiv.org/content/10.1101/2022.12.28.522087v1
713 26. Lockton S, Ross-Ibarra J, Gaut BS. Demography and weak selection drive patterns of
714 transposable element diversity in natural populations of Arabidopsis lyrata. Proc. Natl.
715 Acad. Sci. U S A. 2008;105(37):13965–13970.
716 27. Boissinot S, Davis J, Entezam A, Petrov D, Furano AV. Fitness cost of LINE-1 (L1) activity in
717 humans. Proc. Natl. Acad. Sci. U S A. 2006;103(25):9590-9594.
718 28. Blumenstiel JP, Chen X, He M, Bergman CM. An age-of-allele test of neutrality for
719 transposable element insertions. Genetics. 2014;196(2):523-38.
720 29. Rech GE, Bogaerts-Márquez M, Barrón MG, Merenciano M, Villanueva-Cañas JL, Horváth
721 V, et al. Stress response, behavior, and development are shaped by transposable element-
722 induced mutations in Drosophila. PLoS Genet. 2019;15(2):e1007900.
723 30. Bourgeois Y, Ruggiero RP, Hariyani I, Boissinot S. Disentangling the determinants of
724 transposable elements dynamics in vertebrate genomes using empirical evidences and
725 simulations. PLoS Genet. 2020;16(10):e1009082.

726 31. Mérel V, Gibert P, Buch I, Rodriguez Rada V, Estoup A, Gautier M, et al. The Worldwide
727 Invasion of Drosophila suzukii Is Accompanied by a Large Increase of Transposable
728 Element Load and a Small Number of Putatively Adaptive Insertions. Mol. Biol. Evol.
729 2021;38(10):4252-4267.
730 32. Stritt C, Gordon SP, Wicker T, Vogel JP, Roulin AC. Recent activity in expanding populations
731 and purifying selection have shaped transposable element landscapes across natural
732 accessions of the mediterranean grass Brachypodium distachyon. Genome Biol. Evol.
733 2018;10(1):304-318
734 33. Charlesworth B. Transposable elements in natural populations with a mixture of selected
735 and neutral insertion sites. Genet. Res. 1991;57(2):127-134.
736 34. Charlesworth B, Langley CH, Sniegowski PD. Transposable element distributions in
737 Drosophila. Genetics. 1997;147(4):1993-1995.
738 35. Charlesworth B, Charlesworth D. The population dynamics of transposable elements.
739 Genet. Res. 1983;42(1):1-27.
740 36. Charlesworth B. Background selection and patterns of genetic diversity in Drosophila
741 melanogaster. Genet. Res. 1996;68(2):131-149.
742 37. International Brachypodium Initiative. Genome sequencing and analysis of the model
743 grass Brachypodium distachyon. Nature. 2010;463:763-768.
744 38. Stritt C, Gimmi EL, Wyler M, Bakali AH, Skalska A, Hasterok R, et al. Migration without
745 interbreeding: Evolutionary history of a highly selfing Mediterranean grass inferred from
746 whole genomes. Mol. Ecol. 2022;31(1):70-85.

747 39. Minadakis N, Williams H, Horvath R, Cakovic D, Stritt C, Thieme M, et al. The demographic
748 history of the wild crop relative Brachypodium distachyon is shaped by distinct past and
749 present ecological niches. bioRxiv. 2023. https://doi.org/10.1101/2023.06.01.543285
750 40. Bourgeois Y, Stritt C, Walser JC, Gordon SP, Vogel JP, Roulin AC. Genome-wide scans of
751 selection highlight the impact of biotic and abiotic constraints in natural populations of
752 the model grass Brachypodium distachyon. Plant J. 2018;96(2):438-451.
753 41. Wyler M, Stritt C, Walser JC, Baroux C, Roulin AC. Impact of transposable elements on
754 methylation and gene expression across natural accessions of Brachypodium distachyon.
755 Genome Biol. Evol. 2020;12(11):1994-2001.
756 42. Kimura M, Ohta T. The age of a neutral mutant persisting in a finite population. Genetics.
757 1973;75(1):199–212.
758 43. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the
759 human genome. PLoS Biol. 2006;4:e72.
760 44. Günther T, Coop G. Robust identification of local adaptation from allele frequencies.
761 Genetics. 2013;195(1):205-20.
762 45. Olazcuaga L, Loiseau A, Parrinello H, Paris M, Fraimout A, Guedot C, et al. A whole-genome
763 scan for association with invasion success in the fruit fly Drosophila suzukii using contrasts
764 of allele frequencies corrected for population structure. Mol. Biol. Evol. 2020;37(8):2369-
765 2385.
766 46. Garcia Guerreiro MP. What makes transposable elements move in the Drosophila
767 genome. Heredity. 2012;108(5):461–468.

