# Disentangling the determinants of transposable elements dynamics in vertebrate genomes using empirical evidences and simulations

Yann Bourgeois [1,2], Robert Ruggiero [2,3], Imtiyaz Hariyani [1], Stéphane Boissinot [1].

[1] School of Biological Sciences, University of Portsmouth, Portsmouth PO1 2DY, UK

[2] New York University Abu Dhabi, Saadiyat Island Campus, Abu Dhabi, United Arab Emirates

[3] Department of Biology, Southeast Missouri State University, Cape Girardeau, MO, 63701, USA

E-mail addresses:

Yann Bourgeois: yann.bourgeois@port.ac.uk

Robert Ruggiero: rruggiero@semo.edu

Imtiyaz Hariyani: imtiyaz.hariyani@nyu.edu

Stéphane Boissinot: stephane.boissinot@nyu.edu

Corresponding authors: Yann Bourgeois and Stéphane Boissinot.

# Abstract

**Background** - The interactions between transposable elements (TEs) and their hosts constitute one of the most profound co-evolutionary processes found in nature. The population dynamics of TEs depends on factors specific to each TE families, such as the rate of transposition and insertional preference, the demographic history of the host and the genomic landscape. How these factors interact has yet to be investigated holistically. Here we are addressing this question in the green anole (*Anolis carolinensis*) whose genome contains an extraordinary diversity of TEs (including non-LTR retrotransposons, SINEs, LTR-retrotransposons and DNA transposons).

**Results** - We observe a positive correlation between recombination rate and TEs frequencies and densities for LINEs, SINEs and DNA transposons. For these elements, there was a clear impact of demography on TE frequency and abundance, with a loss of polymorphic elements and skewed frequency spectra in recently expanded populations. On the other hand, some LTR-retrotransposons displayed patterns consistent with a very recent phase of intense amplification. To determine how demography, genomic features and intrinsic properties of TEs interact we ran simulations using SLiM3. We determined that i) short TE insertions are not strongly counter-selected, but long ones are, ii) neutral demographic processes, linked selection and preferential insertion may explain positive correlations between average TE frequency and recombination, iii) TE insertions are unlikely to have been massively recruited in recent adaptation..

**Conclusions** – We demonstrate that deterministic and stochastic processes have different effects on categories of TEs and that a combination of empirical analyses and simulations can disentangle the effects of these processes.

## Keywords

Transposable elements, LINE, SINE, DNA-transposons, LTR-retrotransposons, *Anolis carolinensis*, forward-in-time simulations, linked selection, population genomics, selection.

## Background

Transposable elements (TEs) are among the genomic features that display the most variation across the living world. The nature of the interactions between these genomic 'parasites' and their hosts has likely played a considerable role in determining the size, structure and function of eukaryotic genomes [1–3]. From the perspective of TEs, genomes can be seen as an ecosystem with distinct niches. Borrowing from community ecology concepts [4,5], variation in TE composition and diversity along the genome may be due to competition for resources between clades or constraints linked to changes in environmental conditions (niche-partitioning). An alternative model would posit that TE diversity be driven by stochastic events of expansion and drift that are independent of intrinsic TE properties such as selection or transposition (neutral theory). Within a given host species, these processes can be studied through the prism of population genetics, a field that conceptually inspired the study of ecological communities. Processes linked to niche-partitioning such as varying selection against new insertions [6], variability in the use of cellular machinery and access to chromatin by different TE clades [7,8], or domestication of elements [9], may shape TEs diversity in predictable ways. On the other hand, stochastic processes at the level of individual elements, but also demography at the scale of the host [10–12], may be sufficient to explain variation in the TE landscape [4]. In addition, stochastic processes may not be constant along the genome. For example, recent investigations have highlighted the importance of recombination rates in shaping genomic diversity, due to the

64 effects of selection at linked sites. Because of Hill-Robertson interference and hitch-hiking,
65 regions near a selected site see their genetic diversity drop, an effect that increases in regions of
66 low recombination [13].

67 In vertebrates, most of the knowledge on the micro-evolutionary dynamics of TEs is provided by
68 studies on humans [6]. It seems clear that mechanisms such as drift, selection and migration may
69 play an important role in shaping TEs abundance and frequencies (e.g. [10]). In addition, TEs
70 can insert within regulatory sequences and coding regions, and have a strong potential to reduce
71 fitness. It is therefore likely that they are under purifying selection, which should leave specific
72 signatures such as allele frequency spectra skewed towards rare variants in TEs compared to
73 near-neutral markers such as SNPs [14]. In human, purifying selection acting against long TEs
74 has been demonstrated and this pattern was explained by the greater ability of long elements to
75 mediate deleterious ectopic recombination [15]. While the human model has provided deep
76 insights about the dynamic of LINEs in mammals, it provides only a partial picture of the
77 dynamics of TEs as a whole, given the absence of recent activity of other categories of TEs, such
78 as DNA transposons, in the human genome. In fact, mammalian genomes are unique among
79 vertebrates. They are typically dominated by a single category of autonomous element, *L1*, and
80 related non-autonomous elements (e.g. *Alu* in primates).

81 Non-mammalian vertebrates display a much larger TE diversity, and often include both class I
82 elements (i.e. elements that use an RNA intermediate in their life cycle) and class II elements
83 (i.e. elements that don't use an RNA intermediate). Class I includes LTR-retrotransposons, non-
84 LTR retrotransposons (*i.e.* LINEs and *Penelope*) and their non-autonomous counterparts
85 (SINEs). Class II includes a wide diversity of elements including the widespread DNA
86 transposons. Since TEs vary in their mode of transposition, length, and structure, it is likely that

the effect they have on host fitness and how they are in turn affected by host-specific response will differ. A potentially fruitful approach to this question would be to apply the conceptual and practical tools of population genetics in a model harboring a wide diversity of active TEs. This would facilitate direct comparisons between TE categories while removing the confounding effects of host demography since all elements within the same genome share the same demographic history. The growing availability of whole-genome resequencing data, as well as the development of new computational tools, has revived the interest of the evolutionary genomics community for the analysis of TE polymorphisms within species [16,17].

Whether TEs constitute a substrate for adaptation is another area of interest. Since TEs can lead to substantial regulatory and structural variation, they may constitute targets for fast adaptation and be domesticated by the host's genome [18]. Several possible cases have now been identified at short evolutionary scales, such as the involvement of a TE insertion in industrial melanism trait in peppered moth [19], or the association between some TEs and adaptation to temperate environment or pesticides [9,20] in *Drosophila*. Identifying candidate TEs (and more generally genomic regions) for positive selection is still challenging, and requires stringent filters to keep the number of false positives at a minimum. Combining genome scans obtained from SNP data with a screening of TEs displaying strong difference in frequencies across populations should, fulfill this goal [16,21].

In this study we investigate TE variation in the green anole (*Anolis carolinensis*), which is a particularly relevant model since it is extremely diverse in terms of TE content. Its genome contains four main TE categories, each represented by multiple clades of elements: non-LTR autonomous retrotransposons (nLTR-RT; including the *L1*, *CR1*, *L2* and *Penelope* clades), SINEs, LTR-retrotransposons (LTR-RT; including the *BEL*, *Copia*, *Gypsy* and *Dirs* clade), and

DNA transposons (including *hAT*, *hobo*, *Tc1/Mariner* and *helitrons* clades). There is preliminary evidence that TEs may have been involved in adaptation in this clade, for example by inserting in the *Hox* genes cluster [22]. Previous studies have investigated patterns of genetic structure and past history: the ancestor of the green anole originally colonized Florida from Cuba between 6 and 12 million years ago [23]. A first step of divergence occurred in Florida between 3 and 2 mya (Sup. Figure 1) [24], producing three distinct genetic clusters in Florida, the North-Eastern Florida clade (NEF), the North-Western Florida clade (NWF) and the South Florida clade (SF), the latter being the basal one. The ancestral population of lizards now living in temperate territories diverged from the NEF clade approximately 1 Mya. This divergence was followed by expansion northwards from Florida to the remaining South-Eastern USA, across the Gulf Coastal Plain over the last 100,000-300,000 years [25,26]. This led to the emergence of the two current northern clades, Gulf-Atlantic (GA) and Carolinas (CA). A key aspect of these studies is that they revealed large effective population sizes in all clusters, which should increase the efficiency of selection on TEs and render it easier to detect. In addition, the broad set of environmental conditions encountered by the green anole should provide opportunities for recruitment of TEs by positive selection. At last, genetic diversity is highly variable along the green anole genome, reflecting the joined effects of heterogeneous recombination rates and linked selection [26].

We take advantage of previous studies that investigated the recombination and diversity landscape along the genome to assess i) how does diversity and genomic repartition vary across different TE clades; ii) if direct selection against TE insertions is detectable; iii) how the interaction between demography, counter-selection and linked selection may impact TE frequencies and local abundance; iv) whether there is any clear evidence for positive selection acting on TEs.

# Results

*Description of polymorphic insertions*

A total of 339,149 polymorphic TE insertions were recovered from resequencing data obtained from 28 anoles, including the five genetic clusters identified in previous studies [25,26]. The most abundant category of polymorphic TE found in our dataset consisted in DNA transposons (N=132,370), followed by nLTR-RTs (N=97,586), LTR-RTs (N=78,472), and SINEs (N=30,721). At a finer taxonomic scale, we mostly identified elements belonging to the *CR1*, *L2*, *L1* and *Penelope* clades for nLTR-RTs, *Gypsy* and *DIRS* for LTR-RT, and *Hobo*, *Tc1/Mariner*, *hAT* and *Helitron* for DNA transposons (Table 1). Elements such as *R4*, *RTEX*, *RTE-BovB*, *Vingi* or *Neptune* were rare and mostly fixed (Table 1), probably due to their older age. The same was observed for ancient repeats, classified as *Eulor*, *MER*, *UCON* or *REP* for DNA transposons.

*Diversity within individuals and genetic clusters*

We first examined the possible impact of demography on TEs diversity and abundance. In each individual, we assessed whether heterozygous insertions were found in other green anoles or outgroups. We focused on shared heterozygosity at the individual level to better visualize intra- and inter-individual diversity (Figure 1). Singletons are more likely to be of recent origin, while heterozygous TEs shared between multiple individuals should be older, which may give information about the past and current dynamic of polymorphic elements. Given the low homoplasy of TE insertions, elements shared with the two outgroups were almost certainly found in the common ancestor, and may highlight how past demography impacted individual TE landscapes. An examination of the repartition of polymorphic insertions across individuals

showed a similar pattern across nLTR-RTs, SINEs and DNA transposons. On average, more heterozygous TEs were observed in individuals from the Floridian populations, which became established about two million years and remained stable and large (effective population size, $N_e \sim$ 1 million) since colonization from Cuba. For these three categories, heterozygous TEs (private or shared) are more abundant in the outgroups (which correspond to the 2 Cuban anole species) and in the Floridian populations but become rarer in populations that expanded out of Florida, which is consistent with the loss of genetic variation experienced in those more recently established population. In addition, for the most abundant clades, there were always more fixed insertions in GA and CA than in Floridian groups with similar sample sizes (Table 1). These patterns are consistent with drift leading to faster fixation or elimination of TEs. For LINEs and SINEs, *L2* and *SINE2* elements displayed a large number of heterozygous TEs found only in the two outgroups, but also displayed a large proportion of heterozygous sites shared between *A. carolinensis* and either *A. porcatus* or *A. allisoni*. The same was observed for the DNA transposons *Tc1/Mariner* and *hAT*. This suggests that a substantial proportion of elements inserted before the split between these species, and that drift may have led to gradual loss of shared elements. *Hobo*, *Helitron*, *SINE1*, *L1*, *CR1* and *Penelope* maintained a relatively high proportion of private insertions in individuals from Florida, less shared heterozygous sites and similar number of heterozygous insertions when compared to the outgroups. This is consistent with elements at lower frequencies in the common ancestor, either because of stronger purifying selection or more recent transposition activity, leading to less shared variation between present genetic groups and species.

On the other hand, for LTR-RTs, elements from the *Gypsy* and *BEL* clades displayed a large number of private insertions in the green anole, with many insertions found only in a single

178 individual, and no clear pattern of reduced abundance in bottlenecked populations from the
179 Northern cluster. This can be interpreted as a signature of recent and active transposition in the
180 green anole lineage. This was especially clear for *Gypsy* elements, suggesting a burst of
181 transposition following colonization from Cuba.

182 A visual inspection of allele frequency spectra (AFS) confirmed the effect of demography on
183 TEs (Figure 2, Sup. Figures 2 to 5): for DNA transposons, LINEs and SINEs, spectra were
184 skewed toward singletons in genetic clusters with large population sizes (SF, NEF, NWF), while
185 this trend was less pronounced in clusters having been through a recent bottleneck (GA and CA).
186 This is consistent with the excess of frequent alleles expected in the case of population
187 contraction. These differences were however less clear for LTR-RTs, with spectra strongly
188 skewed towards singletons in all populations. While AFS were clearly U-shaped for the other
189 three types of elements, almost no LTR-RT insertion was found at very high frequencies. Such a
190 pattern is consistent with recent activity and purifying selection preventing insertions to reach
191 high frequencies. There were also differences within different types of elements. For non-LTR
192 retrotransposons, elements such as *Poseidon* or *RTEBovB* were mostly found at high frequencies
193 (Table 1). Elements such as *RTE1*, *L1*, *CR1* and *Penelope* displayed a stronger skew towards
194 singletons than *L2*. In SINEs, *SINE1* had more singletons, while other elements were more
195 frequent. For DNA transposons, the skew towards singletons was strongly pronounced for *Hobo*
196 and *Helitron*, and very few fixed insertion were found (Table 1), suggesting either stronger
197 purifying selection or a recent increase in transposition rate.

