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

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

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

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

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

Results

Genetic variation in *Brachypodium distachyon*

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

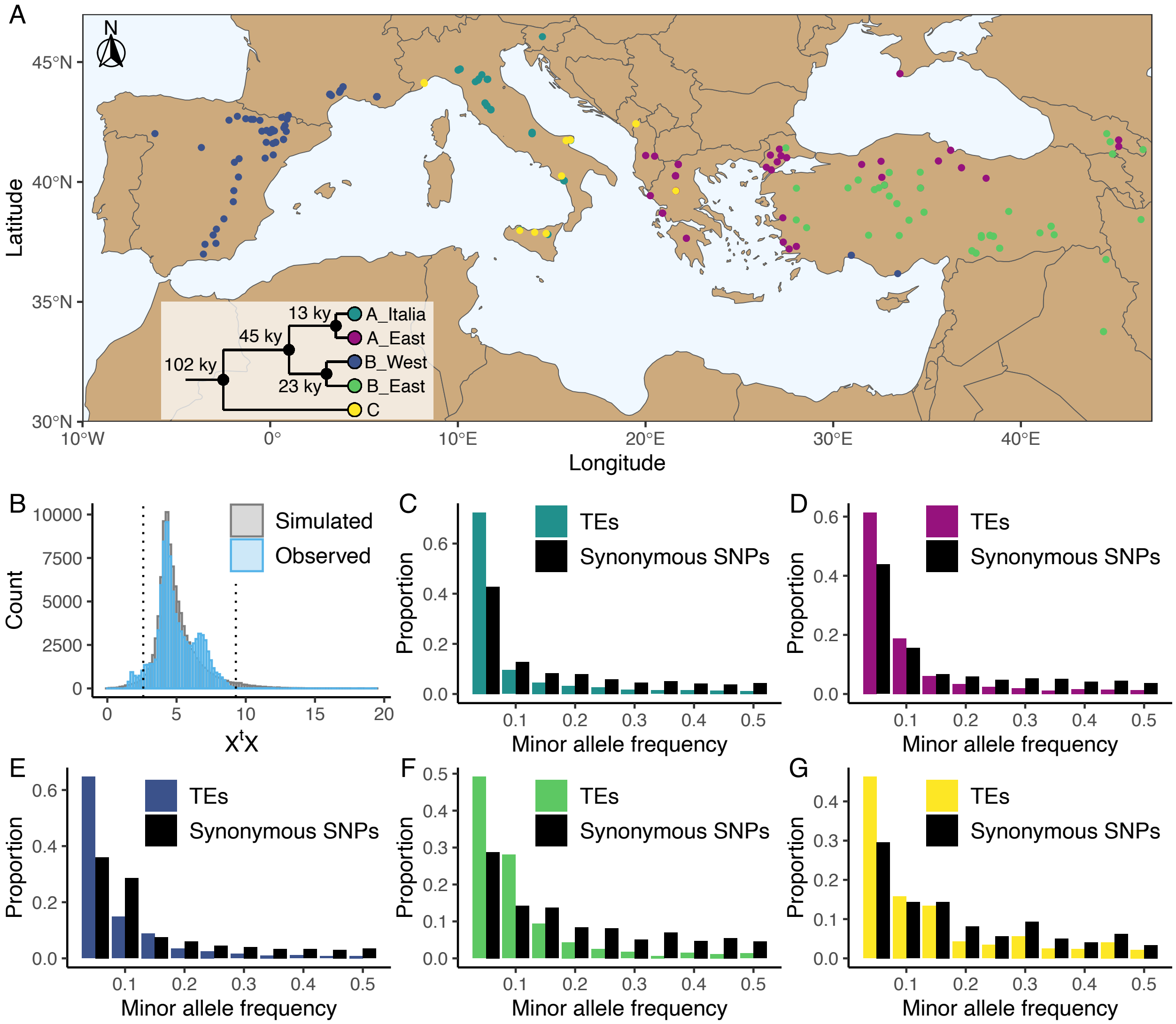

Fig 1. Distribution of the studied accessions and TE polymorphism frequencies. Panel A: Map showing the geographical distribution of the accessions used in the current study. The phylogenetic tree illustrates the phylogeny between the five genetic clades. This panel was made based on the data and results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

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

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

To examine the overall contribution of TEs to evolution and adaptation in *B. distachyon*, we first identified regions of the genomes that were likely affected by recent selective sweeps. The fast increase in the frequency of a beneficial allele is expected to lead to a longer than average haplotype around the mutation under positive selection. Such events (known as selective sweeps) can be identified by computing the integrated haplotype score (iHS) around focal mutations [43]. We therefore computed iHS along the genome for the four derived genetic clades. Regions of the genomes with significantly higher iHS than average are expected to harbor mutations that were under positive selection during evolution and adaptation. We hypothesized that if TEs constitute an important part of the genetic makeup that led to adaptation in a given genetic clade, then they should be more frequently fixed or at higher frequencies in regions with high iHS than in the corresponding regions that did not experience recent selective sweeps in other clades.