768 47. Belyayev A. Bursts of transposable elements as an evolutionary driving force. J. Evol. Biol.
769 2014;27(12):2573–2584.
770 48. Horvath R, Menon M, Stitzer M, Ross-Ibarra J. Controlling for variable transposition rate
771 with an age-adjusted site frequency spectrum. Genome Biol. Evol. 2022;14(2):evac016.
772 49. Wright SI, Agrawal N, Bureau TE. Effects of recombination rate and gene density on
773 transposable element distributions in Arabidopsis thaliana. Genome Res.
774 2003;13(8):1897-1903.
775 50. Horvath R, Slotte T. The Role of small RNA-based epigenetic silencing for purifying
776 selection on transposable elements in Capsella grandiflora. Genome Biol. and Evol.
777 2017;9(10):2911-2920.
778 51. Raissig MT, Woods DP. The wild grass Brachypodium distachyon as a developmental
779 model system. Curr. Top. Dev. Biol. 2022;147:33-71.
780 52. Hasterok R, Catalan P, Hazen SP, Roulin AC, Vogel JP, Wang K, et al. Brachypodium: 20
781 years as a grass biology model system; the way forward? Trends Plant Sci.
782 2022;27(10):1002-1016.
783 53. Casacuberta E, González J. The impact of transposable elements in environmental
784 adaptation. Mol. Ecol. 2013;22(6):1503-1517.
785 54. Bourgeois Y, Boissinot S. On the population dynamics of junk: a review on the population
786 genomics of transposable elements. Genes. 2019;10(419):1-23.
787 55. Langmüller AM, Nolte V, Dolezal M, Schlötterer C. The genomic distribution of
788 transposable elements is driven by spatially variable purifying selection. Nucleic Acids Res.
789 2023;1-11.

790 56. Eyre-Walker A, Woolfit M, Phelps T. The distribution of fitness effects of new deleterious
791 amino acid mutations in humans. Genetics. 2006;173(2):891-900.
792 57. Lockton S, Gaut BS. The evolution of transposable elements in natural populations of self-
793 fertilizing Arabidopsis thaliana and its outcrossing relative Arabidopsis lyrata. BMC Evol.
794 Biol. 2010;10(10):1-11.
795 58. González J, Macpherson JM, Messer PW, Petrov DA. Inferring the strength of selection in
796 drosophila under complex demographic models. Mol. Biol. Evol. 2009;26(3):513-526.
797 59. Xue AT, Ruggiero RP, Hickerson MJ, Boissinot S. Differential effect of selection against
798 LINE retrotransposons among vertebrates inferred from whole-genome data and
799 demographic modeling. Genome Biol. Evol. 2018;10(5):1265-1281.
800 60. Wang X, Weigel D, Smith LM. Transposon variants and their effects on gene expression in
801 Arabidopsis. PLoS Genet. 2013;9(2):e1003255.
802 61. McCue AD, Nuthikattu S, Slotkin RK. Genome-wide identification of genes regulated in
803 trans by transposable element small interfering RNAs. RNA Biol. 2013;10(8):1379-1395.
804 62. McCue AD, Nuthikattu S, Reeder SH, Slotkin RK. Gene expression and stress response
805 mediated by the epigenetic regulation of a transposable element small RNA. PLoS Genet.
806 2012.8(2):e1002474.
807 63. McCue AD, Slotkin RK. Transposable element small RNAs as regulators of gene expression.
808 Trends Genet. 2012;28(12):616-623.
809 64. Cho J. Transposon-derived non-coding RNAs and their function in plants. Front. Plant Sci.
810 2018;9(600):1-6.