198 *Correlation of TE density with recombination and differentiation reveals discordant patterns*

199 Studies focusing on SNPs have revealed that regions of low recombination display lower
200 diversity and stronger differentiation between populations due to the effects of linked selection

[27,28]. We tested whether TEs were also impacted by this phenomenon by examining correlations between TE frequencies with recombination rates, derived allele frequencies, and absolute ($d_{XY}$) and relative ($F_{ST}$) measures of differentiation computed over SNP data in non-overlapping 1Mb windows (Figure 3). We focused on the six main autosomes of the green anole. If linked selection shapes genomic diversity along the genome, there should be 1) positive correlations between diversity indices (average derived allele frequency, $d_{XY}$) and recombination, 2) negative correlations between differentiation measures ($F_{ST}$) and recombination, 3) consistency in genomic regions displaying high or low values for $F_{ST}$ or $d_{XY}$ across all pairwise comparisons. This is in line with our observations, with mostly positive correlations between recombination rate, diversity and absolute divergence for all pairwise comparisons between the five genetic clusters (Figure 3). Pairwise relative measures of differentiation ($F_{ST}$) were negatively correlated with recombination rate, $d_{XY}$, and derived allele frequencies, which is consistent with a role of linked selection reducing diversity in regions of low recombination across all genetic clusters. Indices of differentiation comparing CA or GA with other populations were less correlated with indices of differentiation estimated between pairs of clusters from Florida, suggesting a role for recent expansion in blurring the expected correlations.

Linked selection should have a similar effect on TEs as on SNPs. The stronger lineage sorting observed in regions of low recombination should lead to a lower number of polymorphic TEs in regions of low recombination. On the other hand, purifying selection against elements through ectopic recombination should slow the accumulation of TEs in regions of high recombination compared to regions of low recombination. This should result in decreasing densities of both polymorphic and fixed elements as recombination increases. We examined densities of polymorphic and fixed TEs across four main categories of TEs (Figure 3). TE densities were

positively correlated with recombination rate, diversity and relative measures of differentiation for SINEs and DNA transposons. Correlations were weaker for LINEs, and almost absent for LTR-RTs. The density of fixed LTR-RTs even followed an opposite pattern, with more fixed insertions in regions of low recombination and high $F_{ST}$. For fixed LINEs, correlations were weak or absent. This suggests that purifying selection against LTR-RT and to some extent LINEs may explain their local abundance and diversity.

The higher abundance of some TE categories in regions of low recombination was not explained by a higher density of functional elements that could increase the deleterious effects of LTRs or LINEs (Sup. Figure 6). Exon density was positively correlated with recombination rate (Spearman's *rho*=0.15; p-value=$9.1.10^{-7}$), which suggests that regions of high recombination may also be more frequently transcribed, and are therefore more often in an open chromatin state.

TE densities were positively correlated with each other across genetic groups for all TEs, with correlations strengthening as comparisons involved more closely related pairs of populations. This effect is expected due to a longer shared history for related clades.

*Comparison of TE diversity across TE clades in a demographically stable genetic cluster*

We assessed whether purifying selection had a direct impact against TEs by examining average TE frequencies in 1Mb windows and comparing it to the frequencies of derived SNPs. To obtain a more accurate estimate of frequency, we focused on the population with the largest sample size and with a historically stable effective population size, NEF [26]. We also examined diversity at the clade level to highlight specific dynamics. We excluded TE clades with less than 5000 elements (Table 1), and merged SINEs that were not *SINE2* together to provide a comparison

within the category. We examined these statistics for SNPs and the main clades within the four main TE categories (Figure 4). Average TE frequencies were lower for LTR-RTs and *Dirs* than for SNPs and the differences were statistically significant (frequencies of 0.10, 0.15, 0.13, 0.17, and 0.26 for *BEL*, *Dirs*, *Gypsy*, unclassified LTRs and derived SNPs respectively; paired-samples Wilcoxon tests, all $P<2.2.10^{-16}$) across all clades. This is consistent with either purifying selection against these elements, and/or their younger age. The same was observed for *CR1*, *L1* and *Penelope* (frequencies of 0.19, 0.17 and 0.16), but not *L2* (frequency of 0.30), for which the average frequencies were significantly higher than derived SNPs (all $P<1.10^{-11}$). The average frequency of SINEs other than *SINE2* was 0.28, not substantially different from SNPs ($P=0.88$), and was even higher for SINE2 (0.33, $P=5.5.10^{-12}$). For DNA transposons, *Hobo*, *Helitron*, and to a lesser extent, *Tc1/Mariner* displayed lower frequencies than SNPs (0.13, 0.12 and 0.22 respectively, all $P<2.2.10^{-16}$). On the other hand, *hAT* displayed an average frequency of 0.39, substantially higher than SNPs ($P<2.2.10^{-16}$). Elements at a higher frequency than derived SNPs are likely ancient, and their high frequency is best explained by a non-equilibrium dynamic, with a lack of recent transposition resulting in a depletion in the lower frequencies of the allele frequency spectrum.

Ectopic recombination should lead to stronger purifying selection against TEs in regions of high recombination. This should result in reduced frequency and abundance of elements in regions of high recombination. To test whether TEs from different clades followed this predicted pattern, we assessed whether their average frequency and their density varied with the recombination rate (Figures 5, 6, 7, Table 2). For all LTR-RTs, we observed negative correlations between recombination rate and average frequency (Figure 5). Weak, negative correlations were also observed when replacing frequency by the density of fixed insertions (Figure 7), the strongest

trend being observed with *Gypsy*. For the latter, negative correlation between the density of polymorphic sites and recombination was observed (Figure 6). This pattern is clearly consistent with a stronger deleterious effect of these elements in regions of high recombination. Correlations were however weak (*BEL*, unclassified LTR-RT) for other LTR-RTs. They were significantly positive for *Dirs*. SINEs and DNA transposons (except *Hobo*) showed positive correlations between all three summary statistics and recombination rate, which may be partly explained by linked selection and a lack of strong purifying selection. For *Hobo*, the only significant correlation was found between recombination rate and the density of polymorphic sites, probably because of the rather low number of fixed insertions, obscuring correlations.

For nLTR-RTs, we did not observe significant correlations between recombination and TE frequency or the density of fixed insertions, except for *CR1* (Figure 5 and 7; Table 2). Positive correlations were however observed for *Penelope*, *CR1* and *L2* when examining the density of polymorphic sites. We however suspect that this lack of clear correlation may be due to variation in the strength of purifying selection among nLTR-RTs. Previous studies in vertebrates and *Drosophila*, [6,12,29,30] have shown that the effects of LINEs on fitness may be correlated with their length. This is due to the fact that the odds of homologous recombination rise with the length of homologous fragments [31]. Truncation in LINEs occurs at the 5' end of elements, which makes MELT estimates of their length accurate since it detects TEs based on reads mapping the ends of the insertion. To assess whether purifying selection acted more strongly on longer elements, we examined the correlation between recombination rates and the average length of fixed and polymorphic LINEs in 1Mb windows (Figure 8), and observed a clear negative correlation between these two statistics (Spearman's *rho*=-0.16, -0.26, -0.21 for *CR1*, *L1* and *L2* respectively, $P<5.10\text{-}7$). LINEs that were fixed in the NEF population were also

shorter than the polymorphic ones. We then focused on short LINEs (<20% of the maximum length of their respective clade) to assess whether they were also erased from regions of high recombination. We used 10Mb windows to increase the number of insertions and avoid losing too much information. We then reexamined the correlations between recombination rate and our three summary statistics (Figure 8). We found a positive correlation between frequency and recombination rate for short *CR1* and short *L2*. All short elements showed positive correlations between recombination and the density of polymorphic elements, while no clear correlation was observed for the density of fixed elements (Table 2, Figure 8). For long LINEs (>30% max length), we observed strong negative correlations between TE frequency, the density of fixed insertions, and recombination. The same was observed with the density of polymorphic insertions, except for *L2* (Table 2). These results suggest that weak correlations observed at the scale of the whole clade are explained by non-uniform, length-dependent selection against the elements. Short LINEs are therefore more likely under the influence of linked selection, while long LINEs display patterns that are closer to observations in LTR-RT, suggesting a stronger influence of purifying selection.

*Simulations clarify the relative impact of purifying selection, linked selection and bursts of transposition on autonomous retrotransposons diversity.*

Our results reveal many combinations of correlations between TE diversity and recombination rate. To clarify and illustrate the conditions under which these combinations arise, we built a simple model of retrotransposon evolution in the forward-in-time simulator SLiM3 [32]. We simulated a 4Mb fragment with two recombination rates and negative selection on 10% of the SNPs. Recombination was high on the first and last Mb, and low for the 2Mb in the middle of the fragment. Two categories of TEs were simulated, “short” TEs that were weakly deleterious

(Figure 9, blue boxplots), and "long" TEs (red boxplots) that were more deleterious in regions of high recombination. We then examined the same three summary statistics than earlier: the average frequency of polymorphic insertions, the density of polymorphic insertions, and the density of fixed insertions (Figure 9). Short TEs showed higher average frequencies in regions of high recombination when transposition was kept constant, a pattern consistent with expectations if linked selection increases lineage sorting in regions of low recombination (Figure 9, panels A). This trend was however reversed if transposition occurred as a single ancient burst (panels B). In that case, average TE frequencies were also higher, due to the older age of insertions. Moreover, because linked selection leads to faster lineage sorting in regions of low recombination, polymorphic insertions that survive after the burst reach higher frequencies, explaining the observed correlation. On the other hand, long TEs displayed lower average frequencies in regions of high recombination, due to their stronger deleterious effects, whether transposition was kept constant or not. Models including preference for TEs to insert in regions of high recombination (panels C and D) produced very similar results for this summary statistic.

The density of polymorphic insertions was higher in regions of high recombination for short TEs across all simulations, but the difference was even more pronounced when preference for regions of high recombination was added to the model (panels C and D). The trend was reversed for long TEs (panel A), but including preference for high recombination again led to a positive correlation between recombination rate and the summary statistic (panel C), since more insertions could replace the ones erased by selection. Models where a burst of transposition occurred gave the same trends (panels B and D), although preference for high recombination did not fully reverse the correlation (panel D).

337 The density of fixed insertions was lower in regions of high recombination than in regions of low
338 recombination in models with no preference (panels A and B). This result was observed for both
339 short and long TEs, although the effect was enhanced for long TEs due to their stronger
340 deleterious effect in regions of high recombination. In models where preferential insertion in
341 regions of high recombination was added however, a positive correlation with recombination rate
342 was observed under a constant transposition rate, and differences were less marked in the case of
343 a transposition burst (panels C and D).
344
345 We compared these trends with our actual observations (summarized in Table 2), which are
346 consistent with either strong purifying selection against new insertions through ectopic
347 recombination or predominant effects of linked selection. For short LINEs, and particularly *CR1*,
348 we observed correlations consistent with linked selection, similar to the simulations for short
349 elements highlighted in panels A of Figure 9. A possible effect of preferential insertion may
350 explain the weak correlations observed between the density of fixed elements and recombination
351 for *L1* and *L2* (panels C). For long LINEs, correlations for the three statistics were consistent
352 with simulations obtained for long elements with no preferential insertion (Figure 9, panels A
353 and B). The same was observed for *Gypsy* elements. For long *L2*, the lack of strong correlation
354 between the density of polymorphic elements and recombination may reflect a situation closer
355 from the simulations presented in panels C and D, with some effect of preferential insertion and
356 past burst of transposition. The same reasoning may be applied to LTR-RT elements such as
357 *BEL*. For *Dirs*, observations matched expectations for long elements in simulations shown in
358 panel C, suggesting both selection against ectopic recombination and preferential insertion in
359 regions of high recombination. For SINEs and *Tc1/Mariner*, the observed correlations clearly

matched simulations for short elements including linked selection and preferential insertion (panel C). This scenario is also likely for *Hobo* and *Helitron*, although their weak frequencies obscures correlations between average allele frequencies, density of fixed sites, and recombination. The same issue makes any interpretation of patterns observed for *Penelope* difficult.

Given the high frequencies observed for *SINE2* and particularly *hAT*, it is possible that lower transposition rates in more recent times have led to a situation intermediate between our constant transposition and ancient burst scenarios for short elements (panels C and D respectively), weakening correlations between average frequencies and recombination.