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

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

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

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

To further examine the contribution of TEs to adaptation, we tested whether and how many TE polymorphisms were significantly associated with environmental factors. If the presence of a TE provides an advantage in a certain environment and contributes to adaptation, we expected a correlation between the environment and the presence/absence of this TE. In this context, we performed genome-environment association analyses (GEA) using all TEs and SNPs identified across the 326 samples and 32 environmental factors associated with precipitation, solar radiation, temperature, elevation and aridity (see in Materials and methods for the full list). The GEA revealed that only nine of the 97,660 TE polymorphisms were significantly associated with some environmental factors (Additional file 2: Table S1), confirming that TEs only had a limited contribution to adaptation in *B. distachyon.* Importantly, two of these nine TEs were found in a gene, and three were in the vicinity of genes (less than 2 kilobase (kb) away, Additional file 2: Table S1).

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

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

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

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

As previous studies showed that the distance between TE polymorphisms and the next gene can impact the strength of selection affecting TEs [6, 49, 50], we further split our retrotransposon polymorphisms into three categories based on their distance to the next gene: retrotransposons (i) in and up to 1 kb away from genes, (ii) between 1 kb and 5 kb away and (iii) more than 5 kb away. The age-adjusted SFS of all three categories displayed the same pattern as that observed for the whole retrotransposon polymorphism dataset: older retrotransposon polymorphisms were significantly less frequent than neutral sites of the same age regardless of their distance to genes (one-sided Wilcoxon test, Bonferroni corrected p value < 0.01), indicating that retrotransposons more than 5 kb away from genes are also affected by purifying selection (Fig. 2).

