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

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. 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. 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

35 ## Introduction

36 Transposable elements (TEs) are an intrinsic part of eukaryotic genomes and their evolution [1-
37 12]. In addition to modulating genome size, the ability of TEs to create genetic diversity through
38 insertion and excision events can lead to new phenotypes on which selection can act. TEs can
39 alter phenotypes through various mechanisms, including the functional disruption of genes [1,
40 2], large-scale changes in the regulatory apparatus [3, 4], alteration of epigenetic landscapes [5,
41 6], ectopic recombination and structural rearrangements [7, 8]. In plants, the dynamics of TE loss
42 and proliferation play a major role in genome evolution [e.g., 9-12]. TEs therefore constitute
43 potentially important drivers of plant evolution, both in nature and during domestication [13].

44 Beyond their influence on genome structure, and given that their transpositional activity
45 can be stress-inducible [for review 14], TEs are often regarded as more likely than classical point
46 mutations to produce the diversity needed for individuals to respond quickly to challenging
47 environments [15-17]. For instance, punctual TE polymorphisms can lead to gains of fitness and
48 evolve under positive selection [2, 18-23]. TE polymorphisms can even induce more extreme
49 changes in gene expression than single nucleotide polymorphisms (SNPs) in plants [24, 25].

50 Despite such evidence, whether TE polymorphisms are major contributors to adaptation
51 to changing environments is still debated. Indeed, TE transposition can be disruptive, and
52 purifying selection has been shown to play an important role in TE evolution [e.g., 26, 27]. Based
53 on simulations, it has been suggested that the persistence of TE polymorphisms within a genome
54 without an uncontrolled accumulation, can only be achieved if weak purifying selection is the
55 main force governing TE evolution [28-31]. The uncertainty surrounding this important question
56 in evolutionary genomics results from the limited number of studies that comprehensively tested

57 the extent to which selection shapes TE allele frequencies, both in plants [23, 32] and animals
58 [26, 33-36] and characterized the distribution of fitness effects of new TE insertions. To clarify
59 this question, we used the plant model system *Brachypodium distachyon* [37] to disentangle the
60 effects of purifying and positive selection on TE polymorphisms in natural populations.

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

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

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

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

91

92 # Results

93 **Genetic variation in *Brachypodium distachyon***

94 Using the *B. distachyon* diversity panel (Fig. 1A), we identified 97,660 TE polymorphisms in our
95 *B. distachyon* dataset, of which 9,172 were retrotransposons, 52,249 were DNA-transposons and
96 36,239 were unclassified. We also identified 9 million SNPs across the 326 samples, including
97 182,801 synonymous SNPs. A Principal Component Analysis (PCA) performed either with SNPs or
98 TE polymorphisms reflects the previously described population structure of *B. distachyon* [38,
99 39], with the first two components of the PCA splitting the data according to the demographic
100 structure (Additional file 1: Fig. S1). Investigating the genetic variation caused by

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

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

114

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

116 To examine the overall contribution of TEs to evolution and adaptation in *B. distachyon*, we first
117 identified regions of the genomes that were likely affected by recent selective sweeps. The fast
118 increase in the frequency of a beneficial allele is expected to lead to a longer than average
119 haplotype around the mutation under positive selection. Such events (known as selective
120 sweeps) can be identified by computing the integrated haplotype score (iHS) around focal
121 mutations [43]. We therefore computed iHS along the genome for the four derived genetic
122 clades. Regions of the genomes with significantly higher iHS than average are expected to harbor

123 mutations that were under positive selection during evolution and adaptation. We hypothesized
124 that if TEs constitute an important part of the genetic makeup that led to adaptation in a given
125 genetic clade, then they should be more frequently fixed or at higher frequencies in regions with
126 high iHS than in the corresponding regions that did not experience recent selective sweeps in
127 other clades.

128 First, we tested if more TE polymorphisms were fixed in a specific region of the genome
129 if a genetic clade had a high iHS, and presumably experienced a selective sweep, than in other
130 genetic clades. An analysis of covariance (ANCOVA) revealed that the number of fixed TE
131 polymorphisms per clade did not significantly differ between high iHS regions and the same
132 regions in other clades (Table 1). These results indicate that there is no correlation between the
133 overall number of fixed TE polymorphisms per clade in a region and recent selective sweeps.
134 However, the number of fixed TEs in genomic regions along the genome was significantly affected
135 by the total number of TEs in the region, the TE superfamily, the TE age, the genetic clade and
136 the overall genetic features of the region (e.g., recombination rate, see Materials and methods)
137 but not by the iHS itself (Table 1). Similarly, we tested if the allele frequency of TE polymorphisms
138 was significantly higher in a specific region of the genome if a genetic clade had a high iHS than
139 in other genetic clades. A second ANCOVA revealed that the allele frequency of TE
140 polymorphisms was significantly influenced by the TE superfamily, TE age, clade and overall
141 genetic features of the region but not by the iHS (Table 2). These results indicate that TEs in high
142 iHS regions did not experience a significant increase in their frequency and that TEs in high iHS
143 regions are experiencing the same selective constraints as other TEs. Similar results were
144 obtained when investigating the number of fixed TE polymorphisms (Additional file 2: Table S1)

145 and the allele frequency of TE polymorphisms (Additional file 2: Table S2) in high iHS regions using
146 a subset of our dataset with an expected lower false negative TE call rate, that only included
147 samples with a genome-wide mapping coverage of at least 20x (see Discussion and Materials and
148 Methods for more details).

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

167 To further examine the contribution of TEs to adaptation, we tested whether and how
168 many TE polymorphisms were significantly associated with environmental factors. If the presence
169 of a TE provides an advantage in a certain environment and contributes to adaptation, we
170 expected a correlation between the environment and the presence/absence of this TE. In this
171 context, we performed genome-environment association analyses (GEA) using all TEs and SNPs
172 identified across the 326 samples and 32 environmental factors associated with precipitation,
173 solar radiation, temperature, elevation and aridity (see in Materials and methods for the full list).
174 The GEA revealed that only nine of the 97,660 TE polymorphisms were significantly associated
175 with some environmental factors (Additional file 2: Table S3), confirming that TEs only had a
176 limited contribution to adaptation in *B. distachyon.* Importantly, two of these nine TEs were
177 found in a gene, and three were in the vicinity of genes (less than 2 kilobase (kb) away, Additional
178 file 2: Table S3).

179

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

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

187 These shifts could be the result of purifying selection as the analyses presented above
188 indicate that positive selection has a negligible effect on TE polymorphism frequencies in

189 *B. distachyon*. However, in contrast to SNPs, TEs do not evolve in a clock-like manner, as their
190 transposition rate is known to vary between generations [46, 47]. Changes in transposition rate
191 and purifying selection can lead to similar shifts in the SFS but can be disentangled using age-
192 adjusted SFS [48]. In brief, if TE polymorphisms are evolving neutrally, they are expected to
193 accumulate on average at the same rate in a population as neutral SNPs of the same age. Hence,
194 Δ frequency, the difference between the average frequency of TE polymorphisms and neutral
195 sites in a specific age bin, will remain close to 0 regardless of the polymorphisms' age. In contrast,
196 if TE polymorphisms evolve under purifying selection, they will tend to occur at lower frequencies
197 than neutral SNPs of the same age, as selection will prevent them from accumulating in the
198 population. Consequently, the Δ frequency will reach negative values for older TE polymorphisms
199 [48].

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

206 As previous studies showed that the distance between TE polymorphisms and the next
207 gene can impact the strength of selection affecting TEs [6, 49, 50], we further split our
208 retrotransposon polymorphisms into three categories based on their distance to the next gene:
209 retrotransposons (i) in and up to 1 kb away from genes, (ii) between 1 kb and 5 kb away and (iii)
210 more than 5 kb away. The age-adjusted SFS of all three categories displayed the same pattern as

211 that observed for the whole retrotransposon polymorphism dataset: older retrotransposon
212 polymorphisms were significantly less frequent than neutral sites of the same age regardless of
213 their distance to genes (one-sided Wilcoxon test, Bonferroni corrected *p* value < 0.01), indicating
214 that retrotransposons more than 5 kb away from genes are also affected by purifying selection
215 (Fig. 2).