*Are TEs targeted by strong and recent positive selection in northern genetic clusters?*

Because TEs can cause major regulatory changes, they may be recruited during local adaptation, especially in species encompassing a broad range of environmental conditions. If TE insertions were recruited during the recent colonization of northern environments, they should display a strong change in frequencies between the Floridian source and northern populations, and fall in regions displaying signatures of positive selection that can be detected through the use of SNP data. We first scanned all polymorphic insertions to identify a set of candidate TEs displaying high frequencies in Northern clusters and low frequency in Florida. We used two statistics to identify TEs that were potentially under positive selection, $X^TX$ and *eBPis*. $X^TX$ is a measure of global differentiation that should be higher for markers displaying variation in allele frequencies that are not consistent with demographic expectations drawn from SNPs. *eBPis* is a complementary statistic that specifically contrasts frequencies between Floridian or Northern clusters. We identified a set of 34 insertions that were in the top 1% for both *eBPis* and $X^TX$

383 statistics and showed a shift of at least 0.5 in their frequency compared to all samples in Southern
384 Florida. We then filtered out insertions that did not fall in a set of candidate windows displaying
385 consistent signals of selection across three different approaches (see Methods). Four insertions
386 passed this last filter (Table 3), three of them overlapping two distinct genes, *Neurexin2* and
387 *TCF-1*.

388

389 ## Discussion

390 *Demography shapes TE diversity across populations*

391 We observed a clear effect of genetic drift on TE diversity across the genetic clusters examined
392 in this study. Past work on green anole demography clearly showed that the GA and CA clusters
393 expanded recently after a bottleneck when populations contracted to reach about 10% of their
394 ancestral sizes [26]. This is associated with a reduction in the total number of polymorphic
395 insertions found in these populations (Figure 1, Table 1), but also in an increase in the number of
396 fixed elements compared to Floridian samples. Across families and clades, there were between 5
397 and 20% more fixed insertions in northern samples than in Florida (Table 1). This is a classical
398 expectation: under a bottleneck, rare mutations frequently go extinct while frequent ones tend to
399 reach fixation, leaving an excess of mutations at intermediate frequencies [33]. Fixation may also
400 be facilitated by relatively less efficient selection due to lower effective population size, reducing
401 $N_e s$. The strong impact of demography on TE abundance and frequencies has also been observed
402 in a broad range of species and TE families, such as SINEs, LINEs, *Ac*-like elements and *Gypsy*
403 in several species of *Arabidopsis* [10,34]. In *Drosophila subobscura*, recent bottlenecks may also
404 explain the unusually high frequencies of *Gypsy* and *bilbo* elements [35].

*Linked selection affects TE frequency, but not TE density*

We obtained intriguing results for SINEs, DNA transposons such as *Tc1/Mariner*, and short LINEs. Under the ectopic recombination hypothesis [29,36], which is usually invoked to explain genome-wide patterns of TE diversity, TEs tend to be removed from regions of high recombination through purifying selection. Such correlations have been commonly observed for several TE families in fruit flies and other vertebrates[6,29,37]. This should lead to negative correlations between recombination and TE diversity or abundance, assuming constant transposition. Instead, we observe a positive correlation between recombination and average frequency and density of polymorphic elements. Such positive correlation between allelic diversity and recombination is however a well-known feature of so-called "linked selection" [13,38]. Haplotypes harboring deleterious mutations tend to be longer in regions of low recombination, and competition between them reduces the efficacy of selection [38]. Similarly, the local reduction in diversity that comes with selective sweeps extends over longer genomic distances in regions of low recombination. Altogether, this leads to an effect similar to a local reduction of effective population sizes in regions of low recombination, reducing diversity and increasing the odds that deleterious alleles reach fixation.

While some work has been done in examining whether Hill-Robertson interference between elements may increase the number of fixed insertions in regions of low recombination [39], there is not any study (to our knowledge) that examined the allele frequency spectrum of TE insertions under linked selection. In addition, the latter study considered only TE insertions and did not incorporate background selection or sweeps on SNPs. Our simulations suggest that linked selection may lead to positive correlations between polymorphic TE frequency and abundance: polymorphic TEs would stochastically reach frequencies of 0 or 1 at a faster rate in regions of

lower recombination. This would therefore lead to a rise in the number of polymorphic TEs and average TE frequencies as recombination increases, but also to a reduction in the number of fixed TEs (as expected in the case of Hill-Robertson interference).

Unlike the ectopic recombination and linked-selection hypotheses, preferential insertion in regions of high recombination and open chromatin does predict a positive correlation between recombination rates and TEs density. This mechanism has been proposed to explain why LINEs and LTR-RT may be more abundant in regions of high recombination in *Ficedula* flycatchers and the zebra finch [7]. It is commonly observed for several retrotransposons in a variety of species [8,40,41]. However, in humans, *L1* may actually not display strong preference for open chromatin and is more constrained by local replication timing [42,43]. In the green anole, LINEs and LTR-RTs do not display strong evidence of preferential insertion in regions of high recombination, which tend to harbor less fixed elements. We note that these families are relatively ancient in birds, having accumulated between 55 and 33 Mya [44], while a substantial proportion of these elements display less than 1% divergence from their consensus in green anoles (see repeat landscape at http://www.repeatmasker.org/species/anoCar.html, last accessed 25/03/2020). It is therefore possible that purifying selection has had more time to remove the most deleterious insertions in birds, increasing the signal of preferential insertion that may be masked in the green anole. Further studies at finer genomic scales will be helpful to precisely quantify how local genomic features impact TE abundance.

Our simulations suggest preferential insertion would probably not produce higher average TE frequencies in regions of high recombination. We interpret this as the fact that preferential insertion is analog to locally higher mutation rates for nucleotides: while this may affect local SNP density along the genome, it should have little effect on the shape of the allele frequency

spectrum under mutation-drift equilibrium (under the assumption of infinite sites which should hold for low mutation or transposition rates [45]).

We therefore propose that SINEs, *Tc1/Mariner* and most short LINEs are under the influence of linked selection and preferentially insert into regions of high recombination, possibly because these are more likely to be associated with an open chromatin state. Indeed, combining these mechanisms in our simulations produced correlations matching our observations for SINEs and *Tc1/Mariner*. The average frequency of these elements was quite close from average derived SNP frequency. It is therefore unlikely that strong purifying selection acts against these elements (Figure 4, 5, 6, 9).

For short LINEs, the negative correlation between recombination rate and the density of fixed elements may reflect a residual effect of stronger purifying selection in regions of high recombination, and/or weaker preference for regions of open chromatin. In the case of short *L1*, we observe a positive correlation between the density of polymorphic elements and recombination rate, but this correlation is weak when examining the density of fixed elements or average frequency. We note however that *L1s* are substantially longer than other LINEs, which limits our power to study short elements.

*Combination of bursts of transposition and linked selection leaves a specific signature*

Sudden bursts of transposition are common in TEs, and have been documented in a variety of species [46–50]. This idiosyncrasy limits direct comparisons between TEs and SNPs, since mutation rates are usually considered constant for the latter. A general prediction is that the average frequency of elements should increase with their age, which is observed in *Drosophila* [37]. Our simulations also suggest that the positive correlation between average TE frequency

and recombination rate observed for weakly deleterious TEs could be weakened and even reversed in the case of a sufficiently old transposition burst. This is due to the fact that the rarest elements have already been eliminated through drift, and the effects of linked selection lead to a faster accumulation of elements at high frequency in regions of low recombination.

We found that elements such as *hAT* and *L2* had substantially higher average frequencies, even higher than derived SNPs. For these two elements, correlations between their average frequencies and recombination rate were quite weak, even when considering only short *L2* that should be the least deleterious. This could reflect an intermediate situation compared to the extreme scenarios illustrated in Figure 9, such as multiple waves of transposition, or a younger burst than the one modeled, that may obscure correlations by flattening average allele frequencies. Examining the spectrum from more individuals may have the potential to reveal irregular transposition since local peaks in the spectrum should correspond to the age of each burst.

On the other hand, DNA transposons such as *Helitron* and *Hobo* are at very low frequencies, with almost no fixed insertion, but are more abundant in regions of high recombination. This pattern could be explained by a recent burst of transposition associated with weak purifying selection. Whether these elements share the preference of other DNA transposons for regions of high recombination remains difficult to assess due to the lack of fixed insertions.

*Strong purifying selection against Dirs, LTR-RTs and long LINEs.*

There is evidence that strong purifying selection acts on *Dirs*, LTR-RTs and long LINEs: their average frequency is generally lower than the one of derived SNPs. Recent bursts of transposition alone may also be responsible for an excess of young, therefore rare, alleles [51].

While this seems clearly the case for *Gypsy* elements, which display many singletons and seem to be less impacted by recent demography, we also found evidence for lower average TE frequency and density of fixed TEs in regions of high recombination for long LINEs and LTR-RTs. According to our simulations, such a correlation can only be obtained through stronger purifying selection in regions of high recombination, consistent with the ectopic recombination model. For all LTR-RTs (except *Gypsy*) and long *L2*, we observed weak and even positive correlations between recombination rate and the density of polymorphic elements. This may reflect some preference for regions of high recombination compensating the loss of polymorphic elements through selection.

These results suggest that LTRs and long LINEs may be more harmful in regions of high recombination, which are also richer in functional elements. Assuming that our simulations are reasonably close from the actual processes taking place in the green anole, $N_e s$ against these elements is likely high, and possibly higher for elements with very low frequencies such as *Dirs*, *Gypsy* or *BEL*. The length of an element seems to be strongly correlated with its impact on fitness, since short LINEs display a weakening and even a reversal of correlations with recombination rate. These results are consistent with the ectopic recombination hypothesis, since longer elements are more likely to mediate ectopic recombination events [6].

*Strong recent positive selection on TEs is rare*

Recent colonization of northern climates by the green anole may have been an opportunity for domestication of TEs, either through adaptation to the new selective pressures encountered or selection on dispersal promoting colonization of the new environment [52]. We did not find strong evidence that TEs be involved in adaptation in the northern clades. Only a few TEs displayed substantial differences in frequencies between northern and Floridian clusters. We

found in total four elements that are serious candidate for positive selection, falling in introns of *Neurexin2* and *TCF1-220*. *Neurexin2* is involved in the neurotransmitter release [53], while *TCF20-201* is a transcription factor associated with behavioral abnormalities [54,55]. While this suggests a potential impact on the nervous system and behavior, and echoes our findings from a previous study on positive selection in green anoles [52], further investigations are needed to formally validate the causal role of these elements and discard the possibility that they are only linked to a causal variant under selection. Our results contrast with observations in *Drosophila*, where many TEs display steep clines in frequency that match environmental gradients and adaptive phenotypes [20,56,57]. Further investigations are needed to assess whether higher effective population sizes and more compact genome structure in *Drosophila* may explain higher rates of domestication.

There is a growing body of evidence that intrinsic properties of genomes (e.g. overdominance, Hill-Robertson effects, non-equilibrium demography) may lead to spurious signals of selection. We note that we are extremely stringent in our approach, requiring that at least three distinct tests of positive selection give a consistent signal, one of these tests explicitly incorporating demographic history in its implementation. While this could potentially limit our power to detect more subtle signals of positive selection (e.g. soft or partial sweeps), we caution against over interpreting results obtained from a single test, especially when demographic histories are complex. This is not to say that TEs are not more frequently involved over longer timescales: for example, TEs may be involved in speciation and morphological adaptation by shaping the *Hox* genes cluster in anoles [22]. Future studies on larger sample sizes may provide a more refined picture of the role of TEs in local adaptation.

# Conclusions

Using empirical data in a model species harboring a large diversity of active TEs as well as simulations, we investigated the relative impact of selective and non-selective factors on the population dynamics of all the main TE categories active in vertebrates. We tested how the combination of linked selection in the host, direct selection against TEs and changes in transposition rate may explain heterogeneous TE frequency and abundance along the genome. By comparing the diversity of several of the most common TE categories found in vertebrates within the same organism, we clearly demonstrate that the interaction between these processes lead to sometimes drastically different outputs, even under a shared demographic history. It may be possible to disentangle these different processes using information about elements length, genomic location and frequency.

We created a simple model of TE evolution that incorporated variable purifying selection against TEs, bursts of transposition, preferential insertion of TEs in regions of high recombination, and linked selection. While this model was designed as a way to illustrate how different combinations of parameters may impact correlations for the three main statistics examined in this work, this constitutes a template for future, more details studies of TE evolution. For example, SLiM3 allows the incorporation of detailed maps of genomic features, complex demographic histories, multiple modes of selection, or asexual reproduction. This should facilitate the interpretation of TE diversity in species for which a reference genome is available, and improve our understanding in model species for which extensive genomic information exists. Simulated data could be used in an ABC-like approach [12], or to train machine learning algorithms [58]. Such approaches may have the power to directly quantify for each TE clade the strength of

563 purifying selection and how other processes such as linked selection and transposition process
564 may interact.