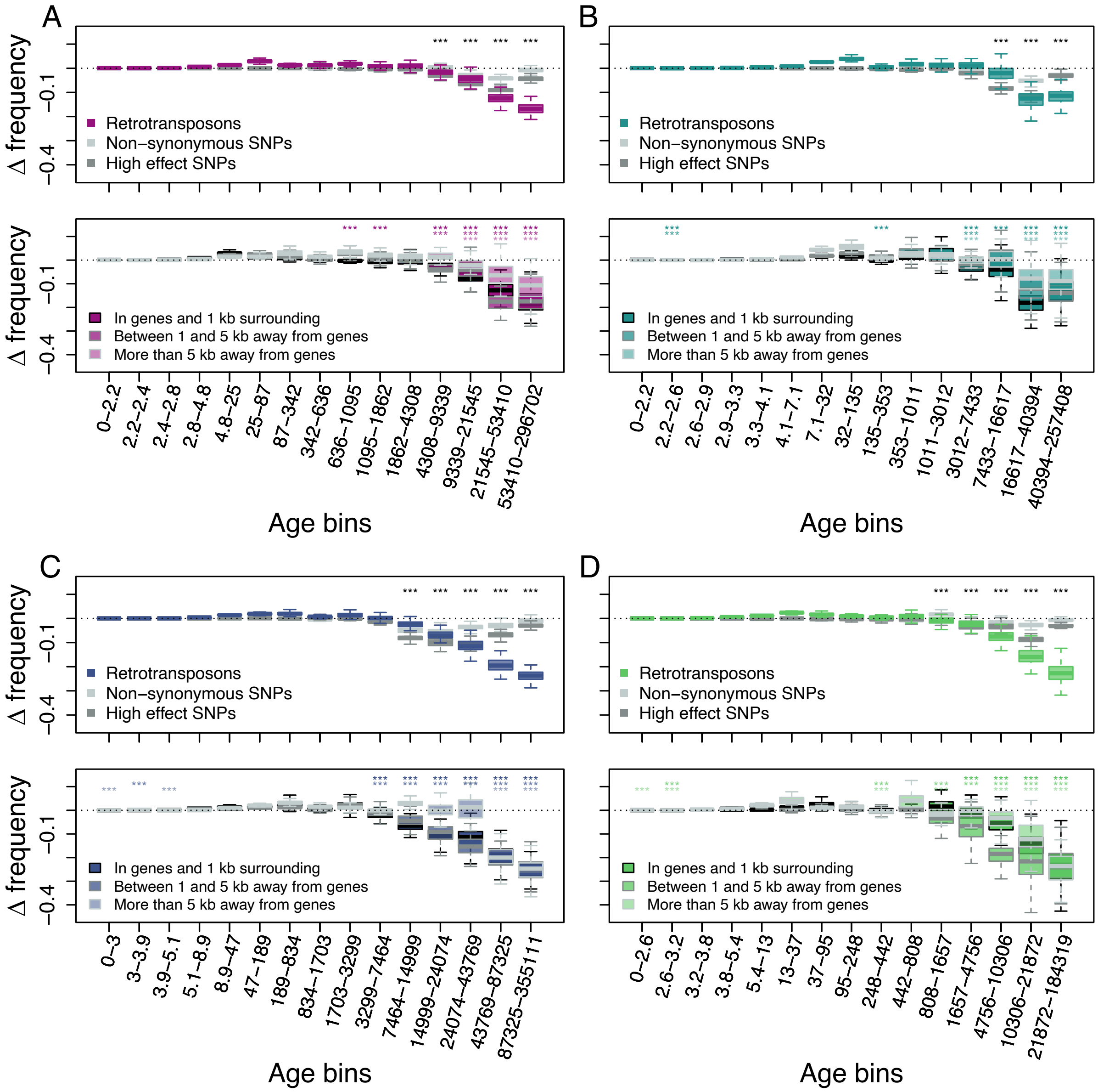

Fig. 2 Age-adjusted SFS of retrotransposons. The top row shows the age-adjusted SFS of all retrotransposons (colored), non-synonymous SNPs (light gray) and high effect SNPs (dark gray) in the four derived clades. The bottom row shows the age-adjusted SFS of retrotransposons based on their distance to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

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

To assess whether similar forces may drive retrotransposon and DNA-transposon evolution, we repeated the analysis for DNA-transposons. The age-adjusted SFS of DNA-transposons revealed very similar patterns, with Δ frequency showing significant deviations from 0 in older age bins (one-sided Wilcoxon test, Bonferroni corrected *p* value < 0.01) and DNA-transposon polymorphisms being more deleterious than non-synonymous SNPs and high effect SNPs (Additional file 1: Fig. S9).

Forward simulations allow us to quantify the strength of purifying selection

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

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

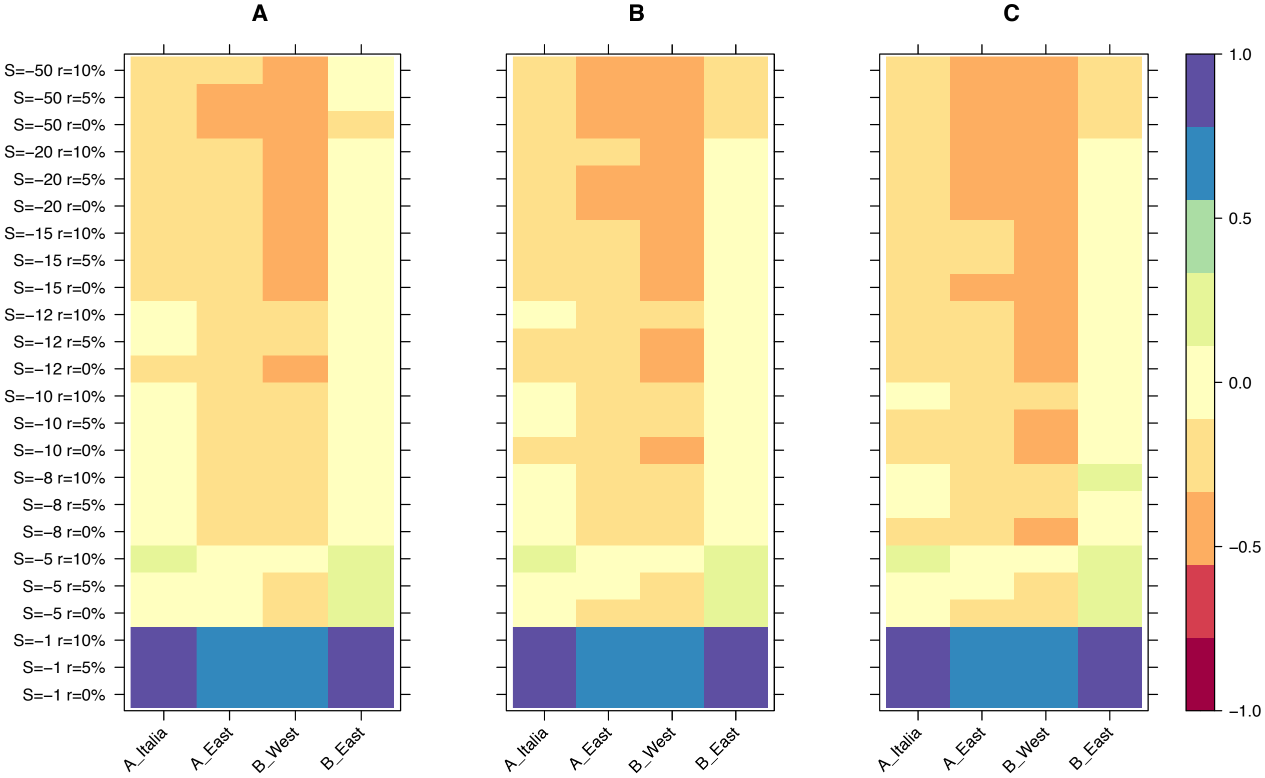

Fig. 3 Relative age difference ((mutation age in simulations - observed mutation age)/maximum absolute age difference) between simulated and observed data in the last bin of the age-adjusted SFS. A: 25% quantile; B: 50% quantile; C: 75% quantile.