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

233 To assess whether similar forces may drive retrotransposon and DNA-transposon
234 evolution, we repeated the analysis for DNA-transposons. The age-adjusted SFS of DNA-
235 transposons revealed very similar patterns, with Δ frequency showing significant deviations from
236 0 in older age bins (one-sided Wilcoxon test, Bonferroni corrected p value < 0.01), but DNA-
237 transposon polymorphisms seemed less deleterious than non-synonymous SNPs and high effect
238 SNPs (Additional file 1: Fig. S7).

239

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

241 To evaluate to what extent the proportion of neutrally evolving mutations in the focal group of
242 mutations affects the shape of the age-adjusted SFS, we ran forward simulation with mutations
243 under multiple selective constraints, and we tested what ratio of neutral to selected mutations
244 can lead to an age-adjusted SFS similar to that observed for retrotransposons in *B. distachyon*.
245 Specifically, we investigated the conditions under which we observed a Δ frequency in the oldest
246 age bin significantly smaller than Δ frequency in all other age bins. Our simulations revealed that
247 the shape of the age-adjusted SFS of retrotransposons could only be reproduced if less than 10%
248 of the mutations were neutrally evolving for most of the selective constraint investigated
249 (Additional file 1: Fig. S8 and Additional file 2: Table S4).

250 Finally, we used the results from our simulations to narrow down the selection strength
251 affecting retrotransposons in *B. distachyon* by investigating the age of the oldest
252 retrotransposons in our dataset. The main difference between the age-adjusted SFS of mutations
253 evolving under weak and strong purifying selection is that the oldest mutations are much older
254 in the simulation with weak purifying selection than in the simulation with strong purifying

255 selection. This age difference arises because mutations under strong purifying selection are
256 removed from the population more effectively and, therefore, cannot persist as long in the
257 population. Examining the age of the last retrotransposon bins in the age-adjusted SFS revealed
258 that the ages of the oldest retrotransposons were the most similar to the expected ages of the
259 oldest mutations in our simulations, with a scaled selection coefficient (S) of -5 and -8 (Fig. 3),
260 indicating that retrotransposons in *B. distachyon* are under moderate purifying selection. In
261 simulations with a nearly neutral selection coefficient (S = -1), the simulated mutations were
262 much older than the oldest observed retrotransposons (Fig. 3). Conversely, in simulations with a
263 strong purifying selection coefficient (S < -10), they were much younger than the oldest observed
264 retrotransposons (Fig. 3).

265

266 Discussion

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

277 in the natural population of this model species are under weak to moderate purifying selection,
278 with only a small minority of TE polymorphisms evolving neutrally.

279

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

281 By combining complementary approaches, we were able to demonstrate that TEs are rarely the
282 target of positive selection in *B. distachyon*. We first probed for footprints of positive selection
283 on TE polymorphisms using the five genetic clades as focal populations. In conducting this
284 analysis, we did not find TE polymorphisms to be at high frequencies or fixed at higher rates than
285 expected, in regions of the genome presumably harboring selective sweeps in at least one of the
286 genetic clades (high iHS regions). This suggested that TEs were rarely the target of positive
287 selection, which we confirmed with a genome-wide scan for overly differentiated TE
288 polymorphisms using X^tX analysis. Indeed, this approach revealed that only a very small
289 proportion of TE polymorphisms are more differentiated than expected under a neutral scenario.

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

299 in close proximity of genes, making these polymorphisms very good candidates for future
300 functional validation.

301 Single TE insertions can have a drastic impact on phenotypic variation and be affected by
302 positive selection [for review, see 16, 53, 54]. For instance, TEs have increased in frequency
303 through positive selection in humans [23] or during range expansion in *Arabidopsis* [25] and
304 *D. melanogaster* [20, 22]. Evidently, *B. distachyon* exhibits a different pattern, as causal
305 mutations for adaptation in this grass species are rarely TEs. Only a few studies have thoroughly
306 quantified the extent to which positive selection influences the evolution of TEs [25, 26, 28, 30,
307 31]. But two of these drew similar conclusions to us, in the green anole *Anolis carolinensis* [30]
308 and in the invasive species *Drosophila Suzukii* [31]. In addition, a large number of candidate genes
309 for adaptation were identified with a similar approach focusing on SNPs [39], indicating that
310 population structure or demographic events are not limiting factors for the methods we used.
311 Altogether, these observations call for a closer investigation of which forces, e.g., purifying
312 selection or neutral evolution, are important in shaping TE allele frequency in natural
313 populations.

314

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

316 Our results suggest that purifying selection is an important factor limiting the ability of TE
317 polymorphisms to fix and increase their frequency in *B. distachyon*. This finding that purifying
318 selection is the main force shaping the landscape of TE polymorphisms in *B. distachyon* is in line
319 with similar observations made for example in maize [55], *Arabidopsis thaliana* [56, 57] and
320 *Drosophila simulans* [58]. Indeed, one of the significant explanatory variables in our ANCOVA

321 models was the genetic clade, a proxy for the effective population size (N_e), which affects the
322 efficiency with which selection can fix beneficial mutations and purge deleterious ones. In *B.*
323 *distachyon*, the number of fixed TE polymorphisms per clade and the frequency of TE
324 polymorphisms were negatively correlated with N_e, indicating that the accumulation of TEs is
325 significantly lower in genetic clades with a larger N_e, potentially because of a greater efficacy of
326 purifying selection.

327 It is widely accepted that most new TE insertions have a deleterious or no effect on the
328 fitness of the host [22, 26, 28, 30, 33-36, 58]. To properly quantify the effect of purifying selection
329 on TE evolution in *B. distachyon*, we used age-adjusted SFS analyses to evaluate the selective
330 constraint experienced by TE polymorphisms while accounting for previously reported changes
331 in their activity [32]. While this method can only be applied to retrotransposons (because the
332 model does not allow back mutations), it provided a first clue on the importance of purifying
333 selection on TE evolution and revealed that overall, retrotransposons evolved under purifying
334 selection in all four derived genetic clades. Indeed, the Δ frequency was significantly smaller than
335 0, especially for older retrotransposons, meaning that old retrotransposons are less common
336 than neutrally evolving SNPs at the same age. This further demonstrates that even after
337 accounting for the different genetic clades and using a large sample size, retrotransposons evolve
338 under purifying selection in *B. distachyon*. One caveat of the approach used in this study is that
339 TE calling pipelines based on short-reads tend to have higher false positive and false negative call
340 rates than SNP calling pipelines, which is also the case for the TEPID TE calling pipeline used here
341 [57, 59]. A high false negative TE calling rate however might bias our TE frequency estimates
342 toward lower frequencies, which could drive the observed patterns in the age-adjusted SFS. To

343 assess if the false negative TE calling rate in our study substantially affected our results, we re-
344 run the age-adjusted SFS on a subset of our dataset only including samples with a genome-wide
345 mapping coverage of at least 20x, as higher mapping coverages are expected to reduce the false
346 negative call rate [27, 59]. Using the TE allele frequencies estimated based on this subset of our
347 data to estimate Δ frequency revealed similar results of the age-adjusted SFS based on the whole
348 dataset (Additional file 1: Fig. S9), indicating that our observation of retrotransposons evolving
349 under purifying selection is not solely driven by a high false negative TE calling rate.

350 We also revealed that only a minority of retrotransposons evolved neutrally, as the
351 observed shape of the Δ frequency curve could only be reproduced in our simulation if the
352 proportion of neutrally evolving mutations in our focal mutations was below 10%. This estimate
353 gives a first glimpse into the distribution of fitness effects of new TE insertions, a fundamental
354 parameter in genetics that describes the way in which new TE insertions can contribute to
355 evolution and adaptation [60]. Here, we show for the first time that new TE insertions have a less
356 than 10% chance to insert into the genome of *B. distachyon* in a way that will allow them to
357 evolve neutrally, advocating for a large potential of TEs to create, through their movement, new
358 phenotypic variation on which selection can act on. PCAs based on TE polymorphisms allowed us
359 to recover the population structure of *B. distachyon*, implying that demographic history and
360 hence neutral processes may indeed partially explain the differences in the TE distribution we
361 observed between genetic clades, as shown in *Arabidopsis thaliana* and *Arabidopsis lyrata* [26,
362 61], *Drosophila melanogaster* [62], humans [63] and the green anole (*Anolis carolinensis*; [30]).
363 However, and in line with our simulations, the first two axes of the PCA explain less than 7% of

364 the variance, indicating that neutrally evolving TEs contribute only mildly to overall TE diversity
365 in our system.