565

566 # Methods

567 *Sampling and SNPs calling*

568 Liver tissue samples from 27 *Anolis carolinensis* individuals were collected between 2009 and
569 2011 (Tollis et al. 2012), and *A. porcatus* and *A. allisoni* were generously provided by Breda
570 Zimkus at Harvard University. Whole genome sequencing libraries were generated from these
571 samples following the laboratory and bioinformatics procedures already presented in [26] and
572 included as supplementary material. Sequencing depth was comprised between 7.22X and
573 16.74X, with an average depth of 11.45X. SNP data included 74,920,333 variants with less than
574 40% missing data. Sequencing data from this study have been submitted to the Sequencing Read
575 Archive (https://www.ncbi.nlm.nih.gov/sra) under the BioProject designation PRJNA376071.
576 We excluded one individual with low depth of coverage from subsequent analyses due to its
577 large amount of missing data.

578 *Calling TEs*

579 We used the Mobile Element Locator Tool (MELT) to identify polymorphic insertions in the
580 green anole genome [59]. This software performs well in identifying and genotyping
581 polymorphic TEs in resequencing data of low and moderate coverage (5-20X), using TE
582 consensus sequences to identify reads mapping to both the reference genome and the consensus.
583 We followed the same pipeline used in previous studies [11,12], but included several clades of
584 transposable elements covering SINEs, nLTR-RT LTR-RT and DNA transposons, using all

available consensus sequences available on Repbase [60] to call TEs. Note that MELT can estimate the most likely breakpoints, insertion length, and strand for each insertion. We followed the MELT-SPLIT pathway, which consists of four main steps. First, TEs are called for each individual separately (IndivAnalysis). Then, calling is performed at the scale of the whole dataset to improve sensitivity and precision when estimating breakpoints and insertion length (GroupAnalysis). This information is then used to genotype each individual (Genotype), after what a VCF file is produced that lists all polymorphisms (makeVCF). To draw an accurate estimate of TE frequency spectra, we also used MELT-DELETION to identify polymorphic insertions found in the reference but not in all sequenced individuals. We called polymorphic TEs for each clade within the four main categories, using a threshold of 5% with the consensus sequence to attribute an element to a specific clade. The resulting VCF files were then merged for each of the four main categories considered. In case of a possible duplicate call (*i.e.* when two insertions were found at less than 2000bp from each other), only the insertion with the lowest divergence was kept. In case of equal divergence, the element with the highest calling rate was kept. We focused on TEs insertions with no missing data. While we acknowledge that these filters may be quite stringent, they should not have any impact on correlations with intrinsic genomic features and demography.

*Correlations with genomic features and SNPs statistics*

From the MELT output, we extracted information about the frequency of each insertion in each of the five genetic clusters found in the green anole, using the option --counts in VCFTOOLS (v0.1.14) [61]. We also estimated the number of heterozygous sites for each individual using the --012 option in VCFTOOLS. For LINEs, we extracted the length of each insertion using shell

scripts. We counted the number of insertions, and the proportion of private and shared alleles for each clade using R scripts [62].

We also investigated how TE diversity correlated with intrinsic features of the genome such as the recombination rate, and statistics related to demographic processes. We focused on three commonly used statistics to describe TE diversity in each genetic cluster: the density of polymorphic TEs, the density of fixed TEs, and the average frequency of polymorphic TEs. Note that we do not include TEs that are fixed in all 29 samples since our interest in on the most recent population dynamics. We averaged TEs frequencies and densities over 1Mb windows, a length chosen to recover enough TEs even at the clade level, while limiting the effects of linkage disequilibrium and autocorrelation between adjacent windows. Windows with no TEs or found on scaffolds not assigned to any of the six main autosomes were excluded. To estimate average TE frequencies, only windows with at least three polymorphic insertions were used. We also extracted the average effective recombination rate $\rho = 4N_e r$ in the NEF clade estimated by LDHat (v2.2) [63] in a previous study, with $N_e$ the effective population size and $r$ the recombination rate between two adjacent sites (see Sup. Material and [26] for details). This population was chosen since it has the largest sample size and has a large, stable effective population size. This rate was divided by another estimator of the effective population size, the average number of pairwise differences ($\theta_\pi = 4N_e\mu$, $\mu$ being the mutation rate per base pair), to obtain an estimate $r/\mu$ less sensitive to local reductions in effective population sizes due to linked selection. Relative and absolute measures of differentiation such as $d_{XY}$ and $F_{ST}$ were also computed over 1 Mb windows, as well as the average frequency of derived SNPs in green anoles, using the two outgroups *A. porcatus* and *A. allisoni* to determine the derived alleles. These last statistics were obtained using the package POPGENOME (v2.7.5) [64]. Correlograms

summarizing correlations between these summary statistics, TE frequencies, and TE densities for the four main orders were obtained using the R package corrplot. Significance and strength of correlations were assessed using Spearman's rank correlation tests. For plots of correlation, regression lines and their confidence intervals were added to improve visibility with the function geom_smooth in ggplot2 (v3.2.1) [65], using a Gaussian model for TE frequencies and a Poisson model for TE densities (which are counts per window).

*SLiM3 simulations*

In order to clarify how factors such as linked selection, bursts of transposition and preferential insertion of TEs may impact the three statistics examined in this study, we performed simulations using the forward-in-time simulator SLiM (v3.3.2) [32]. We simulated a 4Mb genomic fragment with parameters realistic for green anoles (Sup. Figure 7). We simulated 8 diploid individuals drawn from a stable population with a $N_e$ of one million diploid individuals, similar to the NEF clade (Sup. Figure 1). The mutation rate for nucleotides was set at $2.1.10^{-10}$ mutation/generation/site [24]. To simulate the effects of linked selection, we set the recombination rate at $2.10^{-10}$ /generation on the first and last Mb of the fragment, and at $2.10^{-11}$ /generation in the 2Mb between. These rates encompassed those estimated with LDHat in previous studies [26,52]. Of all new point mutations, 10% were deleterious with $2N_es = -10$ ($s = 5.10^{-6}$). There is not much information about the fitness effects of new mutations in vertebrates in general, and the green anole in particular. However, our estimate seems reasonable given that in mice and humans, about 20-40 % of mutations in conserved regions may have an $s > 1.10^{-3.5}$ [66]. While we acknowledge that positive selection may also play a major role in reducing diversity, we did not include it in these simulations for the sake of simplicity and given the difficulties in properly estimating the proportion of positively selected sites in our dataset. We

assumed that there are 10 TE progenitors in the whole genome that can jump and insert at $P$=1.10$^{-3}$ elements/generation/genome at a constant rate, a value chosen to reflect known transposition rates in vertebrates and which produced a number of TEs close from our empirical observations. This gave a probability of insertion in the 4Mb region of $P$ x 4 / 1780, since the green anole genome is 1.78 Gb long. We also modelled bursts of transposition where $P$ was set 100 times higher, but with transposition occurring only during a lapse of 100,000 years, starting 1,000,000 years ago. Half of the newly generated elements were considered "short" and under weak purifying selection, with $2N_es$ = -0.1. The other half were considered "long", and had a stronger impact on fitness when falling in regions of high recombination ($2N_es$ = -10) than in regions of low recombination ($2N_es$ = -1). The justification for this is that long elements have a higher probability of mediating deleterious ectopic recombination events and those events are more likely to occur in regions of high recombination. To improve the speed of simulations, we modelled a population of size $N_e$ = 1000 diploid individuals, and rescaled all parameters accordingly: mutation, recombination and rates of insertion were multiplied by a factor 1000, and times in generation and selection coefficients divided by the same factor. Simulations were run over 20,000 generations to ensure that mutation-selection-drift balance was achieved for nucleotide mutations.

To account for the potential preference of elements to insert in regions of high recombination, which tend to be gene rich and are often associated with open chromatin [7,67], we also added a preference bias $Q$ which could take the values 0.5 (TEs were as likely to insert in regions of low recombination than in regions of high recombination) or 0.7 (in that case, 70% of TEs jumping into the 4Mb region inserted in regions of high recombination and 30% in regions of low recombination). Note that values for selection coefficients and preferential insertion were chosen

676 to better visualize the trends that we observed across a range of other combinations, and because
677 they produced results close from our empirical observations. The scripts used to simulate these
678 data are available on Github (https://github.com/YannBourgeois/SLIM_simulations_TEs), and
679 can be reused to explore in more details other combinations of parameters.

680 *Overlap with scans for positive selection*

681 We used the approach implemented in BAYPASS (v2.1) [68] to detect TEs displaying high
682 differentiation in northern populations. Overall divergence at each locus was first characterized
683 using the $X^TX$ statistics, which is a measure of adaptive differentiation corrected for population
684 structure and demography. Briefly, BAYPASS estimates a variance-covariance matrix reflecting
685 correlations between allele frequencies across populations, a description that can incorporate
686 admixture events and gene flow. This matrix is then used to correct differentiation statistics.
687 BAYPASS offers the option to estimate an empirical Bayesian *p*-value (*eBPis*) and a correlation
688 coefficient, which can be seen as the support for a non-random association between alleles and
689 specific populations. BAYPASS was run using default parameters under the core model and
690 using the matrix inferred from SNP data in [52]. We considered a TE as a candidate for selection
691 in northern populations when belonging to the top 1% $X^TX$ and 1% *eBPis*, and if the difference in
692 frequency with Florida was at least 0.5.

693 We compared our set of candidate TEs with the results obtained from a previous study on
694 positive selection in the same northern populations [52]. Briefly, three different methods were
695 applied and their results compared. We first used diploS/HIC [69], which is a machine-learning
696 approach that uses coalescent simulations with and without selection to estimate which genomic
697 regions are more likely to be under selection. This method has the advantage of incorporating
698 past fluctuations in population sizes, which may reduce the number of false positives due to

699 demography. We also used LSD [70], an approach that compares genealogies along genomic
700 windows and detects those harboring short branches in the focal population compared to its sister
701 clades, a signal of disruptive selection. At last, we also used BAYPASS on SNP data. Further
702 details can be found in Sup. Material and [52]. The set of candidate TEs for selection was
703 compared with the set of candidate windows for positive selection and the intersection was
704 extracted using BEDTOOLS (v2.25.0) [71].

705

706 ## Declarations

707 *Ethics approval and consent to participate*

708 Not applicable.

709 *Consent for publication*

710 Not applicable.

711 *Availability of data and materials*

712 The scripts used to perform simulations using SLiM3 are available on Github
713 (https://github.com/YannBourgeois/SLIM_simulations_TEs). All sequencing data are available
714 on the European Nucleotide Archive (https://www.ncbi.nlm.nih.gov/sra) under the BioProject
715 designation PRJNA376071 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA376071).

716 *Competing interests*

717 The authors declare that they have no competing interests

718 *Funding*

This work was supported by New York University Abu Dhabi (NYUAD) research funds AD180 (to SB). The NYUAD Sequencing Core is supported by NYUAD Research Institute grant G1205-1205A to the NYUAD Center for Genomics and Systems Biology. The funding bodies had no role in designing the study, nor in data collection and interpretation.

*Authors' contributions*

YB and SB designed the study. YB analyzed the data and ran the SLiM3 simulations. RR called TEs using MELT. YB, IH and SB contributed to the interpretation of results. YB and SB wrote the manuscript. All authors read and approved the final manuscript.

*Acknowledgments*

We are grateful to Breda Zimkus from the Museum of Comparative Zoology Cryogenic Collection in Harvard and J. Rosado from the Herpetology Collection for providing the samples of *Anolis porcatus* and *Anolis allisoni*. We thank Marc Arnoux from the Genome Core Facility at NYUAD for assistance with genome sequencing. This research was carried out on the High-Performance Computing resources at New York University Abu Dhabi.

## References

1. Sotero-Caio CG, Platt RN, Suh A, Ray DA. Evolution and diversity of transposable elements in vertebrate genomes. Genome Biol Evol. 2017;9:161–77.

2. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: From conflicts to benefits. Nat Rev Genet . Nature Publishing Group; 2017;18:71–86.

3. Song MJ, Schaack S. Evolutionary Conflict between Mobile DNA and Host Genomes. Am Nat . 2018;192:263–73.

740 4. Venner S, Feschotte C, Biémont C. Dynamics of transposable elements: towards a community
741 ecology of the genome. Trends Genet. 2009;25:317–23.

742 5. Brookfield JFY. The ecology of the genome - Mobile DNA elements and their hosts. Nat Rev
743 Genet. 2005;6:128–36.

744 6. Boissinot S, Davis J, Entezam A, Petrov D, Furano A V. Fitness cost of LINE-1 (L1) activity
745 in humans. Proc Natl Acad Sci . 2006;103:9590–4.

746 7. Kawakami T, Mugal CF, Suh A, Nater A, Burri R, Smeds L, et al. Whole-genome patterns of
747 linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-
748 scale recombination rate variation in birds. Mol Ecol. 2017;26:4158–72.

749 8. Liu S, Yeh CT, Ji T, Ying K, Wu H, Tang HM, et al. Mu transposon insertion sites and
750 meiotic recombination events co-localize with epigenetic marks for open chromatin across the
751 maize genome. PLoS Genet. 2009;5.

752 9. González J, Karasov TL, Messer PW, Petrov DA. Genome-wide patterns of adaptation to
753 temperate environments associated with transposable elements in Drosophila. PLoS Genet.
754 2010;6:33–5.