Discussion

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

Rare instances of positive selection on TEs

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

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

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

Moderate purifying selection is the dominant force during TE evolution

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

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

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

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

Because TEs can cause phenotypic variation through new insertions [1-8], it is not surprising that most new insertions interfere with the function of the genome, especially in a species with a small genome, such as *B. distachyon* (272 Mb) [37]. The proportion of neutrally evolving TE polymorphisms is expected to be very small in genes, as insertions in genic regions are likely to result in loss-of-function [1, 2]. Similarly, TE insertions in close proximity to genes are expected to be highly disruptive, as regulatory elements such as *cis*-regulatory elements are predominantly located in the proximity of genes. In *A. thaliana*, for instance, TEs located in the vicinity of genes (less than 2 kb) globally result in downregulation [60]. Although only specific families alter gene expression in *B. distachyon* [41], the observed Δ frequency for retrotransposon polymorphisms in genes and in their 1 kb surroundings matched our expectations. The fact that TE polymorphisms located more than 5 kb away from genes are also evolving under purifying selection was more surprising. That said, little is known about the distance between *cis*-regulatory sequences and genes in *B. distachyon.* In plants, TEs are believed to affect gene expression in *trans* through the production of small-interfering RNA [61-65]. Hence, the fact that only a small proportion of TEs can accumulate neutrally indicates that, in a gene-dense genome such as that of *B. distachyon* (42.5% of the genome are genes) [65], TE insertions in any genomic compartment may result in some *cis*- or *trans*-regulatory effects visible to selection.

To further ascertain the strength of purifying selection, we used forward simulation and showed that simulations assuming a moderately weak selection pressure (S = -5 or S = -8) against TE polymorphisms best fitted our observed data. In theory, no TE polymorphisms under strong purifying selection should be present in a natural population, as such mutations are expected to be quickly lost, especially in a predominantly selfing species where most loci are expected to be homozygous. Therefore, it is not surprising that TE polymorphisms in *B. distachyon* are under weak to moderate selection, as also shown, for example, for the L1 retrotransposons in humans [27] or the BS retrotransposon family in *Drosophila melanogaster* [58].

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

While the analyses of positive selection and GEA were based on both DNA-transposons and retrotransposons, we only used retrotransposons to assess the strength of selection on TE polymorphisms, as the age-adjusted SFS was developed with the assumption of no back mutations [48]. Yet, DNA-transposons do not solely transpose through cut and paste mechanisms as they would otherwise not be so abundant in Eukaryotic genomes. DNA-transposons can also create extra copies of themselves by transposing during chromosome replication or repair from a position that has already been replicated, or repaired [66]. We therefore repeated the age-adjusted SFS analyses using DNA-transposons to evaluate whether DNA-transposons were affected by similar selective constraints. The folded SFS of DNA-transposons and retrotransposons display similar shifts toward high proportions of rare alleles and Δ frequency deviations from 0 in the age-adjusted SFS of DNA-transposons and retrotransposons are comparable. Hence, we argue that the conclusion drawn for retrotransposons also holds for DNA-transposons, and that purifying selection affect TEs broadly.

Conclusion

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

Materials and methods

Whole-genome resequencing data

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

Data processing

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

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

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

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

Analyses of positive selection

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

A first ANCOVA was used to model the number of fixed TE polymorphisms in each clade found in the candidate region under positive selection based on the following genetic features: total number of TEs, TE superfamily, TE age (split into three categories: young: age < 10,000 generations; intermediate: age between 10,000 generations and 60,000 generations; old: age > 60,000 generations), clade, genomic region (a unique ID for each candidate region under positive selection) and iHS classification of the regions in each clade (high or average). A second ANCOVA was used to model the allele frequency of TE polymorphisms found in the candidate region under positive selection based on the following genetic features: TE superfamily, TE age, clade, genomic region and iHS classification of the regions in each clade. The TE superfamily was included to account for different evolutionary behaviors of TEs from different superfamilies. Age accounted for differences in the fixation rate and frequency distribution between young and old TEs. The clade was included to account for clade-specific differences such as differences in N_e. Finally, a unique ID for each candidate region under positive selection was included to account for region-specific differences such as differences in the recombination rate and GC content. In the end, regions that were found to have a high iHS in some clades were compared to the same regions in the other clades. All ANCOVAs were run in R using the car package [85].