366 Because TEs can cause phenotypic variation through new insertions [1-8], it is not
367 surprising that most new insertions interfere with the function of the genome, especially in a
368 species with a small genome, such as *B. distachyon* (272 Mb) [37]. The proportion of neutrally
369 evolving TE polymorphisms is expected to be very small in genes, as insertions in genic regions
370 are likely to result in loss-of-function [1, 2]. Similarly, TE insertions in close proximity to genes are
371 expected to be highly disruptive, as regulatory elements such as *cis*-regulatory elements are
372 predominantly located in the proximity of genes. In *A. thaliana*, for instance, TEs located in the
373 vicinity of genes (less than 2 kb) globally result in downregulation [64]. Although only specific
374 families alter gene expression in *B. distachyon* [41], the observed Δ frequency for
375 retrotransposon polymorphisms in genes and in their 1 kb surroundings matched our
376 expectations. The fact that TE polymorphisms located more than 5 kb away from genes are also
377 evolving under purifying selection was more surprising. That said, little is known about the
378 distance between *cis*-regulatory sequences and genes in *B. distachyon.* In plants, TEs are believed
379 to affect gene expression in *trans* through the production of small-interfering RNA [65-69].
380 Hence, the fact that only a small proportion of TEs can accumulate neutrally indicates that, in a
381 gene-dense genome such as that of *B. distachyon* (42.5% of the genome are genes) [69], TE
382 insertions in any genomic compartment may result in some *cis*- or *trans*-regulatory effects visible
383 to selection. Finally, we tested whether TE polymorphisms located more than 5 kb away from
384 genes are evolving under purifying selection could be due to mapping or other artefacts by
385 comparing the shape of the age-adjusted SFS of retrotransposons and SNPs more than 5 kb away

386 from genes. However, the age-adjusted SFS of SNPs 5 kb away from genes differs from the one
387 of retrotransposons (Additional file 1: Fig. S10), indicating that the shape of the age-adjusted SFS
388 of retrotransposons more than 5 kb away from genes is not likely to be the result of artefacts in
389 regions of the genome far away from genes.

390 To further ascertain the strength of purifying selection, we used forward simulation and
391 showed that simulations assuming a moderately weak selection pressure (S = -5 or S = -8) against
392 TE polymorphisms best fitted our observed data. In theory, no TE polymorphisms under strong
393 purifying selection should be present in a natural population, as such mutations are expected to
394 be quickly lost, especially in a predominantly selfing species where most loci are expected to be
395 homozygous. Therefore, it is not surprising that TE polymorphisms which persist in *B. distachyon*
396 are under weak to moderate selection, as also shown, for example, for the L1 retrotransposons
397 in humans [27] or the BS retrotransposon family in *Drosophila melanogaster* [62].

398 While some of the parameters we chose for our simulations, such as the dominance or
399 selfing rate, can affect the efficiency of TE purging, it is unlikely that discrepancies in the true and
400 assumed values for these parameters would have led to drastically different results. For example,
401 we assumed codominance for all mutations, which might not hold true for each TE
402 polymorphism. However, because of the high selfing rate observed in *B. distachyon* [38],
403 heterozygous loci are expected to be rare, and dominance is unlikely to have a strong impact on
404 our observations. Similarly, with a higher selfing rate, deleterious TE polymorphisms should be
405 removed more efficiently by purifying selection. To check whether a lower selfing rate could
406 allow a higher proportion of TE polymorphisms to evolve neutrally, we reran the simulations
407 assuming fully outcrossing individuals. This also resulted in simulation with weak to moderate

408 selection strength on TE polymorphisms best fitting the observed data, further strengthening our
409 results (Additional file 1: Fig. S11).

410 While the analyses of positive selection and GEA were based on both DNA-transposons
411 and retrotransposons, we only used retrotransposons to assess the strength of selection on TE
412 polymorphisms, as the age-adjusted SFS was developed with the assumption of no back
413 mutations [48]. Yet, DNA-transposons do not solely transpose through cut and paste mechanisms
414 as they would otherwise not be so abundant in Eukaryotic genomes. DNA-transposons can also
415 create extra copies of themselves by transposing during chromosome replication or repair from
416 a position that has already been replicated, or repaired [70]. We therefore repeated the age-
417 adjusted SFS analyses using DNA-transposons to evaluate whether DNA-transposons were
418 affected by similar selective constraints. The folded SFS of DNA-transposons and
419 retrotransposons display similar shifts toward high proportions of rare alleles and Δ frequency
420 deviations from 0 in the age-adjusted SFS of DNA-transposons and retrotransposons are
421 comparable. Hence, we argue that the conclusion drawn for retrotransposons also holds for DNA-
422 transposons, and that purifying selection affect TEs broadly. To further examine our conclusion
423 on purifying selection, we investigated the selective regime affecting different retrotransposons
424 and DNA-transposons superfamilies. Thereby, we generated age-adjusted SFS for the four most
425 common TE superfamilies Copia, Ty3 (also known under the name Gypsy, but we will avoid using
426 this name because of its problematic nature see [71]), Helitron and MITE and found similar
427 deviations of the Δ frequency from 0 in the four investigated TE superfamilies (Additional file 1:
428 Fig. S12–S15). These results indicate that our conclusion on the broad effect of purifying selection

429 is not driven by a single TE superfamily but is at least common among the four most numerous
430 TE superfamilies.

431

432 **Conclusion**

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

443

444 Materials and methods

445 **Whole-genome resequencing data**

446 In this study, we analyzed a total of 326 publicly available whole-genome sequencing data from
447 *Brachypodium distachyon* accessions sampled around the Mediterranean Basin (Fig. 1A;
448 Additional file 2: Table S5). Our *B. distachyon* dataset consisted of 47 samples published by
449 Gordon et al. [8], 57 samples published by Skalska et al. [72], 65 samples published by Gordon et

450 al. [73], 86 samples published by Stritt et al. [38] and 71 samples published by Minadakis et al.
451 [39], covering all five genetic clades previously described in this species [38, 39]. Each sample was
452 assigned to a genetic clade based on previously published results [39].

453

454 **Data processing**

455 Raw reads were trimmed using Trimmomatic 0.36 [74] and mapped to the *B. distachyon*
456 reference genome version 3.0 [37] using bowtie2 [75] and yaha [76], and TE polymorphisms were
457 identified using the TEPID pipeline [59] and the recently updated TE annotation by Stritt et al.
458 [77] and Wyler et al. [69]. TE polymorphisms include both TE insertion polymorphisms (TIPs;
459 insertions absent from the reference genome but present in at least one natural accession) and
460 TE absence polymorphisms (TAPs; insertions present in the reference genome but absent from
461 at least one natural accession). The class, superfamily and family of each TE call were assigned
462 based on the TEPID results and the TE annotation from the reference genome. TIPs that were
463 less than 100 base pairs (bp) apart in different samples and assigned to the same TE family were
464 merged.

465 Single nucleotide polymorphisms (SNPs) were called using GATK v.4.0.2.1 [78] using
466 HaplotypeCaller [79] following Minadakis et al. [39]. The SNP calls were hard filtered using the
467 following conditions: QD < 5.0; FS > 20.0; SOR > 3.0; MQ < 50.0; MQRankSum < 2.5; MQRankSum
468 > -2.5; ReadPosRankSum < 2.0; ReadPosRankSum > -2.0. Because *B. distachyon* displays a high
469 selfing rate [38], most genetic variants are expected to be homozygous within an individual.
470 Hence, all TE calls were treated as homozygous, and heterozygous SNP calls were removed from
471 our dataset to reduce false variant calls. Additionally, all sites with multiallelic TE and SNP calls

472 were removed. SNPs were classified as synonymous, non-synonymous and of high fitness effect
473 using SnpEff [80]. SNPs and TE polymorphisms were merged into a single vcf file using custom
474 scripts provided in github (see section Availability of data and materials).