755 10. Lockton S, Ross-Ibarra J, Gaut BS. Demography and weak selection drive patterns of
756 transposable element diversity in natural populations of Arabidopsis lyrata. Proc Natl Acad Sci U
757 S A . 2008;105:13965–70.

758 11. Ruggiero RP, Bourgeois Y, Boissinot S. LINE Insertion Polymorphisms Are Abundant but at
759 Low Frequencies across Populations of Anolis carolinensis. Front Genet. 2017;8:1–14.

760 12. Xue AT, Ruggiero RP, Hickerson MJ, Boissinot S. Differential effect of selection against

761 LINE retrotransposons among vertebrates inferred from whole-genome data and demographic
762 modeling. Genome Biol Evol. 2018;10:1265–81.

763 13. Burri R. Interpreting differentiation landscapes in the light of long-term linked selection.
764 Evol Lett. 2017;1:118–31.

765 14. Charlesworth B, Charlesworth D. Elements of evolutionary genetics. Roberts Co. Publ. 2010.

766 15. Boissinot S, Entezam A, Furano A V. Selection Against Deleterious LINE-1-Containing Loci
767 in the Human Lineage. Mol Biol. 2001;18:926–35.

768 16. Villanueva-Cañas JL, Rech GE, de Cara MAR, González J. Beyond SNPs: how to detect
769 selection on transposable element insertions. Methods Ecol Evol . 2017;8:728–37.

770 17. Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, et al. Finding the
771 Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. Am Nat
772 . 2016;188:379–97.

773 18. Jangam D, Feschotte C, Betrán E. Transposable Element Domestication As an Adaptation to
774 Evolutionary Conflicts. Trends Genet . 2017;33:817–31.

775 19. Hof AEV t., Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, et al. The industrial
776 melanism mutation in British peppered moths is a transposable element. Nature . Nature
777 Publishing Group; 2016;534:102–5.

778 20. González J, Petrov DA. The adaptive role of transposable elements in the Drosophila
779 genome. Gene . Elsevier B.V.; 2009;448:124–33.

780 21. Bourgeois Y, Boissinot S. On the Population Dynamics of Junk: A Review on the Population
781 Genomics of Transposable Elements. Genes . 2019;10:419.

782 22. Feiner N. Accumulation of transposable elements in Hox gene clusters during adaptive
783 radiation of Anolis lizards. Proceedings Biol Sci . 2016;283.

784 23. Glor RE, Losos JB, Larson A. Out of Cuba: Overwater dispersal and speciation among
785 lizards in the Anolis carolinensis subgroup. Mol Ecol. 2005;14:2419–32.

786 24. Tollis M, Boissinot S. Genetic Variation in the Green Anole Lizard (Anolis carolinensis)
787 Reveals Island Refugia and a Fragmented Florida During the Quaternary. Genetica. 2014;1:59–
788 72.

789 25. Manthey JD, Tollis M, Lemmon AR, Moriarty Lemmon E, Boissinot S. Diversification in
790 wild populations of the model organism Anolis carolinensis : A genome-wide phylogeographic
791 investigation. Ecol Evol . 2016;6:8115–25.

792 26. Bourgeois Y, Ruggiero RP, Manthey JD, Boissinot S. Recent Secondary Contacts, Linked
793 Selection, and Variable Recombination Rates Shape Genomic Diversity in the Model Species
794 Anolis carolinensis. Genome Biol Evol. 2019;11:2009–22.

795 27. Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, Smeds L, et al. Linked selection and
796 recombination rate variation drive the evolution of the genomic landscape of differentiation
797 across the speciation continuum of Ficedula flycatchers. Genome Res. 2015;25:1656–65.

798 28. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due
799 to reduced diversity, not reduced gene flow. Mol Ecol. 2014;23:3133–57.

800 29. Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. Size matters: Non-LTR
801 retrotransposable elements and ectopic recombination in Drosophila. Mol Biol Evol.
802 2003;20:880–92.

803 30. Boissinot S, Sookdeo A. The Evolution of Line-1 in Vertebrates. Genome Biol Evol .
804 2016;evw247.

805 31. Cooper DM, Schimenti KJ, Schimenti JC. Factors affecting ectopic gene conversion in mice.
806 Mamm Genome. 1998;9:355–60.

807 32. Haller BC, Messer PW. SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher
808 Model. Mol Biol Evol. 2019;36:632–7.

809 33. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA
810 polymorphism. Genetics . 1989;123:585–95.

811 34. Hazzouri KM, Mohajer A, Dejak SI, Otto SP, Wright SI. Contrasting patterns of
812 transposable-element insertion polymorphism and nucleotide diversity in autotetraploid and
813 allotetraploid Arabidopsis species. Genetics. 2008;179:581–92.

814 35. García Guerreiro MP, Chávez-Sandoval BE, Balanyà J, Serra L, Fontdevila A. Distribution
815 of the transposable elements bilbo and gypsy in original and colonizing populations of
816 Drosophila subobscura. BMC Evol Biol. 2008;8:doi:10.1186/1471-2148-8-234.

817 36. Petrov DA, Fiston-Lavier A-S, Lipatov M, Lenkov K, Gonzalez J. Population Genomics of
818 Transposable Elements in Drosophila melanogaster. Mol Biol Evol . 2011;28:1633–44.
819 Available from: http://mbe.oxfordjournals.org/content/28/5/1633

820 37. Kofler R, Betancourt AJ, Schlötterer C. Sequencing of pooled DNA samples (Pool-Seq)
821 uncovers complex dynamics of transposable element insertions in Drosophila melanogaster.
822 PLoS Genet. 2012;8.

823 38. Hill WG, Robertson A. Local effects of limited recombination. Genet Res. 1966;8:269–94.

824 39. Dolgin ES, Charlesworth B. The effects of recombination rate on the distribution and
825 abundance of transposable elements. Genetics. 2008;178:2169–77.

826 40. Baller JA, Gao J, Voytas DF. Access to DNA establishes a secondary target site bias for the
827 yeast retrotransposon Ty5. Proc Natl Acad Sci U S A. 2011;108:20351–6.

828 41. Yoshida J, Akagi K, Misawa R, Kokubu C, Takeda J, Horie K. Chromatin states shape
829 insertion profiles of the piggyBac, Tol2 and Sleeping Beauty transposons and murine leukemia
830 virus. Sci Rep. Nature Publishing Group; 2017;7:1–18.

831 42. Flasch DA, Macia Á, Sánchez L, Ljungman M, Heras SR, García-Pérez JL, et al. Genome-
832 wide de novo L1 Retrotransposition Connects Endonuclease Activity with Replication. Cell.
833 2019;177:837-851.e28.

834 43. Sultana T, van Essen D, Siol O, Bailly-Bechet M, Philippe C, Zine El Aabidine A, et al. The
835 Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence
836 Biases and Post-insertion Selection. Mol Cell. 2019;74:555-570.e7.

837 44. Suh A, Smeds L, Ellegren H. Abundant recent activity of retrovirus-like retrotransposons
838 within and among flycatcher species implies a rich source of structural variation in songbird
839 genomes. Mol Ecol. 2018;27:99–111.

840 45. Hudson RR. Properties of a neutral allele model with intragenic recombination. Theor Popul
841 Biol. 1983;23:183–201.

842 46. de Boer JG, Yazawa R, Davidson WS, Koop BF. Bursts and horizontal evolution of DNA
843 transposons in the speciation of pseudotetraploid salmonids. BMC Genomics. 2007;8:1–10.

844 47. Hellen EHB, Brookfield JFY. Transposable element invasions. Mob Genet Elements .

2013;3:e23920.

48. Vieira C, Lepetit D, Dumont S, Biémont C. Wake up of transposable elements following Drosophila simulans worldwide colonization. Mol Biol Evol. 1999;16:1251–5.

49. Piegu B, Guyot R, Picault N, Roulin A, Saniyal 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;16:1262–9.

50. Manthey JD, Moyle RG, Boissinot S. Multiple and independent phases of transposable element amplification in the genomes of piciformes (woodpeckers and allies). Genome Biol Evol. 2018;10:1445–56.

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

52. Bourgeois Y, Boissinot S. Selection at behavioural, developmental and metabolic genes is associated with the northward expansion of a successful tropical colonizer. Mol Ecol. 2019;28:3523–43.

53. Missler M, Zhang W, Rohlmann A, Kattenstroth G, Hammer RE, Gottmann K, et al. α-neurexins couple Ca 2+ channels to synaptic vesicle exocytosis. Nature. 2003;423:939–48.

54. Vetrini F, McKee S, Rosenfeld JA, Suri M, Lewis AM, Nugent KM, et al. De novo and inherited TCF20 pathogenic variants are associated with intellectual disability, dysmorphic features, hypotonia, and neurological impairments with similarities to Smith-Magenis syndrome. Genome Med. Genome Medicine; 2019;11:1–17.

55. Schäfgen J, Cremer K, Becker J, Wieland T, Zink AM, Kim S, et al. De novo nonsense and

866 frameshift variants of TCF20 in individuals with intellectual disability and postnatal overgrowth.
867 Eur J Hum Genet. 2016;24:1739–45.

868 56. González J, Lenkov K, Lipatov M, Macpherson JM, Petrov DA. High rate of recent
869 transposable element-induced adaptation in Drosophila melanogaster. PLoS Biol. 2008;6:2109–
870 29.

871 57. Rech GE, Bogaerts-Marquez M, Barron MG, Merenciano M, Villanueva-Canas JL, Horvath
872 V, et al. Stress response, behavior, and development are shaped by transposable element-induced
873 mutations in Drosophila. PloS Genet. 2018;15:e1007900.

874 58. Schrider DR, Kern AD. Supervised Machine Learning for Population Genetics: A New
875 Paradigm. Trends Genet . 2018;34:301–12.

876 59. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, et al. The Mobile
877 Element Locator Tool (MELT): Population-scale mobile element discovery and biology.
878 Genome Res . 2017;gr.218032.116.

879 60. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in
880 eukaryotic genomes. Mob DNA. 2015;6–11.

881 61. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call
882 format and VCFtools. Bioinformatics. 2011;27:2156–8.

883 62. R Development Core Team R. R: A Language and Environment for Statistical Computing . R
884 Found. Stat. Comput. 2011. Available from: http://www.r-project.org

885 63. McVean G, Awadalla P, Fearnhead P. A coalescent-based method for detecting and
886 estimating recombination from gene sequences. Genetics. 2002;160:1231–41.

887 64. Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ. PopGenome: An efficient swiss
888 army knife for population genomic analyses in R. Mol Biol Evol. 2014;31:1929–36.

889 65. Ginestet C. ggplot2: Elegant Graphics for Data Analysis. J R Stat Soc Ser A (Statistics Soc.
890 2011;

891 66. Kryukov G V., Schmidt S, Sunyaev S. Small fitness effect of mutations in highly conserved
892 non-coding regions. Hum Mol Genet. 2005;14:2221–9.

893 67. Myers S, Freeman C, Auton A, Donnelly P, McVean G. A common sequence motif
894 associated with recombination hot spots and genome instability in humans. Nat Genet.
895 2008;40:1124–9.

896 68. Gautier M. Genome-Wide Scan for Adaptive Divergence and Association with Population-
897 Specific Covariates. Genetics. 2015;201:1555–79.

898 69. Kern AD, Schrider DR. diploS/HIC: An Updated Approach to Classifying Selective Sweeps.
899 G3; Genes|Genomes|Genetics . 2018;g3.200262.2018.

900 70. Librado P, Orlando L. Detecting signatures of positive selection along defined branches of a
901 population tree using LSD. Mol Biol Evol . 2018;1–16.

902 71. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
903 Bioinformatics . 2010;26:841–2.

904 72. Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from
905 hundreds of unphased whole genomes. Nat Genet . Nature Publishing Group; 2016;49:303–9.