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

Genome-environment association analyses

We identified TE polymorphisms significantly associated with environmental factors using genome-environment association analyses (GEA) following Minadakis et al. [39]. GEAs were run with GEMMA 0.98.5 [87] using the combined TE and SNP vcf file against the following 32 environmental factors extracted by Minadakis et al. [39]: altitude, aridity from March to June, aridity from November to February, annual mean temperature, mean temperature of warmest quarter, mean temperature of coldest quarter, annual precipitation, precipitation of wettest month, precipitation of driest month, precipitation seasonality, precipitation of wettest quarter, precipitation of driest quarter, precipitation of warmest quarter, precipitation of coldest quarter, mean diurnal Range, isothermality, temperature seasonality, maximum temperature of warmest month, minimum temperature of coldest month, temperature annual range, mean temperature of wettest quarter, mean temperature of driest quarter, precipitation from March to June, precipitation from November to February, solar radiation from March to June, solar radiation from November to February, mean temperature between March and June, mean temperature between November and February, maximum temperature between March and June, maximum temperature between November and February, minimum temperature between March and June and minimum temperature between November and February. We applied a False Discovery Rate (FDR, [88]) threshold of 5% to control for false positive rates.

Age-adjusted frequency spectra and analyses of purifying selection

Footprints of purifying selection on TE polymorphisms were first evaluated using folded SFS. An age-adjusted site frequency spectrum (age-adjusted SFS) approach was used to further investigate the impact of purifying selection on retrotransposons while accounting for nonconstant transposition rates. Briefly, the age-adjusted SFS is a summary statistic that describes the difference between the average frequency of TEs at a specific age and the average frequency of neutral sites of the same age [48]. Therefore, the TE dataset was sorted by age and split into equally large bins with respect to the number of observations in each age bin. Neutral sites were then randomly down-sampled to match the number of observations in the TE dataset and its age distribution [48].

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

Forward simulation

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

Supplementary Information

Additional file 1: Supplemental Figures. Figure S1. Principal Component Analyses using TE (left panel) and SNP (right panel) polymorphisms. **Figure S2.** Principal Component Analyses using retrotransposon (left panel) and DNA-transposon (right panel) polymorphisms. **Figure S3.** Observed correlation between age in generations and frequency of synonymous SNPs in the four derived genetic clades. The red points show the expected age of a neutrally evolving mutation at a specific frequency based on the predictions of Kimura and Ohta (1973). Panel A: clade A_East; panel B: clade A_Italia; panel C: clade B_West and panel D: clade B_East. **Figure S4.** Distribution of the observed TE age scaled by the effective population size (N_e) in the four derived genetic clades of *B. distachyon*. **Figure S5.** Folded site frequency spectrum of DNA-transposons and synonymous SNPs in all genetic clades. Panel A: A_East; panel B: A_Italia; panel C: B_West; panel D: B_East; panel E: C. **Figure S6.** Folded site frequency spectrum of retrotransposons and synonymous SNPs in all genetic clades. Panel A: A_East; panel B: A_Italia; panel C: B_West; panel D: B_East; panel E: C. **Figure S7.** Age-adjusted SFS of simulated mutations under negative selection in the four derived clades. The four columns show the results for the A_East, A_Italia, B_West and B_East genetic clades, respectively. Each line shows the results for the different scaled selection coefficients (S). The five colored curves in each plot show the shape of the age-adjusted SFS with varying ratios of neutrally evolving mutations, and the gray curves show variation within one standard deviation based on the 20 runs for each simulation. The X axes show the age bin from the youngest to the oldest, with each age bin including the same number of observations for each simulation. **Figure S8.** Relative age difference ((mutation age in simulations - observed mutation age)/maximum absolute age difference) between simulated data assuming fully outcrossing individuals and observed data in the last bin of the age-adjusted SFS. A: 25% quantile; B: 50% quantile; C: 75% quantile. **Figure S9.** Age-adjusted SFS of DNA-transposons (colored), non-synonymous SNPs (light gray) and high effect SNPs (dark gray) 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 DNA-transposons observations. A: A_East genetic clades; B: A_Italia genetic clades; C: B_West genetic clades; D: B_East genetic clades. Boxplots are based on 100 estimations of Δ frequency.

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

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 for professional editing.

Declarations

Ethics approval and consent to participate

Not applicable.

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 S2. The scripts generated are available on GitHub https://github.com/Roberthorv/TE_in_Brachypodium/tree/main.

Competing interests

The authors declare no competing interests.

Funding

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

Authors' contributions

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

References

1. Bhattacharyya MK, Smith AM, Ellis TH, Hedley C, Martin C. The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme. Cell. 1990;60(1):115-22.
2. Hof AE, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, et al. The industrial melanism mutation in British peppered moths is a transposable element. Nature. 2016;534:102-105.
3. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat. Rev. Genet. 2008;9:397-405.
4. Qiu Y, Köhler C. Mobility connects: transposable elements wire new transcriptional networks by transferring transcription factor binding motifs. Biochem. Soc. Trans. 2020;48(3):1005-1017.
5. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. 2007;8(4):272-85.