475 To estimate the age of each SNP and TE polymorphism, the SNPs and TEs found in the
476 A_East, A_Italia, B_East and B_West clades were polarized using the C clade, which was identified
477 as the most ancestral *B. distachyon* clade [38] and used as the outgroup throughout this study.
478 An estimate for the time of origin of all SNPs and TE polymorphisms was calculated with GEVA, a
479 nonparametric approach that relies on pairwise differences in identity by descent (IBD) regions
480 around the focal mutation to estimate the time of origin [81]. GEVA was run separately for each
481 clade using the genetic map produced by Huo et al. [82] and a mutation rate of $7x10^{-9}$
482 substitutions/generation. The theoretical prediction of the correlation between allele age and
483 allele frequency of neutrally evolving mutations based on N_e [42] was compared to the observed
484 correlation between allele age and frequency of synonymous SNPs to check the sanity of the age
485 estimates.

486 The observed SNP and TE diversity was first examined using a principal component
487 analysis (PCA), and correlations between TE diversity and genetic clades were tested with a
488 mantel test using the ade4 package version 1.7-22 [83] in R version 4.1.2 [84]. The folded site
489 frequency spectrum (SFS) was computed for TE polymorphisms and SNPs using the minor allele
490 frequency in R version 4.1.2 [84]. Finally, the map of the geographical distribution of the used
491 accessions was done in R using the rnaturalearth package 0.3.3 [85].

492 Mapping coverage is known to influence false discovery rate [27, 59]. To investigate the
493 impact of false positive and false negative TE calls on our results, we down sampled the TE dataset

494 to only include TEs that have been called in samples that had at least an average mapping
495 coverage of 20x. The allele frequencies of TEs present in our high coverage dataset was
496 recalculated only considering samples with at least an average mapping coverage of 20x. This
497 second TE dataset was then used to check if using a dataset with a higher mapping coverage and
498 presumably a lower false TE calling rate impacted our results.

499

500 **Analyses of positive selection**

501 Regions of the genome affected by positive selection were identified using the integrated
502 haplotype score (iHS), a measure of the amount of extended haplotype homozygosity along the
503 ancestral allele relative to the derived allele for a given polymorphic site [86]. iHS was calculated
504 using the SNP dataset, and regions displaying longer haplotypes and hence high iHS were
505 identified in R using the rehh package [87, 88]. The threshold to distinguish between regions of
506 high iHS and other regions was selected such that less than 5% of the *B. distachyon* genome was
507 classified as high iHS regions in each clade (Additional file 2: Table S6). Candidate regions under
508 positive selection were defined as all regions that were found to have high iHS in each clade
509 separately.

510 A first ANCOVA was used to model the number of fixed TE polymorphisms in each clade
511 found in the candidate region under positive selection based on the following genetic features:
512 total number of TEs, TE superfamily, TE age (split into three categories: young: age < 10,000
513 generations; intermediate: age between 10,000 generations and 60,000 generations; old: age >
514 60,000 generations), clade, genomic region (a unique ID for each candidate region under positive
515 selection) and iHS classification of the regions in each clade (high or average). A second ANCOVA

516 was used to model the allele frequency of TE polymorphisms found in the candidate region under
517 positive selection based on the following genetic features: TE superfamily, TE age, clade, genomic
518 region and iHS classification of the regions in each clade. The TE superfamily was included to
519 account for different evolutionary behaviors of TEs from different superfamilies. Age accounted
520 for differences in the fixation rate and frequency distribution between young and old TEs. The
521 clade was included to account for clade-specific differences such as differences in N_e. Finally, a
522 unique ID for each candidate region under positive selection was included to account for region-
523 specific differences such as differences in the recombination rate and GC content. In the end,
524 regions that were found to have a high iHS in some clades were compared to the same regions
525 in the other clades. All ANCOVAs were run in R using the car package [89].

526 The standardized allele frequency of a mutation across populations (X^tX) values [44] were
527 computed for the combined TE and SNP dataset using Baypass version 2.3 [45, 90]. The X^tX values
528 were used to identify over- and under differentiated TE polymorphisms between clades. A
529 pseudo-observed dataset (POD) of 100,000 SNPs was simulated under the demographic model
530 inferred from the covariance matrix of the SNP dataset. The POD was then used to determine the
531 97.5% (over-differentiated polymorphisms) and 2.5% (under differentiated polymorphisms)
532 quantiles.

533

534 **Genome-environment association analyses**

535 We identified TE polymorphisms significantly associated with environmental factors using
536 genome-environment association analyses (GEA) following Minadakis et al. [39]. GEAs were run
537 with GEMMA 0.98.5 [91] using the combined TE and SNP vcf file against the following 32

538 environmental factors extracted by Minadakis et al. [39]: altitude, aridity from March to June,
539 aridity from November to February, annual mean temperature, mean temperature of warmest
540 quarter, mean temperature of coldest quarter, annual precipitation, precipitation of wettest
541 month, precipitation of driest month, precipitation seasonality, precipitation of wettest quarter,
542 precipitation of driest quarter, precipitation of warmest quarter, precipitation of coldest quarter,
543 mean diurnal Range, isothermality, temperature seasonality, maximum temperature of warmest
544 month, minimum temperature of coldest month, temperature annual range, mean temperature
545 of wettest quarter, mean temperature of driest quarter, precipitation from March to June,
546 precipitation from November to February, solar radiation from March to June, solar radiation
547 from November to February, mean temperature between March and June, mean temperature
548 between November and February, maximum temperature between March and June, maximum
549 temperature between November and February, minimum temperature between March and June
550 and minimum temperature between November and February. We applied a False Discovery Rate
551 (FDR, [92]) threshold of 5% to control for false positive rates.

552

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

554 Footprints of purifying selection on TE polymorphisms were first evaluated using folded SFS. An
555 age-adjusted site frequency spectrum (age-adjusted SFS) approach was used to further
556 investigate the impact of purifying selection on retrotransposons while accounting for
557 nonconstant transposition rates. Briefly, the age-adjusted SFS is a summary statistic that
558 describes the difference between the average frequency of TEs at a specific age and the average
559 frequency of neutral sites of the same age [48]. Therefore, the TE dataset was sorted by age and

560 split into equally large bins with respect to the number of observations in each age bin. Neutral
561 sites were then randomly down-sampled to match the number of observations in the TE dataset
562 and its age distribution [48].

563 The difference between the average TE and neutral site frequency, or Δ frequency, was
564 computed for each age bin [48]. This method allows for an unbiased comparison between the
565 allele frequencies of TEs and neutral sites, and is robust to transposition rate changes and
566 demographic changes [48]. However, the theory behind this method was developed assuming
567 no back mutations and is therefore best suited for retrotransposons, as DNA-transposons can
568 exit an insertion site [48]. We used the synonymous SNPs identified with SnpEff as the neutrally
569 evolving sites. However, because estimating the population wide frequency of TEs is more
570 challenging than estimating SNP frequencies, putative biases in frequency estimates need to be
571 assessed before performing age-adjusted SFS analyses. To do so, the SNP dataset was resampled
572 so that the SNP dataset used in the age-adjusted SFS had a frequency distribution that matched
573 the observed TE frequency distribution. The age-adjusted SFS of retrotransposons was
574 contrasted against the age-adjusted SFS of non-synonymous, as well as against high fitness effect
575 SNPs. Therefore, 10,000 non-synonymous and high fitness effect SNPs were randomly selected
576 for each clade to reach approximately the same number of retrotransposon polymorphisms, non-
577 synonymous and high fitness effect SNPs for final comparisons. To estimate the variation in Δ
578 frequency estimates, all age-adjusted SFS were computed 100 times. All Wilcoxon tests and
579 Bonferroni *p* value corrections were done in R version 4.1.2 [84].