906

907

# Tables

Table 1: Summary of TE polymorphisms in the five genetic clusters identified in the green anole, and its two Cuban counterparts. For each cluster/outgroup, the number of polymorphic or fixed elements is given.

| Category | Clade | N | Fixed | Heterozygous | Fixed | Heterozygous | Fixed | Polymorphic | Fixed | Polymorphic | Fixed | Polymorphic | Fixed | Polymorphic | Fixed | Polymorphic |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **nLTR-RT** | *CR1* | 32804 | 3892 | 2613 | 3343 | 2795 | 3557 | 6712 | 3488 | 6783 | 3328 | 13601 | 4059 | 5175 | 4357 | 3476 |
| | *L2* | 26392 | 4261 | 4396 | 4469 | 5051 | 6577 | 5196 | 6604 | 4962 | 6259 | 6250 | 7004 | 2929 | 7266 | 2064 |
| | *Penelope* | 16208 | 978 | 1375 | 1313 | 1470 | 1074 | 2935 | 1056 | 3611 | 915 | 6426 | 1081 | 2670 | 1180 | 1876 |
| | *L1* | 14181 | 1057 | 1474 | 1088 | 1594 | 1023 | 2842 | 1012 | 3231 | 914 | 5414 | 1128 | 2137 | 1243 | 1491 |
| | *RTE1* | 3709 | 418 | 232 | 370 | 372 | 348 | 309 | 352 | 943 | 332 | 1152 | 362 | 606 | 375 | 221 |
| | *R4* | 1516 | 166 | 345 | 253 | 427 | 265 | 308 | 274 | 310 | 286 | 243 | 303 | 134 | 301 | 98 |
| | *RTE_BovB* | 920 | 240 | 209 | 302 | 235 | 358 | 167 | 377 | 151 | 347 | 187 | 380 | 83 | 392 | 66 |
| | *Vingi* | 860 | 360 | 174 | 306 | 151 | 777 | 75 | 783 | 77 | 758 | 102 | 823 | 37 | 826 | 34 |
| | *RTEX* | 496 | 128 | 134 | 206 | 146 | 416 | 73 | 425 | 68 | 380 | 116 | 447 | 49 | 465 | 31 |
| | *Neptune* | 376 | 14 | 4 | 4 | 12 | 6 | 37 | 4 | 57 | 3 | 215 | 3 | 56 | 4 | 28 |
| | Other | 124 | 14 | 18 | 18 | 24 | 25 | 22 | 26 | 18 | 24 | 45 | 25 | 28 | 28 | 15 |
| **DNA transposons** | *Hobo* | 45421 | 1380 | 6693 | 986 | 6900 | 244 | 11869 | 31 | 13042 | 2 | 19344 | 6 | 9900 | 39 | 6509 |
| | *Tc1/Mariner* | 37718 | 8380 | 6692 | 8072 | 8133 | 4533 | 6681 | 4600 | 6386 | 4190 | 11070 | 4759 | 4661 | 4935 | 3112 |
| | *hAT* | 25165 | 3520 | 5115 | 6705 | 5043 | 8679 | 5348 | 8778 | 5115 | 7841 | 7515 | 9058 | 4252 | 9461 | 2797 |
| | *Helitron* | 19266 | 1730 | 2008 | 1093 | 2491 | 147 | 3899 | 21 | 5007 | 2 | 7779 | 4 | 2424 | 24 | 1499 |
| | Other | 3517 | 569 | 1015 | 1783 | 878 | 2847 | 636 | 2958 | 554 | 2666 | 850 | 3098 | 419 | 3203 | 314 |
| | *Chapaev* | 1229 | 154 | 329 | 491 | 284 | 783 | 286 | 833 | 204 | 728 | 360 | 837 | 198 | 896 | 123 |
| | *MER* | 17 | 6 | 3 | 8 | 4 | 13 | 4 | 14 | 3 | 12 | 5 | 16 | 1 | 15 | 2 |
| | *Eulor* | 16 | 2 | 5 | 10 | 3 | 13 | 3 | 13 | 3 | 13 | 3 | 15 | 1 | 15 | 1 |
| | *UCON* | 11 | 0 | 1 | 4 | 5 | 7 | 4 | 9 | 2 | 9 | 2 | 10 | 1 | 10 | 1 |
| | *Chompy* | 5 | 1 | 2 | 3 | 1 | 4 | 1 | 5 | 0 | 5 | 0 | 4 | 1 | 5 | 0 |
| | *Harbinger* | 3 | 2 | 1 | 0 | 1 | 1 | 2 | 2 | 1 | 3 | 0 | 2 | 1 | 3 | 0 |
| | *REP* | 2 | 0 | 1 | 1 | 1 | 1 | 1 | 2 | 0 | 1 | 1 | 1 | 1 | 2 | 0 |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **LTR-RT** | *Gypsy* | 45625 | 1037 | 1223 | 985 | 1299 | 1338 | 7408 | 1029 | 9186 | 940 | 15157 | 1106 | 12939 | 1218 | 8020 |
| | Other LTRs | 13946 | 219 | 349 | 241 | 322 | 753 | 3719 | 441 | 4237 | 366 | 6215 | 419 | 2958 | 509 | 2153 |
| | *BEL* | 9391 | 183 | 189 | 125 | 234 | 166 | 1329 | 157 | 1154 | 148 | 4380 | 175 | 1929 | 194 | 1369 |
| | *Dirs* | 6873 | 391 | 577 | 297 | 607 | 362 | 1226 | 320 | 1618 | 315 | 1796 | 363 | 1332 | 393 | 930 |
| | *Copia* | 1962 | 324 | 268 | 315 | 550 | 257 | 279 | 239 | 392 | 233 | 465 | 256 | 233 | 276 | 132 |
| | *ERV* | 674 | 79 | 62 | 80 | 63 | 92 | 112 | 89 | 135 | 87 | 178 | 94 | 130 | 102 | 75 |
| | Ultra-conserved | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| **SINEs** | *SINE2* | 20716 | 4121 | 5185 | 4568 | 5757 | 5600 | 3275 | 5661 | 2954 | 5331 | 3846 | 5837 | 1769 | 5960 | 1294 |
| | Non-assigned (ACASINE) | 4802 | 997 | 1577 | 1427 | 1564 | 1908 | 818 | 1941 | 836 | 1839 | 1024 | 1984 | 554 | 2024 | 418 |
| | *SINE1* | 4115 | 271 | 440 | 195 | 569 | 57 | 1031 | 27 | 1030 | 23 | 1423 | 25 | 473 | 29 | 367 |
| | *SINE3* | 1083 | 87 | 365 | 297 | 380 | 500 | 191 | 518 | 173 | 463 | 243 | 536 | 115 | 549 | 84 |
| | *MIR*-like | 5 | 0 | 2 | 4 | 1 | 5 | 0 | 5 | 0 | 2 | 3 | 5 | 0 | 5 | 0 |

Table 2: Summary of correlations observed between average recombination rate, the average frequency of TEs, the density of polymorphic TEs and the density of fixed elements. For short and long LINEs, due to the low number of fixed insertions in 1Mb windows, we present results for 10MB windows instead. The last column provides an interpretation of the correlations obtained in simulations and observed in empirical data. For simulated TEs, we distinguish between outcomes where TEs are at high frequency (higher than SNPs) and low frequency (lower than SNPs). PS: Purifying Selection against ectopic recombination; LS: Linked Selection; PI: Preferential Insertion in regions of high recombination/open chromatin; ABT: Ancient Burst of Transposition; (): the process may occur but does not impact the direction of correlations (for simulations), or is possible but no conclusive evidence is provided by the three summary statistics (for empirical observations). NA: for *Helitron* and *Hobo*, the lack of fixed insertions prevents the computation of these statistics. *: *P*-value<0.05; **: *P*-value<0.01; ***: *P*-value<0.001.

| Category | Superfamily/simulation | Correlation between recombination rate and: | | | Dominant proces |
|---|---|---|---|---|---|
| | | Average frequency | Polymorphic density | Fixed density | |
| **simulations** | simulated TE | + | + | + | LS + PI |
| | simulated TE | + | + | - | LS |
| | simulated TE (high frequency) | - | + | + | LS + ABT + SPI |
| | simulated TE (high frequency) | - | + | - | LS + ABT |
| | simulated TE (low frequency) | - | + | - | PS + (ABT) + PI |
| | simulated TE | - | - | - | PS + (ABT) |
| **nLTR-RTs** | *CR1* | 0.05 | 0.15*** | -0.08 * | Mixture |
| | *CR1* (short) | 0.30** | 0.25** | -0.24* | LS |
| | *CR1* (long) | -0.28 ** | -0.14 | -0.32 ** | PS |
| | *L1* | 0.02 | 0.01 | 0.08 | Mixture |
| | *L1* (short) | -0.006 | 0.35 *** | 0.08 | LS + PI ? |
| | *L1* (long) | -0.25 * | -0.29 ** | -0.57 *** | PS |
| | *L2* | 0.05 | 0.29 *** | -0.06 | Mixture |
| | *L2* (short) | 0.21 * | 0.50 *** | -0.10 | LS + (PI) |
| | *L2* (long) | -0.24 * | 0.02 | -0.32 ** | PS + (PI) |
| | *Penelope* | 0.04 | 0.15*** | 0.08 | LS + PI ? |

| | | | | | |
|---|---|---|---|---|---|
| **SINEs** | *SINE2* | 0.19 *** | 0.42 *** | 0.28 *** | LS + PI |
| | Other SINEs | 0.24 *** | 0.37 *** | 0.35 *** | LS + PI |
| **LTR-RT** | *Dirs* | -0.21 *** | 0.14*** | -0.09 | PS + PI |
| | *BEL* | -0.28 *** | 0.01 | -0.05 | PS + (PI) |
| | *Gypsy* | -0.18 *** | -0.13 *** | -0.11 * | PS |
| | Other LTRs | -0.18 *** | -0.04 | -0.04 | PS + (PI) |
| **DNA transposons** | *hAT* | 0.04 | 0.51 *** | 0.29 *** | LS + PI + ABT |
| | *Helitron* | 0.10 ** | 0.32 *** | NA | LS + (PI) |
| | *Hobo* | -0.03 | 0.32 *** | NA | LS + (PI) |
| | *Tc1/Mariner* | 0.20 *** | 0.59 *** | 0.20 *** | LS + PI |

939

940

941

942

943

944 Table 3: Summary of the four TE insertions candidate for positive selection. None of these insertions was found in *A. allisoni* and *A.*

945 *porcatus*.

946

| Chromosome | Position | Clade | Gene | Frequency in Florida | Frequency in North |
|---|---|---|---|---|---|
| 1 | 260526442 | *L1* | *Neurexin2* | 1/30 | 21/22 |
| 1 | 260754973 | *CR1* | *Neurexin2* | 2/30 | 22/22 |
| 2 | 128671912 | *ERV* | between *FNIP1* and *RAPGEF6* | 0/30 | 16/22 |
| 5 | 27319544 | *CR1* | *TCF20-201* | 1/30 | 21/22 |

947

948

# Figures

Figure 1: Count of heterozygous sites across all 28 individuals included in this study. Vertical dotted lined delimit the five main genetic clusters and the two outgroups in this order: *A. allisoni* and *A. porcatus*, SF, NWF, NEF, GA and CA. See Sup. Figure 1 for more details about these clusters.

Figure 2: Allele frequency spectra for TEs belonging to two genetic clusters identified in the green anole. NEF (N=8 diploid individuals) corresponds to a large, stable population from Florida, and GA (N=7 diploid individuals) corresponds to a more recently established population having colonized northern environments in the last 100,000 years. A: nLTR-RT; B: SINEs; C: LTR-RT; D: DNA-transposons.

Figure 3: Correlograms illustrating Spearman's rank correlation coefficients between TE densities and genomic features such as recombination rate (measured as $r/\mu$, see Methods), pairwise relative ($F_{ST}$) and absolute ($d_{XY}$) measures of differentiation, and derived SNP frequency in the NEF cluster (DAF). Correlations with $P>0.05$ are indicated with a cross.

Figure 4: Boxplots of average TE frequency for each main TE category in the NEF population. For SNPs, the derived allele frequency was obtained by assigning variants to ancestral and derived states using *A. allisoni* and *A. porcatus*.

Figure 5: Plots of average TE frequency against recombination rate computed over 1Mb windows for each main TE clade in the NEF population.

Figure 6: Plots of polymorphic TE density against recombination rate computed over 1Mb windows for each main TE clade in the NEF population.

970 Figure 7: Plots of fixed TE density against recombination rate computed over 1Mb windows for
971 each main clade in the NEF population. For *Helitron* and *Hobo*, there are not enough fixed
972 insertions.

973 Figure 8: Top: plots of LINEs length against recombination rate. Middle: Plots of average
974 frequency, density of polymorphic insertions and density of fixed insertions for short LINEs,
975 Bottom: same as middle row, for long LINEs. For middle and bottom plots, average frequencies
976 and densities are computed for 10Mb windows.

977 Figure 9: Summary of simulations of TEs using SLiM3, using parameters realistic for the NEF
978 cluster. Eight diploid individuals were sampled to mimic our sampling scheme. Boxplots
979 correspond to the results obtained over 100 simulations of a 4Mb fragment, divided into three
980 regions of 1, 2 and 1 Mb. The first and last Mb correspond to regions of high recombination (10
981 times higher than the 2Mb central region). Coefficients of selection and other parameters are
982 scaled using an effective population size of 1000 instead of 1,000,000 to reduce computation
983 time (see Methods). Blue and red dotted lines correspond to average derived SNP frequencies in
984 regions of low and high recombination respectively.

985

986

987

988

989

990

# Supplementary figures

Sup. Figure 1. Summary of population structure and environmental variation in green anoles (see [26] for further details). A: RAxML phylogeny on one million random SNPs. B: Demographic evolution of the five genetic clusters of green anoles reconstructed by SMC++ [72]. C: Sampling locations used in this study. Units for temperature are in tenth of Celsius degrees. D: PCA over environmental variables (BIOCLIM data) for the locations used in this study. Larger dots highlight the northern clades (GA and CA) and their sister Floridian clade (NEF).

Sup. Figure 2-5: Allele frequency spectra for TEs belonging to all five genetic clusters identified in the green anole. Sup. Figure 3: nLTR-RT; Sup. Figure 4: SINEs; Sup. Figure 5: LTR-RT; Sup. Figure 6: DNA-transposons.

Sup. Figure 6. Plot of the correlation between exon density and scaled recombination rate.