6. Hollister JD, Gaut BS. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. 2009;19(8):1419-1428.

7. Xiao H, Jiang N, Schaffner E, Stockinger EJ, van der Knaap E. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Science. 2008;319(5869):1527-1530.

8. Gordon SP, Contreras-Moreira B, Woods DP, Des Marais DL, Burgess D, Shu S, et al. Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. Nat. Commun. 2017;8:2184.

9. Bennetzen JL, Kellogg EA. Do plants have a one-way ticket to genomic obesity? Plant Cell. 1997;9(9):1509-1514.

10. Vitte C, Panaud O. Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice Oryza sativa L. Mol. Biol. Evol. 2003;20(4):528–540.

11. Piégu B, Guyot R, Picault N, Roulin A, Sanyal A, Kim H, et al. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Res. 2006;16(10):1262-1269.

12. Wendel JF, Jackson SA, Meyers BC, Wing RA. Evolution of plant genome architecture. Genome Biol. 2016;17(37):1-14.

13. Lisch D. How important are transposons for plant evolution? Nat. Rev. Genet. 2013;14:49-61.

14. Negi P, Rai AN, Suprasanna P. Moving through the stressed genome: emerging regulatory roles for transposons in plant stress response. Front. Plant Sci. 2016;7(1448):1-20.

15. Rey O, Danchin E, Mirouze M, Loot C, Blanchet S. Adaptation to Global Change: A Transposable Element-Epigenetics Perspective. Trends Ecol. Evol. 2016;31(7):514-526.

16. Dubin MJ, Mittelsten Scheid O, Becker C. Transposons: a blessing curse. Curr. Opin. Plant Biol. 2018;42:23-29.

17. Quadrana L, Etcheverry M, Gilly A, Caillieux E, Madoui MA, Guy J, et al. Transposition favors the generation of large effect mutations that may facilitate rapid adaption. Nat. Commun. 2019;10(3421):1-10.

18. Uzunović J, Josephs EB, Stinchcombe JR, Wright SI. Transposable elements are important contributors to standing variation in gene expression in Capsella grandiflora. Mol. Biol. Evol. 2019;36(8):1734-1745.

19. Castanera R, Morales-Díaz N, Gupta S, Purugganan M, Casacuberta JM. Transposons are a major contributor to gene expression variability under selection in rice populationse. Life. 2023;12:RP86324.

20. González J, Karasov TL, Messer PW, Petrov DA. Genome-wide patterns of adaptation to temperate environments associated with transposable elements in Drosophila. PLoS Genet. 2010;6(4):e1000905.

21. Studer A, Zhao Q, Ross-Ibarra J, Doebley J. Identification of a functional transposon insertion in the maize domestication gene tb1. Nat. Genet. 2011;43(11):1160-1163.

22. Barrón MG, Fiston-Lavier AS, Petrov DA, González J. Population genomics of transposable elements in Drosophila. Annu. Rev. Genet. 2014;48:561-581.

23. Rishishwar L, Wang L, Wang J, Yi SV, Lachance J, Jordan IK. Evidence for positive selection on recent human transposable element insertions. Gene. 2018;675:69-79.

24. Niu XM, Xu YG, Li ZW, Bian YT, Hou XH, Chen JF, et al. Transposable elements drive rapid phenotypic variation in Capsella rubella. Proc. Nati. Acad. Sci. USA. 2019;116(14):6908-6913.

25. Jiang J, Xu YC, Zhang ZQ, Chen JF, Niu XM, Hou XH, et al. Forces driving transposable element load variation during Arabidopsis range expansion. bioRxiv. 2022;1-12. https://www.biorxiv.org/content/10.1101/2022.12.28.522087v1

26. Lockton S, Ross-Ibarra J, Gaut BS. Demography and weak selection drive patterns of transposable element diversity in natural populations of Arabidopsis lyrata. Proc. Natl. Acad. Sci. U S A. 2008;105(37):13965–13970.

27. Boissinot S, Davis J, Entezam A, Petrov D, Furano AV. Fitness cost of LINE-1 (L1) activity in humans. Proc. Natl. Acad. Sci. U S A. 2006;103(25):9590-9594.

28. Blumenstiel JP, Chen X, He M, Bergman CM. An age-of-allele test of neutrality for transposable element insertions. Genetics. 2014;196(2):523-38.

29. Rech GE, Bogaerts-Márquez M, Barrón MG, Merenciano M, Villanueva-Cañas JL, Horváth V, et al. Stress response, behavior, and development are shaped by transposable element-induced mutations in Drosophila. PLoS Genet. 2019;15(2):e1007900.

30. Bourgeois Y, Ruggiero RP, Hariyani I, Boissinot S. Disentangling the determinants of transposable elements dynamics in vertebrate genomes using empirical evidences and simulations. PLoS Genet. 2020;16(10):e1009082.