580

581 **Forward simulation**

582 We used SLiM 4.0.1 [93, 94] to run forward simulations and assess the proportion of neutrally
583 evolving retrotransposons and the average selection strength affecting them. The simulations
584 were designed to reflect the population size and demographic history of *B. distachyon*. The
585 simulated genomic fragment was 1 megabase (Mb) long and included neutral (synonymous)
586 mutations as well as focal mutations that evolved under different selective constraints. The focal
587 mutations were a mix of neutrally evolving mutations and mutations evolving under a constant
588 selection pressure. Therefore, the ratio (r) of focal mutations that evolved neutrally was either
589 0%, 5%, 10%, 25% or 50%. The scaled selection coefficient (S, defined as $N_e s$, with s the strength
590 of selection and N_e the effective population size) affecting the remaining focal mutations was set
591 at the beginning of the simulation to be either -1, -5, -8, -10, -12, -15, -20 or -50 to cover
592 effectively neutral ($0 > S \geq -1$), intermediate ($-1 > S \geq -10$) and strongly deleterious ($-10 > S$)
593 selective constraints. The selfing rate was set to 70%, as *B. distachyon* is a highly selfing species
594 with occasional outcrossing [38, 39]. In addition, a high recombination rate was chosen to
595 minimize the effects of linked selection in the small genomic fragment simulated. Simulations for
596 each combination of these two parameters were run 20 times to assess the variation in the
597 resulting age-adjusted SFS. The shape of the resulting age-adjusted SFS was used to narrow down
598 the ratio of neutrally evolving TE polymorphisms. Similarly, the age distribution of the mutations
599 in the oldest bin of the age-adjusted SFS was used to narrow down the strength of selection
600 affecting TE polymorphisms.

601

602 # Acknowledgments

We would like to thank Jeffrey Ross-Ibarra and Mitra Menon as well as Fabrizio Menardo, Michael Thieme, Wenbo Xu, Jigisha, Lars Kaderli and Serafin Schefer for all the discussions and their comments on this project. We thank Emmanuelle Botté (https://manuscribe.com.au) for professional editing of the manuscript.

Declarations

Consent for publication

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

Availability of data and materials

The datasets supporting the conclusions of this article are publicly available on the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home) and National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/sra/), and the archive numbers of the accessions used are listed in the Additional file 1: Table S5. The scripts generated are available on GitHub https://github.com/Roberthorv/TE_in_Brachypodium/tree/main.

Funding

This project was funded by the Swiss National Science Foundation (project 31003A_182785) and the Research Priority Program Evolution in Action from the University of Zürich. Data analyzed in this paper were generated in collaboration with the Genetic Diversity Center (GDC), ETH Zürich.

624 **Authors' contributions**

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

630

631 # References

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

635 2. Hof AE, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, et al. The industrial melanism
636 mutation in British peppered moths is a transposable element. Nature. 2016;534:102-
637 105.

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

640 4. Qiu Y, Köhler C. Mobility connects: transposable elements wire new transcriptional
641 networks by transferring transcription factor binding motifs. Biochem. Soc. Trans.
642 2020;48(3):1005-1017.

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

645 6. Hollister JD, Gaut BS. Epigenetic silencing of transposable elements: a trade-off between
646 reduced transposition and deleterious effects on neighboring gene expression. Genome
647 Res. 2009;19(8):1419-1428.
648 7. Xiao H, Jiang N, Schaffner E, Stockinger EJ, van der Knaap E. A retrotransposon-mediated
649 gene duplication underlies morphological variation of tomato fruit. Science.
650 2008;319(5869):1527-1530.
651 8. Gordon SP, Contreras-Moreira B, Woods DP, Des Marais DL, Burgess D, Shu S, et al.
652 Extensive gene content variation in the Brachypodium distachyon pan-genome correlates
653 with population structure. Nat. Commun. 2017;8:2184.
654 9. Bennetzen JL, Kellogg EA. Do plants have a one-way ticket to genomic obesity? Plant Cell.
655 1997;9(9):1509-1514.
656 10. Vitte C, Panaud O. Formation of solo-LTRs through unequal homologous recombination
657 counterbalances amplifications of LTR retrotransposons in rice Oryza sativa L. Mol. Biol.
658 Evol. 2003;20(4):528–540.
659 11. Piégu B, Guyot R, Picault N, Roulin A, Sanyal A, Kim H, et al. Doubling genome size without
660 polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza
661 australiensis, a wild relative of rice. Genome Res. 2006;16(10):1262-1269.
662 12. Wendel JF, Jackson SA, Meyers BC, Wing RA. Evolution of plant genome architecture.
663 Genome Biol. 2016;17(37):1-14.
664 13. Lisch D. How important are transposons for plant evolution? Nat. Rev. Genet. 2013;14:49-
665 61.

666 14. Negi P, Rai AN, Suprasanna P. Moving through the stressed genome: emerging regulatory
667 roles for transposons in plant stress response. Front. Plant Sci. 2016;7(1448):1-20.
668 15. Rey O, Danchin E, Mirouze M, Loot C, Blanchet S. Adaptation to Global Change: A
669 Transposable Element-Epigenetics Perspective. Trends Ecol. Evol. 2016;31(7):514-526.
670 16. Dubin MJ, Mittelsten Scheid O, Becker C. Transposons: a blessing curse. Curr. Opin. Plant
671 Biol. 2018;42:23-29.
672 17. Quadrana L, Etcheverry M, Gilly A, Caillieux E, Madoui MA, Guy J, et al. Transposition
673 favors the generation of large effect mutations that may facilitate rapid adaption. Nat.
674 Commun. 2019;10(3421):1-10.
675 18. González J, Karasov TL, Messer PW, Petrov DA. Genome-wide patterns of adaptation to
676 temperate environments associated with transposable elements in Drosophila. PLoS
677 Genet. 2010;6(4):e1000905.
678 19. Studer A, Zhao Q, Ross-Ibarra J, Doebley J. Identification of a functional transposon
679 insertion in the maize domestication gene tb1. Nat. Genet. 2011;43(11):1160-1163.
680 20. Barrón MG, Fiston-Lavier AS, Petrov DA, González J. Population genomics of transposable
681 elements in Drosophila. Annu. Rev. Genet. 2014;48:561-581.
682 21. Rishishwar L, Wang L, Wang J, Yi SV, Lachance J, Jordan IK. Evidence for positive selection
683 on recent human transposable element insertions. Gene. 2018;675:69-79.
684 22. Niu XM, Xu YG, Li ZW, Bian YT, Hou XH, Chen JF, et al. Transposable elements drive rapid
685 phenotypic variation in Capsella rubella. Proc. Nati. Acad. Sci. USA. 2019;116(14):6908-
686 6913.

687 23. Jiang J, Xu YC, Zhang ZQ, Chen JF, Niu XM, Hou XH, et al. Forces driving transposable
688 element load variation during Arabidopsis range expansion. bioRxiv. 2022;1-12.
689 https://www.biorxiv.org/content/10.1101/2022.12.28.522087v1
690 24. Uzunović J, Josephs EB, Stinchcombe JR, Wright SI. Transposable elements are important
691 contributors to standing variation in gene expression in Capsella grandiflora. Mol. Biol.
692 Evol. 2019;36(8):1734-1745.
693 25. Castanera R, Morales-Díaz N, Gupta S, Purugganan M, Casacuberta JM. Transposons are
694 a major contributor to gene expression variability under selection in rice populationse.
695 Life. 2023;12:RP86324.
696 26. Bourgeois Y, Ruggiero RP, Hariyani I, Boissinot S. Disentangling the determinants of
697 transposable elements dynamics in vertebrate genomes using empirical evidences and
698 simulations. PLoS Genet. 2020;16(10):e1009082.
699 27. Stritt C, Gordon SP, Wicker T, Vogel JP, Roulin AC. Recent activity in expanding populations
700 and purifying selection have shaped transposable element landscapes across natural
701 accessions of the mediterranean grass Brachypodium distachyon. Genome Biol. Evol.
702 2018;10(1):304-318
703 28. Charlesworth B. Transposable elements in natural populations with a mixture of selected
704 and neutral insertion sites. Genet. Res. 1991;57(2):127-134.
705 29. Charlesworth B, Langley CH, Sniegowski PD. Transposable element distributions in
706 Drosophila. Genetics. 1997;147(4):1993-1995.
707 30. Charlesworth B, Charlesworth D. The population dynamics of transposable elements.
708 Genet. Res. 1983;42(1):1-27.