811 65. Wyler M, Keller B, Roulin AC. Potential impact of TE-derived sRNA on gene regulation in
812 the grass Brachypodium distachyon. bioRxiv. 2022.
813 https://www.biorxiv.org/content/10.1101/2022.04.05.487121v1
814 66. Wicker T, Sabot F, Hua-Van A., Bennetzen JL, Capy P, Chalhoub B et al. A unified
815 classification system for eukaryotic transposable elements. Nat. Rev. Genet. 2007;8 973-
816 982.
817 67. Skalska A, Stritt C, Wyler M, Williams HW, Vickers M, Han J, et al. Genetic and methylome
818 variation in turkish Brachypodium distachyon accessions differentiate two geographically
819 distinct subpopulations. Int. J. Mol. Sci. 2020;21(18):6700.
820 68. Gordon SP, Contreras-Moreira B, Levy JJ, Djamei A, Czedik-Eysenberg A, Tartaglio VS, et
821 al. Gradual polyploid genome evolution revealed by pan-genomic analysis of
822 Brachypodium hybridum and its diploid progenitors. Nat. Commun. 2020;11:3670.
823 69. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence
824 data. Bioinformatics. 2014;30(15):2114-2120.
825 70. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie2. Nat. Methods.
826 2012;9:357-359.
827 71. Faust GG, Hall IM. YAHA: fast and flexible long-read alignment with optimal breakpoint
828 detection. Bioinformatics. 2012;28(19):2417-2424.
829 72. Stuart T, Eichten SR, Cahn J, Karpievitch Y, Borevitz JO, Lister R. Population scale mapping
830 of transposable element diversity reveals links to gene regulation and epigenomic
831 variation. eLife. 2016;5:e20777.

832 73. Stritt C, Wyler M, Gimmi EL, Pippel M, Roulin AC. Diversity, dynamics and effects of long
833 terminal repeat retrotransposons in the model grass Brachypodium distachyon. New
834 Phytol. 2020;227(6):1736-1748.
835 74. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome
836 analysis toolkit: a MapReduce framework for analyzing next- generation DNA sequencing
837 data. Genome Res. 2010;20(9):1297-1303.
838 75. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et
839 al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv.
840 2017. https://www.biorxiv.org/content/10.1101/201178v3
841 76. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for
842 annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in
843 the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6(2):80-92.
844 77. Albers PK, McVean G. Dating genomic variants and shared ancestry in population-scale
845 sequencing data. PLoS Biol. 2020;18(1):e3000586.
846 78. Huo N, Garvin DF, You FM, McMahon S, Luo MC, Gu YQ, et al. Comparison of a high-
847 density genetic linkage map to genome features in the model grass Brachypodium
848 distachyon. Theor. Appl. Genet. 2011;123:455-464.
849 79. Dray S, Dufour AB. The ade4 Package: Implementing the duality diagram for ecologists. J.
850 Stat. Softw. 2007;22(4):1-20.
851 80. R Core Team. R: A language and environment for statistical computing. R Foundation for
852 Statistical Computing, Vienna, Austria. 2021. https://www.R-project.org/ Accessed 9
853 August 2023.

854 81. Massicotte P, South A. rnaturalearth: World map data from natural earth. R package
855 version 0.3.3.9000. 2023. https://docs.ropensci.org/rnaturalearth/. Accessed 9 August
856 2023.
857 82. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting
858 recent positive selection in the human genome from haplotype structure. Nature.
859 2002;419:832-837.
860 83. Gautier M, Klassmann A, Vitalis R. rehh 2.0: a reimplementation of the R package rehh to
861 detect positive selection from haplotype structure. Mol. Ecol. Resour. 2017;17(1):78-90.
862 84. Gautier M, Vitalis R. rehh: An R package to detect footprints of selection in genome-wide
863 SNP data from haplotype structure. Bioinformatics. 2012;28(8):1176-1177.
864 85. Fox J, Weisberg S. An R companion to applied regression. Third Edition. Sage. Thousand
865 Oaks CA. 2019. https://socialsciences.mcmaster.ca/jfox/Books/Companion/. Accessed 9
866 August 2023.
867 86. Gautier M. Genome-wide scan for adaptive divergence and association with population-
868 specific covariates. Genetics. 2015;201(4):1555-1579.
869 87. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies.
870 Nat. Genet. 2012;44(7):821-824.
871 88. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful
872 approach to multiple testing. J. R. Stat. Soc. 1995;57(1):289-300.
873 89. Haller BC, Messer PW. Evolutionary modeling in SLiM 3 for beginners. Mol. Biol. Evol.
874 2019;36(5):1101–1109.

875 90. Haller BC, Messer PW. SLiM 3: forward genetic simulations beyond the Wright-Fisher
876 smodel. Mol. Biol. Evol. 2019;36(3):632–637.