31. Mérel V, Gibert P, Buch I, Rodriguez Rada V, Estoup A, Gautier M, et al. The Worldwide Invasion of Drosophila suzukii Is Accompanied by a Large Increase of Transposable Element Load and a Small Number of Putatively Adaptive Insertions. Mol. Biol. Evol. 2021;38(10):4252-4267.

32. Stritt C, Gordon SP, Wicker T, Vogel JP, Roulin AC. Recent activity in expanding populations and purifying selection have shaped transposable element landscapes across natural accessions of the mediterranean grass Brachypodium distachyon. Genome Biol. Evol. 2018;10(1):304-318

33. Charlesworth B. Transposable elements in natural populations with a mixture of selected and neutral insertion sites. Genet. Res. 1991;57(2):127-134.

34. Charlesworth B, Langley CH, Sniegowski PD. Transposable element distributions in Drosophila. Genetics. 1997;147(4):1993-1995.

35. Charlesworth B, Charlesworth D. The population dynamics of transposable elements. Genet. Res. 1983;42(1):1-27.

36. Charlesworth B. Background selection and patterns of genetic diversity in Drosophila melanogaster. Genet. Res. 1996;68(2):131-149.

37. International Brachypodium Initiative. Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature. 2010;463:763-768.

38. Stritt C, Gimmi EL, Wyler M, Bakali AH, Skalska A, Hasterok R, et al. Migration without interbreeding: Evolutionary history of a highly selfing Mediterranean grass inferred from whole genomes. Mol. Ecol. 2022;31(1):70-85.

39. Minadakis N, Williams H, Horvath R, Cakovic D, Stritt C, Thieme M, et al. The demographic history of the wild crop relative Brachypodium distachyon is shaped by distinct past and present ecological niches. bioRxiv. 2023. https://doi.org/10.1101/2023.06.01.543285

40. Bourgeois Y, Stritt C, Walser JC, Gordon SP, Vogel JP, Roulin AC. Genome-wide scans of selection highlight the impact of biotic and abiotic constraints in natural populations of the model grass Brachypodium distachyon. Plant J. 2018;96(2):438-451.

41. Wyler M, Stritt C, Walser JC, Baroux C, Roulin AC. Impact of transposable elements on methylation and gene expression across natural accessions of Brachypodium distachyon. Genome Biol. Evol. 2020;12(11):1994-2001.

42. Kimura M, Ohta T. The age of a neutral mutant persisting in a finite population. Genetics. 1973;75(1):199–212.

43. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4:e72.

44. Günther T, Coop G. Robust identification of local adaptation from allele frequencies. Genetics. 2013;195(1):205-20.

45. Olazcuaga L, Loiseau A, Parrinello H, Paris M, Fraimout A, Guedot C, et al. A whole-genome scan for association with invasion success in the fruit fly Drosophila suzukii using contrasts of allele frequencies corrected for population structure. Mol. Biol. Evol. 2020;37(8):2369-2385.

46. Garcia Guerreiro MP. What makes transposable elements move in the Drosophila genome. Heredity. 2012;108(5):461–468.

47. Belyayev A. Bursts of transposable elements as an evolutionary driving force. J. Evol. Biol. 2014;27(12):2573–2584.

48. Horvath R, Menon M, Stitzer M, Ross-Ibarra J. Controlling for variable transposition rate with an age-adjusted site frequency spectrum. Genome Biol. Evol. 2022;14(2):evac016.

49. Wright SI, Agrawal N, Bureau TE. Effects of recombination rate and gene density on transposable element distributions in Arabidopsis thaliana. Genome Res. 2003;13(8):1897-1903.

50. Horvath R, Slotte T. The Role of small RNA-based epigenetic silencing for purifying selection on transposable elements in Capsella grandiflora. Genome Biol. and Evol. 2017;9(10):2911-2920.

51. Raissig MT, Woods DP. The wild grass Brachypodium distachyon as a developmental model system. Curr. Top. Dev. Biol. 2022;147:33-71.

52. Hasterok R, Catalan P, Hazen SP, Roulin AC, Vogel JP, Wang K, et al. Brachypodium: 20 years as a grass biology model system; the way forward? Trends Plant Sci. 2022;27(10):1002-1016.

53. Casacuberta E, González J. The impact of transposable elements in environmental adaptation. Mol. Ecol. 2013;22(6):1503-1517.

54. Bourgeois Y, Boissinot S. On the population dynamics of junk: a review on the population genomics of transposable elements. Genes. 2019;10(419):1-23.

55. Langmüller AM, Nolte V, Dolezal M, Schlötterer C. The genomic distribution of transposable elements is driven by spatially variable purifying selection. Nucleic Acids Res. 2023;1-11.

56. Eyre-Walker A, Woolfit M, Phelps T. The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics. 2006;173(2):891-900.

57. Lockton S, Gaut BS. The evolution of transposable elements in natural populations of self-fertilizing Arabidopsis thaliana and its outcrossing relative Arabidopsis lyrata. BMC Evol. Biol. 2010;10(10):1-11.