709 31. Charlesworth B. Background selection and patterns of genetic diversity in Drosophila
710 melanogaster. Genet. Res. 1996;68(2):131-149.
711 32. Lockton S, Ross-Ibarra J, Gaut BS. Demography and weak selection drive patterns of
712 transposable element diversity in natural populations of Arabidopsis lyrata. Proc. Natl.
713 Acad. Sci. U S A. 2008;105(37):13965–13970.
714 33. Boissinot S, Davis J, Entezam A, Petrov D, Furano AV. Fitness cost of LINE-1 (L1) activity in
715 humans. Proc. Natl. Acad. Sci. U S A. 2006;103(25):9590-9594.
716 34. Blumenstiel JP, Chen X, He M, Bergman CM. An age-of-allele test of neutrality for
717 transposable element insertions. Genetics. 2014;196(2):523-38.
718 35. Rech GE, Bogaerts-Márquez M, Barrón MG, Merenciano M, Villanueva-Cañas JL, Horváth
719 V, et al. Stress response, behavior, and development are shaped by transposable element-
720 induced mutations in Drosophila. PLoS Genet. 2019;15(2):e1007900.
721 36. Mérel V, Gibert P, Buch I, Rodriguez Rada V, Estoup A, Gautier M, et al. The Worldwide
722 Invasion of Drosophila suzukii Is Accompanied by a Large Increase of Transposable
723 Element Load and a Small Number of Putatively Adaptive Insertions. Mol. Biol. Evol.
724 2021;38(10):4252-4267.
725 37. International Brachypodium Initiative. Genome sequencing and analysis of the model
726 grass Brachypodium distachyon. Nature. 2010;463:763-768.
727 38. Stritt C, Gimmi EL, Wyler M, Bakali AH, Skalska A, Hasterok R, et al. Migration without
728 interbreeding: Evolutionary history of a highly selfing Mediterranean grass inferred from
729 whole genomes. Mol. Ecol. 2022;31(1):70-85.

730 39. Minadakis N, Williams H, Horvath R, Caković D, Stritt C, Thieme M, et al. The demographic
731 history of the wild crop relative Brachypodium distachyon is shaped by distinct past and
732 present ecological niches. Peer Community J. 2023;3(e84):1-19.
733 40. Bourgeois Y, Stritt C, Walser JC, Gordon SP, Vogel JP, Roulin AC. Genome-wide scans of
734 selection highlight the impact of biotic and abiotic constraints in natural populations of
735 the model grass Brachypodium distachyon. Plant J. 2018;96(2):438-451.
736 41. Wyler M, Stritt C, Walser JC, Baroux C, Roulin AC. Impact of transposable elements on
737 methylation and gene expression across natural accessions of Brachypodium distachyon.
738 Genome Biol. Evol. 2020;12(11):1994-2001.
739 42. Kimura M, Ohta T. The age of a neutral mutant persisting in a finite population. Genetics.
740 1973;75(1):199–212.
741 43. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the
742 human genome. PLoS Biol. 2006;4:e72.
743 44. Günther T, Coop G. Robust identification of local adaptation from allele frequencies.
744 Genetics. 2013;195(1):205-20.
745 45. Olazcuaga L, Loiseau A, Parrinello H, Paris M, Fraimout A, Guedot C, et al. A whole-genome
746 scan for association with invasion success in the fruit fly Drosophila suzukii using contrasts
747 of allele frequencies corrected for population structure. Mol. Biol. Evol. 2020;37(8):2369-
748 2385.
749 46. Garcia Guerreiro MP. What makes transposable elements move in the Drosophila
750 genome. Heredity. 2012;108(5):461–468.

751 47. Belyayev A. Bursts of transposable elements as an evolutionary driving force. J. Evol. Biol.
752 2014;27(12):2573–2584.
753 48. Horvath R, Menon M, Stitzer M, Ross-Ibarra J. Controlling for variable transposition rate
754 with an age-adjusted site frequency spectrum. Genome Biol. Evol. 2022;14(2):evac016.
755 49. Wright SI, Agrawal N, Bureau TE. Effects of recombination rate and gene density on
756 transposable element distributions in Arabidopsis thaliana. Genome Res.
757 2003;13(8):1897-1903.
758 50. Horvath R, Slotte T. The Role of small RNA-based epigenetic silencing for purifying
759 selection on transposable elements in Capsella grandiflora. Genome Biol. and Evol.
760 2017;9(10):2911-2920.
761 51. Raissig MT, Woods DP. The wild grass Brachypodium distachyon as a developmental
762 model system. Curr. Top. Dev. Biol. 2022;147:33-71.
763 52. Hasterok R, Catalan P, Hazen SP, Roulin AC, Vogel JP, Wang K, et al. Brachypodium: 20
764 years as a grass biology model system; the way forward? Trends Plant Sci.
765 2022;27(10):1002-1016.
766 53. Casacuberta E, González J. The impact of transposable elements in environmental
767 adaptation. Mol. Ecol. 2013;22(6):1503-1517.
768 54. Bourgeois Y, Boissinot S. On the population dynamics of junk: a review on the population
769 genomics of transposable elements. Genes. 2019;10(419):1-23.
770 55. Stitzer MC, Khaipho-Burch MB, Hudson AI, Song B, Valdez-Franco JA, Ramstein G et al.
771 Transposable element abundance subtly contributes to lower fitness in maize. bioRxiv.
772 2023. https://www.biorxiv.org/content/10.1101/2023.09.18.557618v1

773 56. Quadrana L, Silveira AB, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh JA et al. The
774 Arabidopsis thaliana mobilome and its impact at the species level. eLife 2016;5:e15716.
775 57. Baduel P, Leduque B, Ignace A, Gy I, Gil JJr, Loudet O et al. Genetic and environmental
776 modulation of transposition shapes the evolutionary potential of Arabidopsis thaliana.
777 Genome Biol. 2021;22(138):1-26.
778 58. Langmüller AM, Nolte V, Dolezal M, Schlötterer C. The genomic distribution of
779 transposable elements is driven by spatially variable purifying selection. Nucleic Acids Res.
780 2023;1-11.
781 59. Stuart T, Eichten SR, Cahn J, Karpievitch Y, Borevitz JO, Lister R. Population scale mapping
782 of transposable element diversity reveals links to gene regulation and epigenomic
783 variation. eLife. 2016;5:e20777.
784 60. Eyre-Walker A, Woolfit M, Phelps T. The distribution of fitness effects of new deleterious
785 amino acid mutations in humans. Genetics. 2006;173(2):891-900.
786 61. Lockton S, Gaut BS. The evolution of transposable elements in natural populations of self-
787 fertilizing Arabidopsis thaliana and its outcrossing relative Arabidopsis lyrata. BMC Evol.
788 Biol. 2010;10(10):1-11.
789 62. González J, Macpherson JM, Messer PW, Petrov DA. Inferring the strength of selection in
790 drosophila under complex demographic models. Mol. Biol. Evol. 2009;26(3):513-526.
791 63. Xue AT, Ruggiero RP, Hickerson MJ, Boissinot S. Differential effect of selection against
792 LINE retrotransposons among vertebrates inferred from whole-genome data and
793 demographic modeling. Genome Biol. Evol. 2018;10(5):1265-1281.

794 64. Wang X, Weigel D, Smith LM. Transposon variants and their effects on gene expression in
795 Arabidopsis. PLoS Genet. 2013;9(2):e1003255.
796 65. McCue AD, Nuthikattu S, Slotkin RK. Genome-wide identification of genes regulated in
797 trans by transposable element small interfering RNAs. RNA Biol. 2013;10(8):1379-1395.
798 66. McCue AD, Nuthikattu S, Reeder SH, Slotkin RK. Gene expression and stress response
799 mediated by the epigenetic regulation of a transposable element small RNA. PLoS Genet.
800 2012.8(2):e1002474.
801 67. McCue AD, Slotkin RK. Transposable element small RNAs as regulators of gene expression.
802 Trends Genet. 2012;28(12):616-623.
803 68. Cho J. Transposon-derived non-coding RNAs and their function in plants. Front. Plant Sci.
804 2018;9(600):1-6.
805 69. Wyler M, Keller B, Roulin AC. Potential impact of TE-derived sRNA on gene regulation in
806 the grass Brachypodium distachyon. bioRxiv. 2022.
807 https://www.biorxiv.org/content/10.1101/2022.04.05.487121v1
808 70. Wicker T, Sabot F, Hua-Van A., Bennetzen JL, Capy P, Chalhoub B et al. A unified
809 classification system for eukaryotic transposable elements. Nat. Rev. Genet. 2007;8 973-
810 982.
811 71. Wei K, Aldaimalani R, Mai D, Zinshteyn D, Satyaki PRV, Blumenstiel J, et al. Rethinking the
812 "gypsy" retrotransposon: A roadmap for community-driven reconsideration of
813 problematic gene names. OSF Preprints: 2022;1-13.
814 https://doi.org/10.31219/osf.io/fma57

815 72. Skalska A, Stritt C, Wyler M, Williams HW, Vickers M, Han J, et al. Genetic and methylome
816 variation in turkish Brachypodium distachyon accessions differentiate two geographically
817 distinct subpopulations. Int. J. Mol. Sci. 2020;21(18):6700.
818 73. Gordon SP, Contreras-Moreira B, Levy JJ, Djamei A, Czedik-Eysenberg A, Tartaglio VS, et
819 al. Gradual polyploid genome evolution revealed by pan-genomic analysis of
820 Brachypodium hybridum and its diploid progenitors. Nat. Commun. 2020;11:3670.
821 74. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence
822 data. Bioinformatics. 2014;30(15):2114-2120.
823 75. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie2. Nat. Methods.
824 2012;9:357-359.
825 76. Faust GG, Hall IM. YAHA: fast and flexible long-read alignment with optimal breakpoint
826 detection. Bioinformatics. 2012;28(19):2417-2424.
827 77. Stritt C, Wyler M, Gimmi EL, Pippel M, Roulin AC. Diversity, dynamics and effects of long
828 terminal repeat retrotransposons in the model grass Brachypodium distachyon. New
829 Phytol. 2020;227(6):1736-1748.
830 78. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome
831 analysis toolkit: a MapReduce framework for analyzing next- generation DNA sequencing
832 data. Genome Res. 2010;20(9):1297-1303.
833 79. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et
834 al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv.
835 2017. https://www.biorxiv.org/content/10.1101/201178v3

836 80. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for
837 annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in
838 the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6(2):80-92.
839 81. Albers PK, McVean G. Dating genomic variants and shared ancestry in population-scale
840 sequencing data. PLoS Biol. 2020;18(1):e3000586.
841 82. Huo N, Garvin DF, You FM, McMahon S, Luo MC, Gu YQ, et al. Comparison of a high-
842 density genetic linkage map to genome features in the model grass Brachypodium
843 distachyon. Theor. Appl. Genet. 2011;123:455-464.
844 83. Dray S, Dufour AB. The ade4 Package: Implementing the duality diagram for ecologists. J.
845 Stat. Softw. 2007;22(4):1-20.
846 84. R Core Team. R: A language and environment for statistical computing. R Foundation for
847 Statistical Computing, Vienna, Austria. 2021. https://www.R-project.org/ Accessed 9
848 August 2023.
849 85. Massicotte P, South A. rnaturalearth: World map data from natural earth. R package
850 version 0.3.3.9000. 2023. https://docs.ropensci.org/rnaturalearth/. Accessed 9 August
851 2023.
852 86. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting
853 recent positive selection in the human genome from haplotype structure. Nature.
854 2002;419:832-837.
855 87. Gautier M, Klassmann A, Vitalis R. rehh 2.0: a reimplementation of the R package rehh to
856 detect positive selection from haplotype structure. Mol. Ecol. Resour. 2017;17(1):78-90.

857 88. Gautier M, Vitalis R. rehh: An R package to detect footprints of selection in genome-wide
858 SNP data from haplotype structure. Bioinformatics. 2012;28(8):1176-1177.
859 89. Fox J, Weisberg S. An R companion to applied regression. Third Edition. Sage. Thousand
860 Oaks CA. 2019. https://socialsciences.mcmaster.ca/jfox/Books/Companion/. Accessed 9
861 August 2023.
862 90. Gautier M. Genome-wide scan for adaptive divergence and association with population-
863 specific covariates. Genetics. 2015;201(4):1555-1579.
864 91. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies.
865 Nat. Genet. 2012;44(7):821-824.
866 92. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful
867 approach to multiple testing. J. R. Stat. Soc. 1995;57(1):289-300.
868 93. Haller BC, Messer PW. Evolutionary modeling in SLiM 3 for beginners. Mol. Biol. Evol.
869 2019;36(5):1101–1109.
870 94. Haller BC, Messer PW. SLiM 3: forward genetic simulations beyond the Wright-Fisher
871 smodel. Mol. Biol. Evol. 2019;36(3):632–637.
872

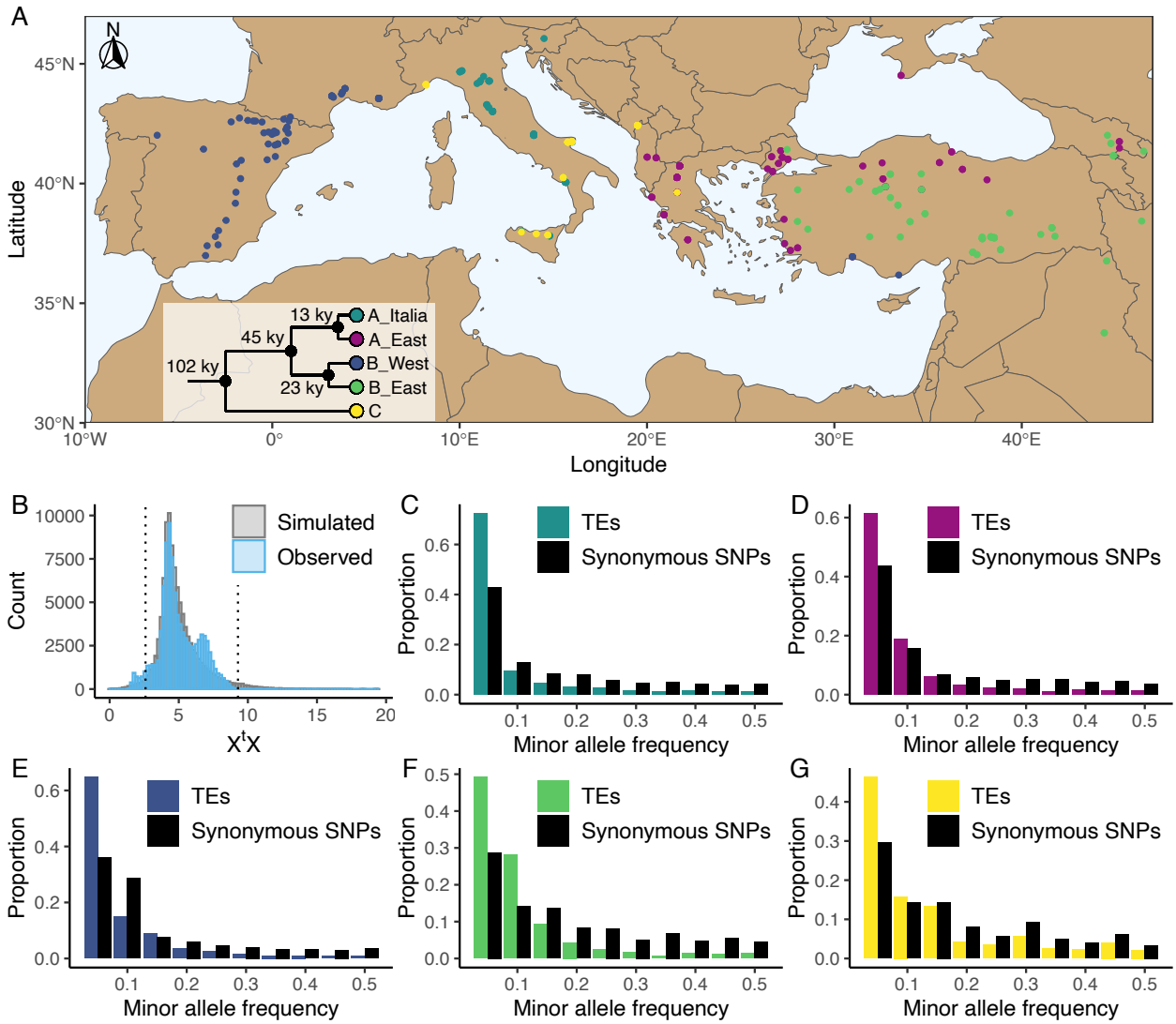

Fig 1. Distribution of the studied accessions and TE polymorphism frequencies. Panel A: Map showing the geographical distribution of the accessions used in the current study. The phylogenetic tree illustrates the phylogeny between the five genetic clades. This panel was made based on the data and results 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.