58. González J, Macpherson JM, Messer PW, Petrov DA. Inferring the strength of selection in drosophila under complex demographic models. Mol. Biol. Evol. 2009;26(3):513-526.

59. Xue AT, Ruggiero RP, Hickerson MJ, Boissinot S. Differential effect of selection against LINE retrotransposons among vertebrates inferred from whole-genome data and demographic modeling. Genome Biol. Evol. 2018;10(5):1265-1281.

60. Wang X, Weigel D, Smith LM. Transposon variants and their effects on gene expression in Arabidopsis. PLoS Genet. 2013;9(2):e1003255.

61. McCue AD, Nuthikattu S, Slotkin RK. Genome-wide identification of genes regulated in trans by transposable element small interfering RNAs. RNA Biol. 2013;10(8):1379-1395.

62. McCue AD, Nuthikattu S, Reeder SH, Slotkin RK. Gene expression and stress response mediated by the epigenetic regulation of a transposable element small RNA. PLoS Genet. 2012.8(2):e1002474.

63. McCue AD, Slotkin RK. Transposable element small RNAs as regulators of gene expression. Trends Genet. 2012;28(12):616-623.

64. Cho J. Transposon-derived non-coding RNAs and their function in plants. Front. Plant Sci. 2018;9(600):1-6.

65. Wyler M, Keller B, Roulin AC. Potential impact of TE-derived sRNA on gene regulation in the grass Brachypodium distachyon. bioRxiv. 2022. https://www.biorxiv.org/content/10.1101/2022.04.05.487121v1

66. Wicker T, Sabot F, Hua-Van A., Bennetzen JL, Capy P, Chalhoub B et al. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 2007;8 973-982.

67. Skalska A, Stritt C, Wyler M, Williams HW, Vickers M, Han J, et al. Genetic and methylome variation in turkish Brachypodium distachyon accessions differentiate two geographically distinct subpopulations. Int. J. Mol. Sci. 2020;21(18):6700.

68. Gordon SP, Contreras-Moreira B, Levy JJ, Djamei A, Czedik-Eysenberg A, Tartaglio VS, et al. Gradual polyploid genome evolution revealed by pan-genomic analysis of Brachypodium hybridum and its diploid progenitors. Nat. Commun. 2020;11:3670.

69. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-2120.

70. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie2. Nat. Methods. 2012;9:357-359.

71. Faust GG, Hall IM. YAHA: fast and flexible long-read alignment with optimal breakpoint detection. Bioinformatics. 2012;28(19):2417-2424.

72. Stuart T, Eichten SR, Cahn J, Karpievitch Y, Borevitz JO, Lister R. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. eLife. 2016;5:e20777.

73. Stritt C, Wyler M, Gimmi EL, Pippel M, Roulin AC. Diversity, dynamics and effects of long terminal repeat retrotransposons in the model grass Brachypodium distachyon. New Phytol. 2020;227(6):1736-1748.

74. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next- generation DNA sequencing data. Genome Res. 2010;20(9):1297-1303.

75. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2017. https://www.biorxiv.org/content/10.1101/201178v3

76. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6(2):80-92.

77. Albers PK, McVean G. Dating genomic variants and shared ancestry in population-scale sequencing data. PLoS Biol. 2020;18(1):e3000586.

78. Huo N, Garvin DF, You FM, McMahon S, Luo MC, Gu YQ, et al. Comparison of a high-density genetic linkage map to genome features in the model grass Brachypodium distachyon. Theor. Appl. Genet. 2011;123:455-464.

79. Dray S, Dufour AB. The ade4 Package: Implementing the duality diagram for ecologists. J. Stat. Softw. 2007;22(4):1-20.

80. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2021. https://www.R-project.org/ Accessed 9 August 2023.

81. Massicotte P, South A. rnaturalearth: World map data from natural earth. R package version 0.3.3.9000. 2023. https://docs.ropensci.org/rnaturalearth/. Accessed 9 August 2023.

82. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419:832-837.

83. Gautier M, Klassmann A, Vitalis R. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. Mol. Ecol. Resour. 2017;17(1):78-90.

84. Gautier M, Vitalis R. rehh: An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics. 2012;28(8):1176-1177.

85. Fox J, Weisberg S. An R companion to applied regression. Third Edition. Sage. Thousand Oaks CA. 2019. https://socialsciences.mcmaster.ca/jfox/Books/Companion/. Accessed 9 August 2023.

86. Gautier M. Genome-wide scan for adaptive divergence and association with population-specific covariates. Genetics. 2015;201(4):1555-1579.

87. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 2012;44(7):821-824.

88. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. 1995;57(1):289-300.

89. Haller BC, Messer PW. Evolutionary modeling in SLiM 3 for beginners. Mol. Biol. Evol. 2019;36(5):1101–1109.

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