875

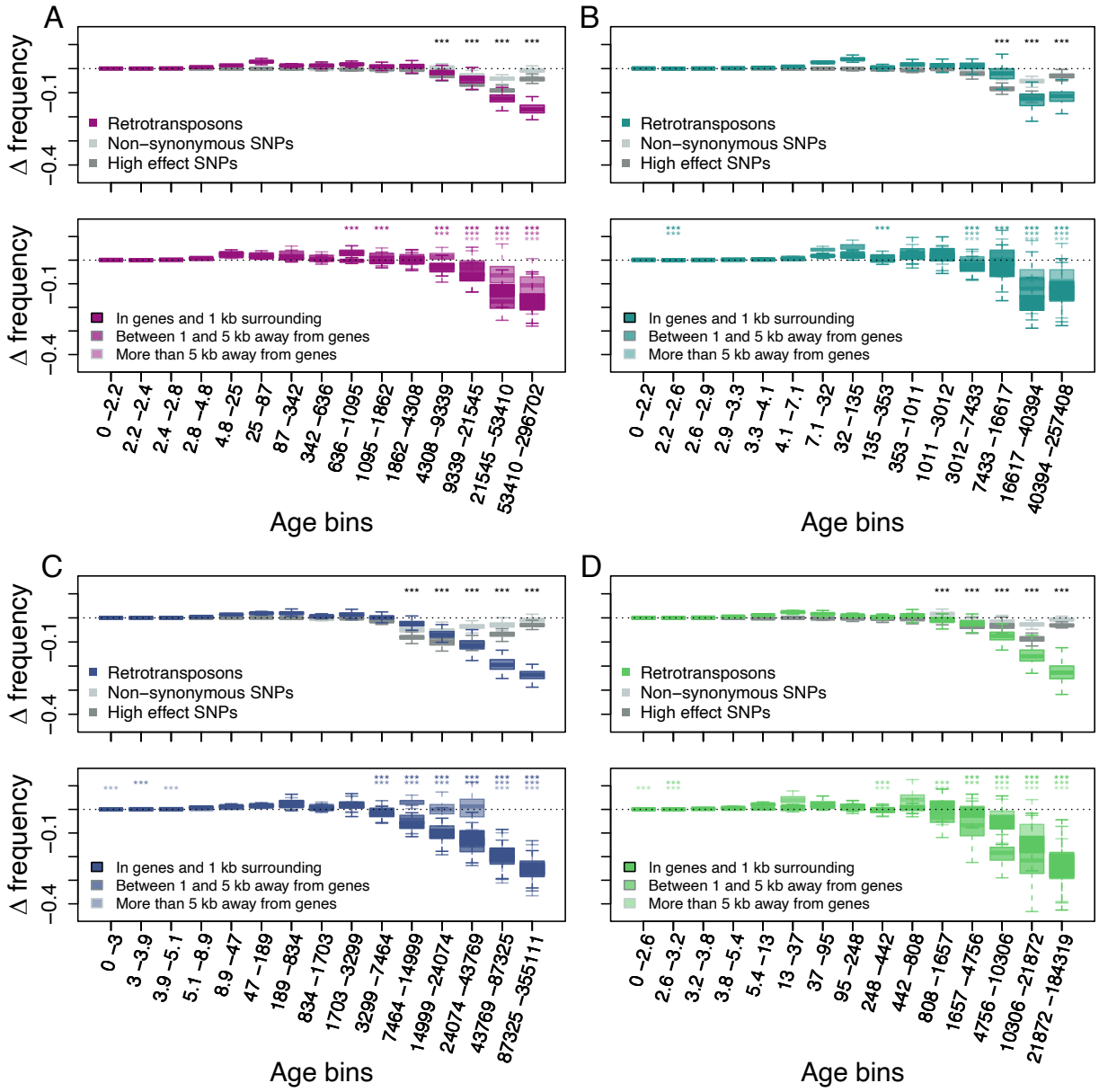

Fig. 2 Age-adjusted SFS of retrotransposons. The top row shows the age-adjusted SFS of all retrotransposons (colored), non-synonymous SNPs (light gray) and high effect SNPs (dark gray) in the four derived clades. The bottom row shows the age-adjusted SFS of retrotransposons based on their distance to 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: ***).

876

877

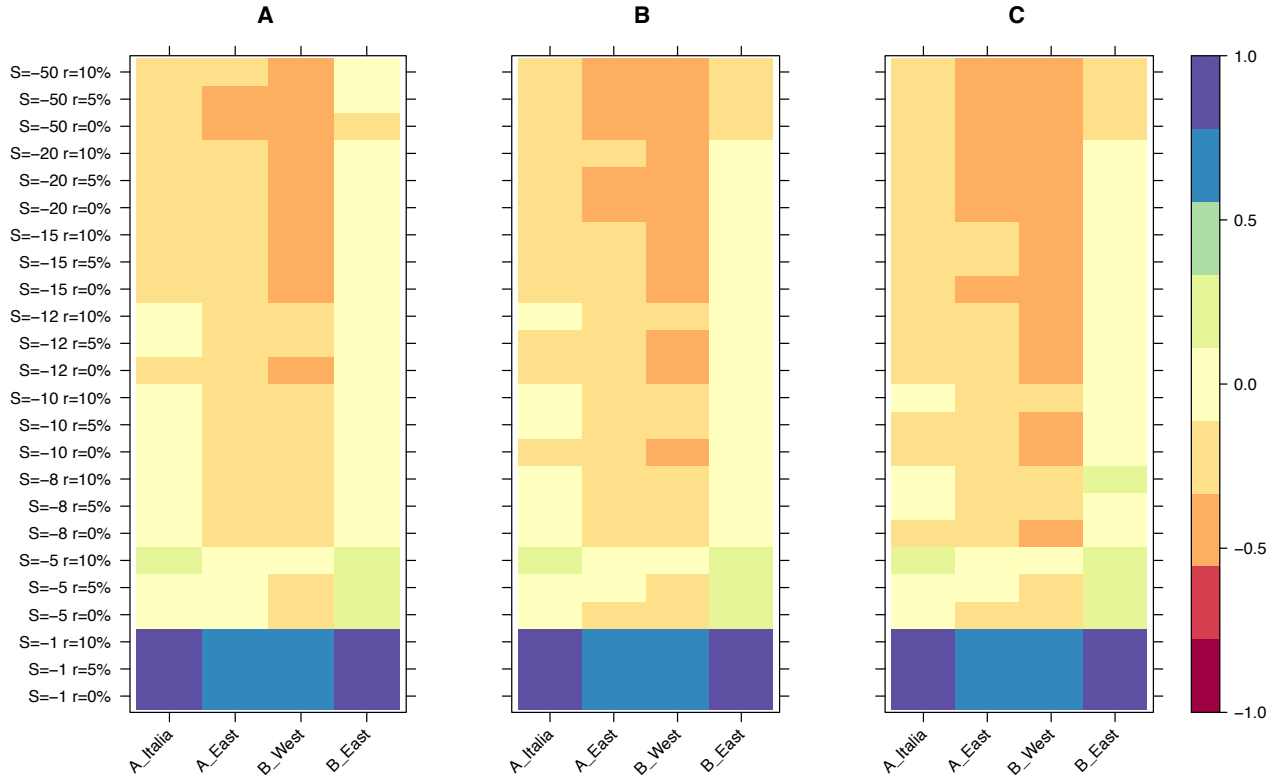

Fig. 3 Relative age difference ((mutation age in simulations - observed mutation age)/maximum absolute age difference) between simulated and observed data in the last bin of the age-adjusted SFS. A: 25% quantile; B: 50% quantile; C: 75% quantile.

878

879 **Table 1** ANCOVA predicting the number of fixed TE polymorphisms per clade in candidate regions
880 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

881

882

883 **Table 2** ANCOVA predicting the allele frequency of TE polymorphisms per clade in candidate
884 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

885