Sup. Figure 7. Graphical summary of SLiM3 simulation parameters. We simulate a 4Mb fragment, assuming the following unscaled parameters (see Methods for details about scaling): a stable effective population size of 1 million individuals, a mutation rate of $2.1.10^{-10}$/year, high recombination in the first and last Mb ($r$=$2.10^{-10}$ /year), low recombination in the 2 Mb in the middle (r= $2.10^{-11}$ /generation). Linked selection is modelled by introducing 10% of deleterious mutations with $2.Ne.s$ = -10. We assume that there are 10 TE progenitors in the whole genome that can jump $P$ generations/genome (constant rate). We also model bursts of transposition where the probability of jumping is 100X higher, but transposition occurs during a lapse of 100,000 years, deviating from transposition-drift balance. We also add an insertion bias $Q$ to model preferential insertion in regions of high recombination.

## nLTR−RTs

CR1

L1

L2

Penelope

Count

*A. porcatus*

*A. allisoni*

*SF*

*NEF*

*NWF*

*GA*

*CA*

Genetic clusters in the green anole (*A. carolinensis*)

## SINEs

ACASINE

SINE1

SINE2

SINE3

Count

Heterozygous TE

Private

Shared with at least one outgroup

Shared with green anole only

## LTR−RTs

BEL

Copia

Dirs

Gypsy

Other LTRs

Count

## DNA transposons

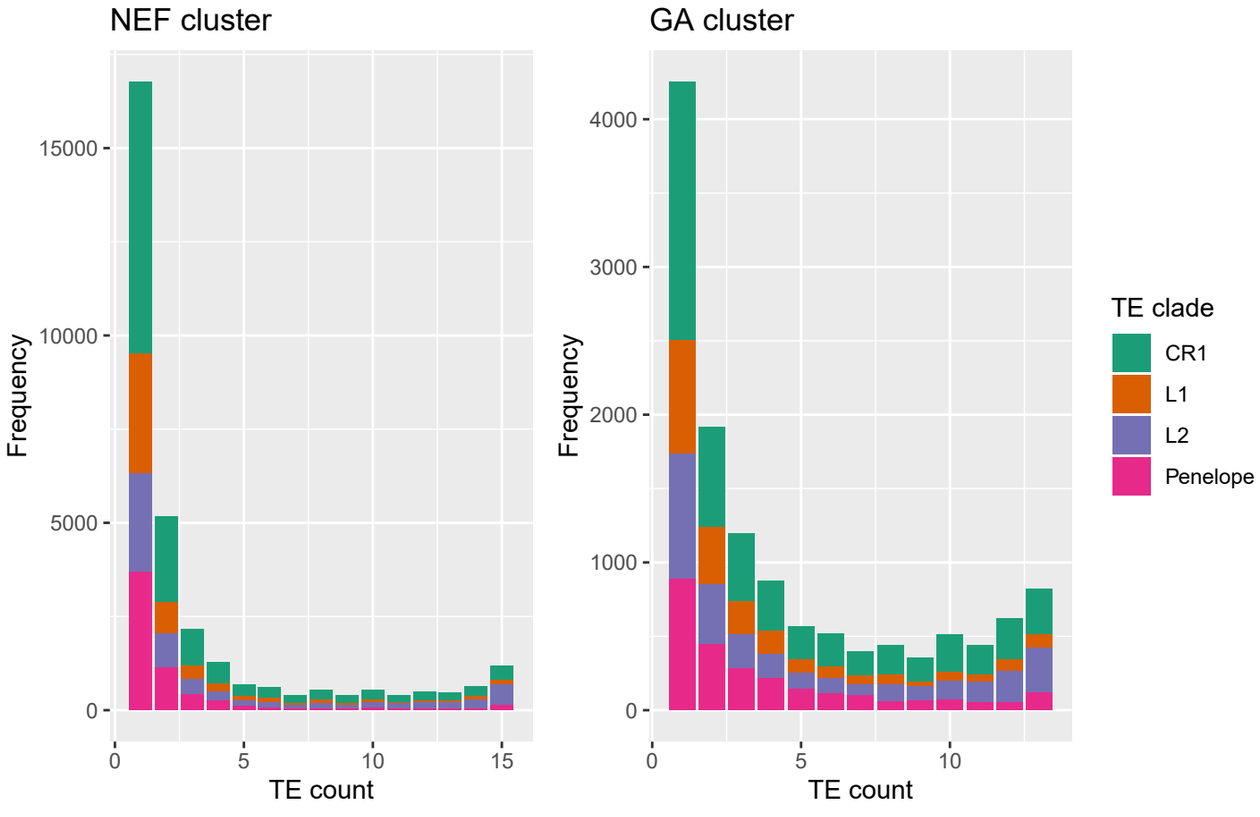
hAT

Helitron

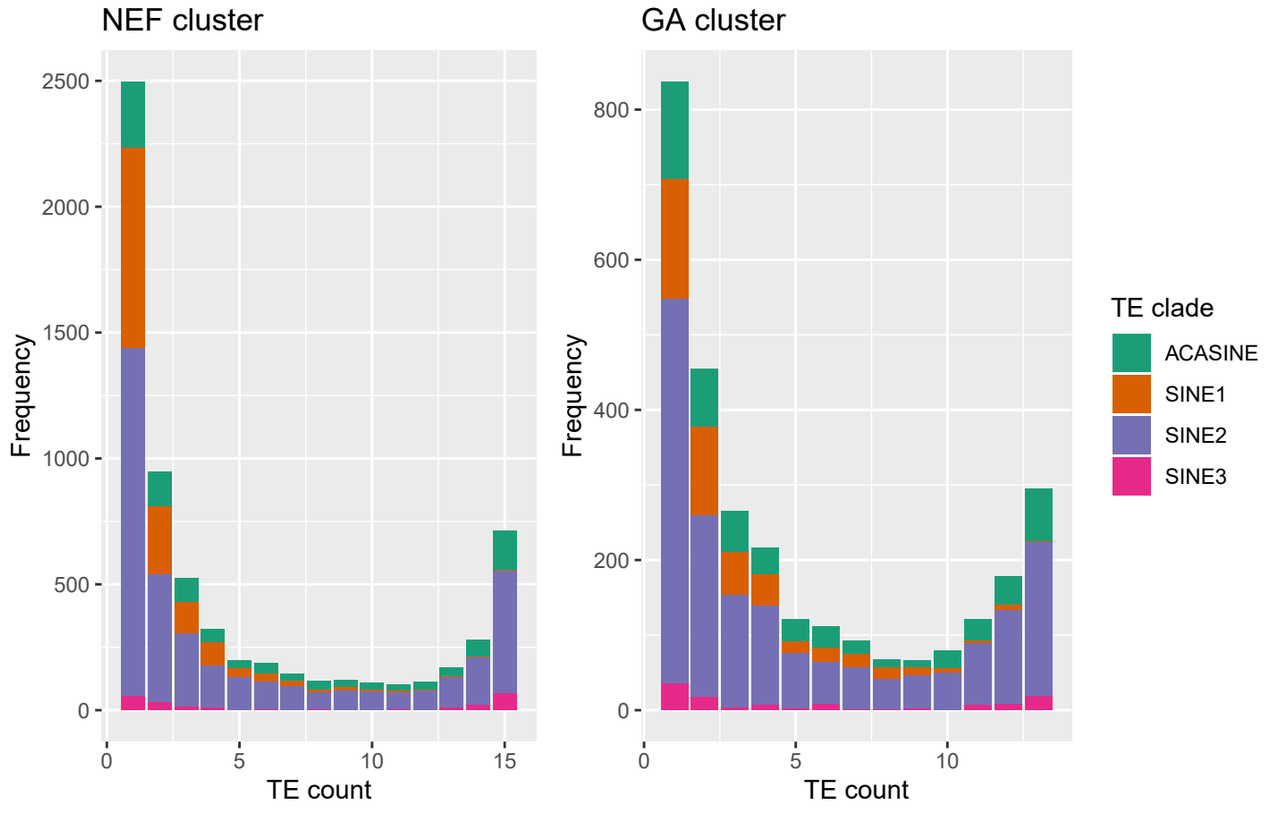
Hobo

Tc1/Mariner

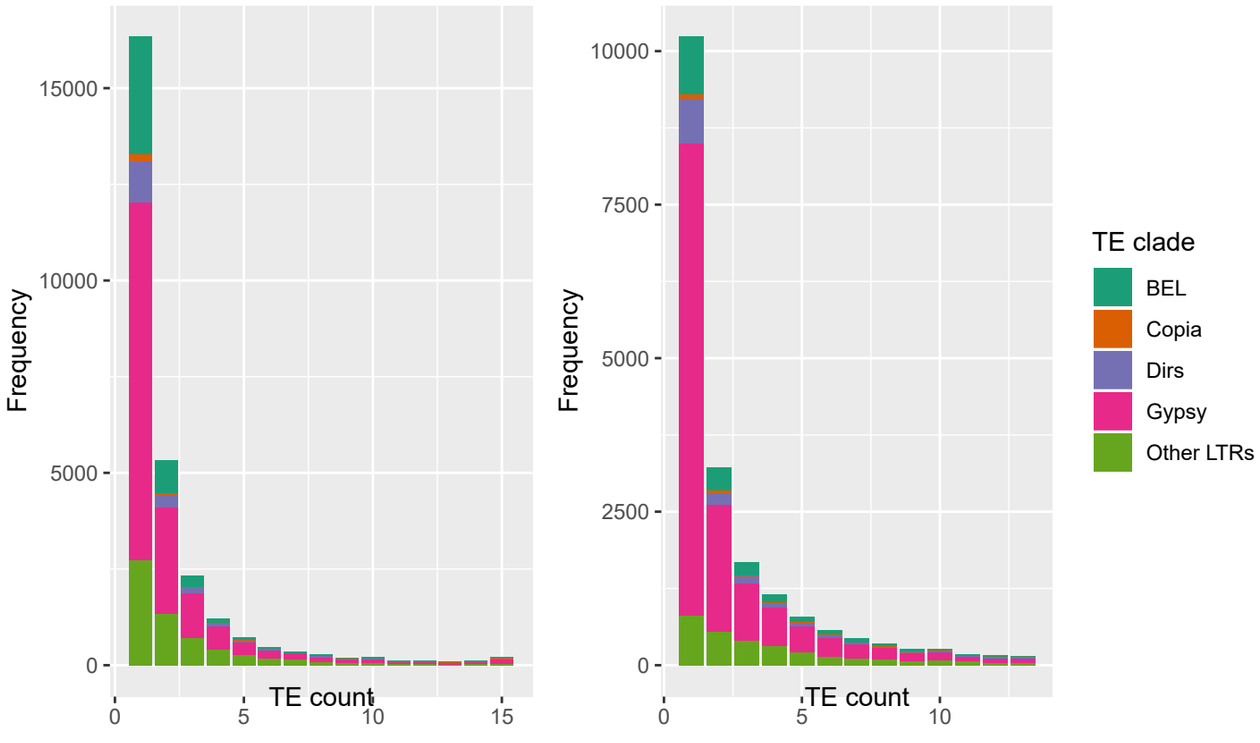
A


NEF cluster
GA cluster
Frequency
TE count
TE clade
CR1
L1
L2
Penelope

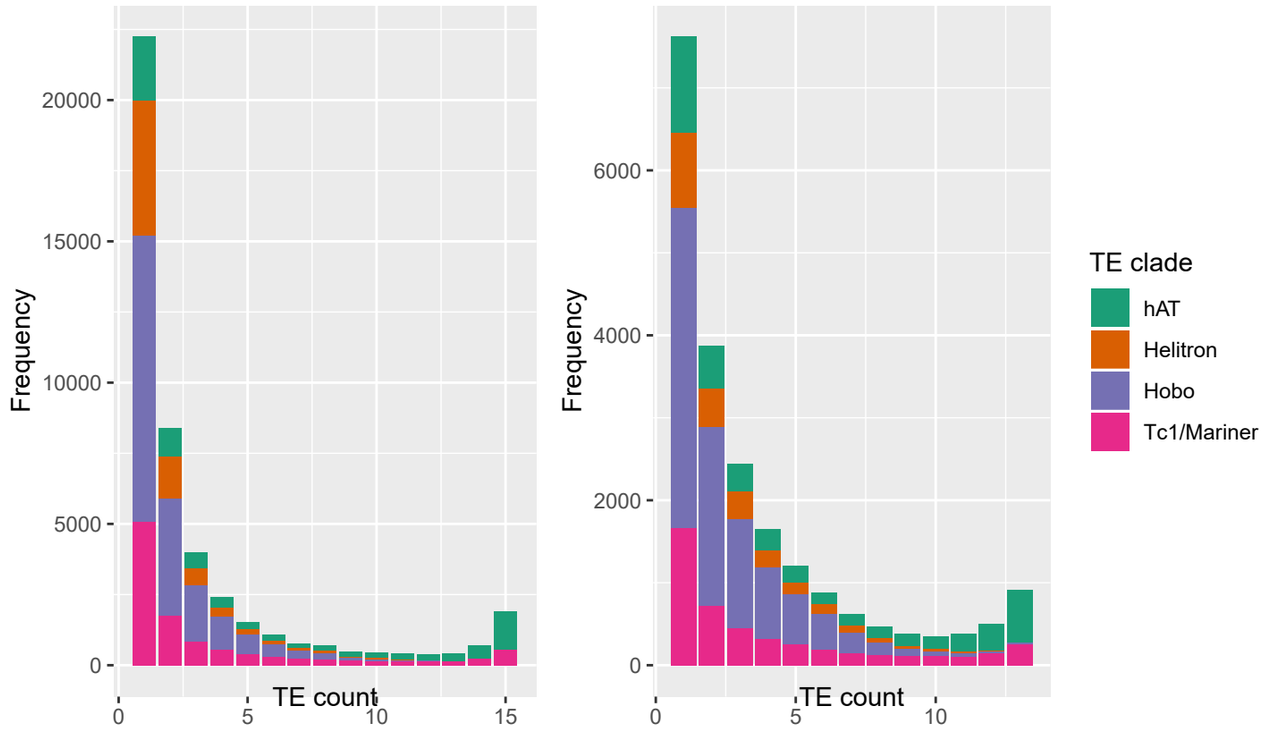
B

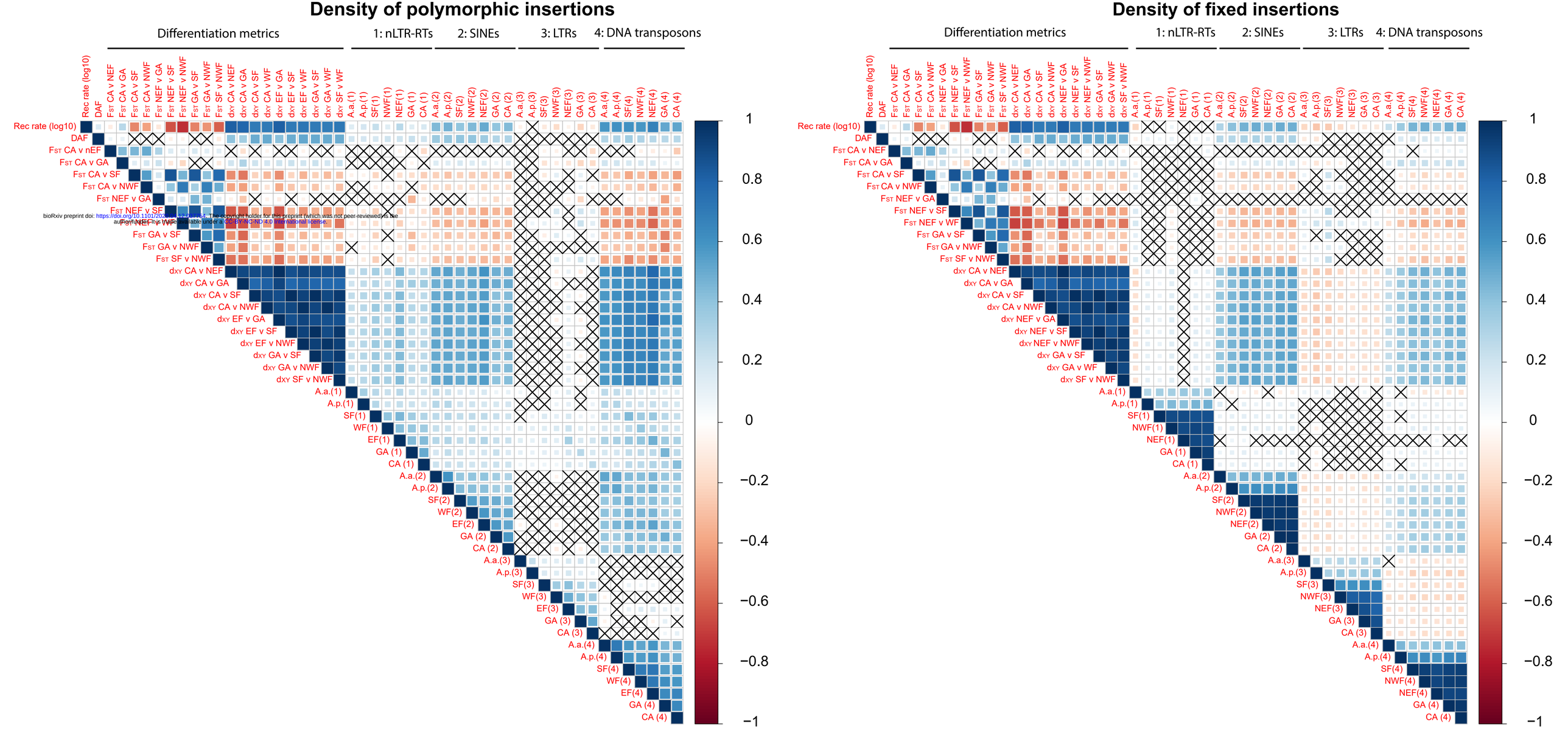

NEF cluster
GA cluster
Frequency
TE count
TE clade
ACASINE
SINE1
SINE2
SINE3

C

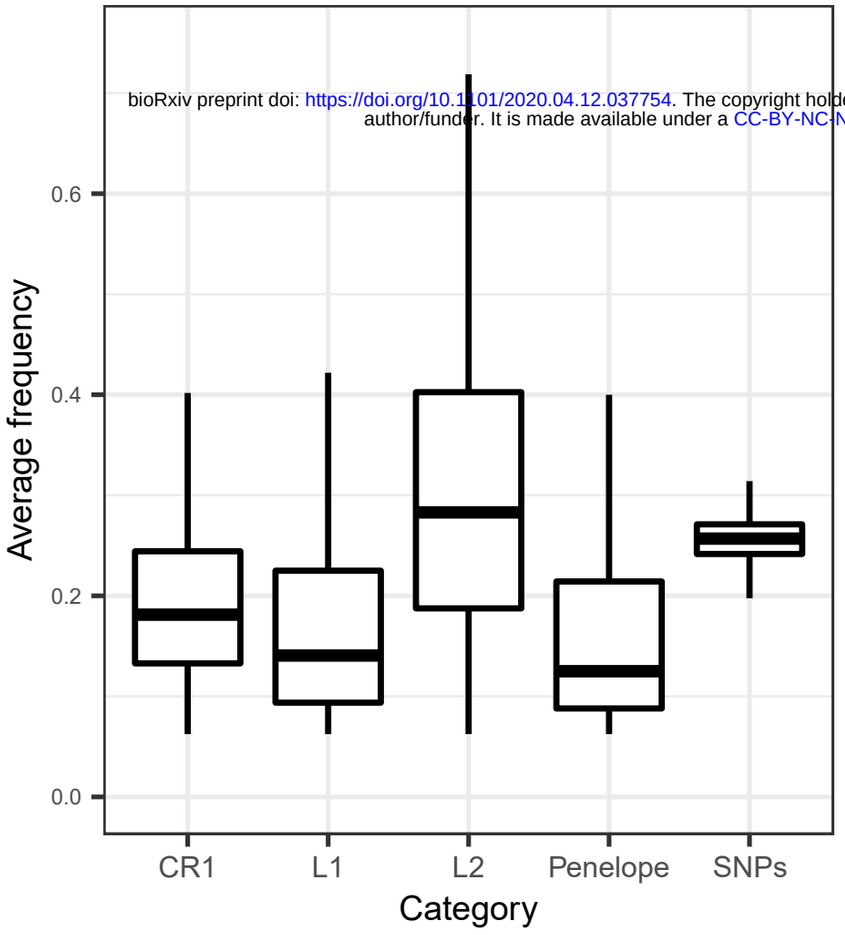

NEF cluster
GA cluster
Frequency
TE count
TE clade
BEL
Copia
Dirs
Gypsy
Other LTRs

D

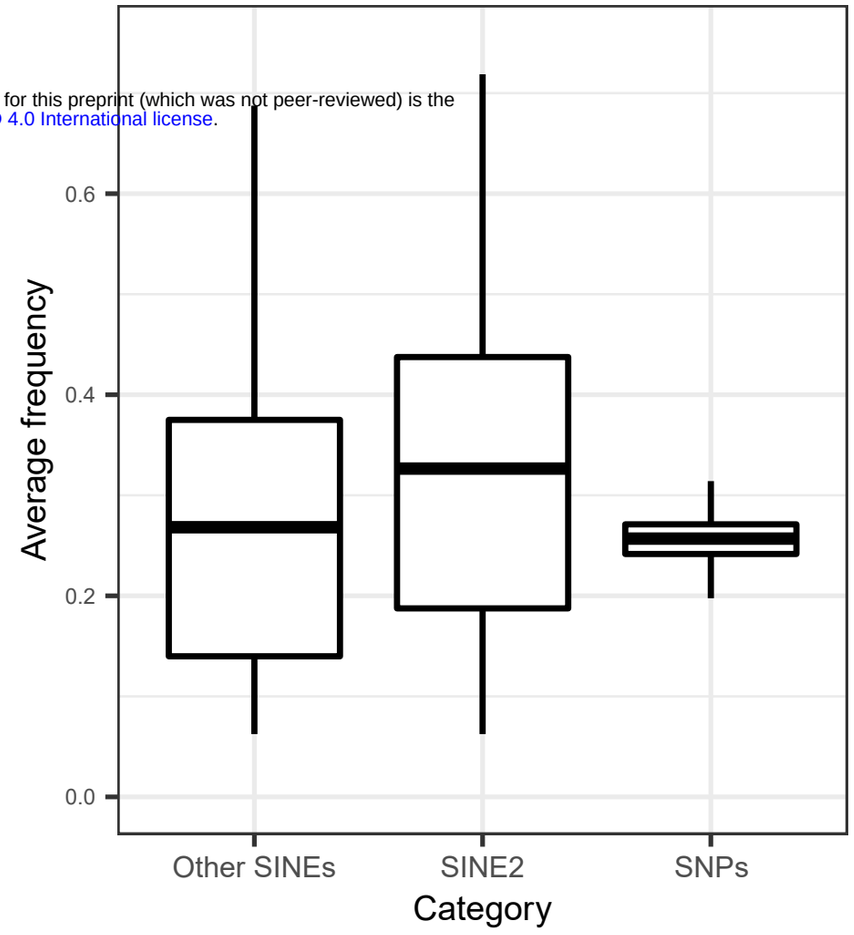

NEF cluster
GA cluster
Frequency
TE count
TE clade
hAT
Helitron
Hobo
Tc1/Mariner



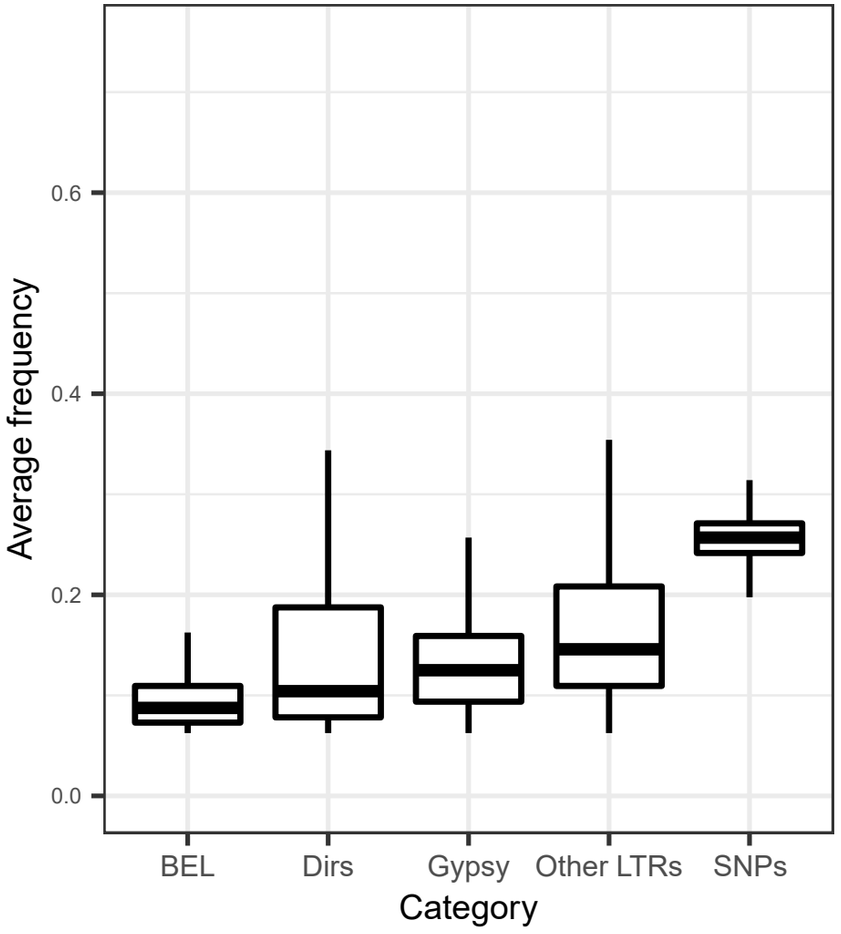
## LINEs and PLP (NEF)



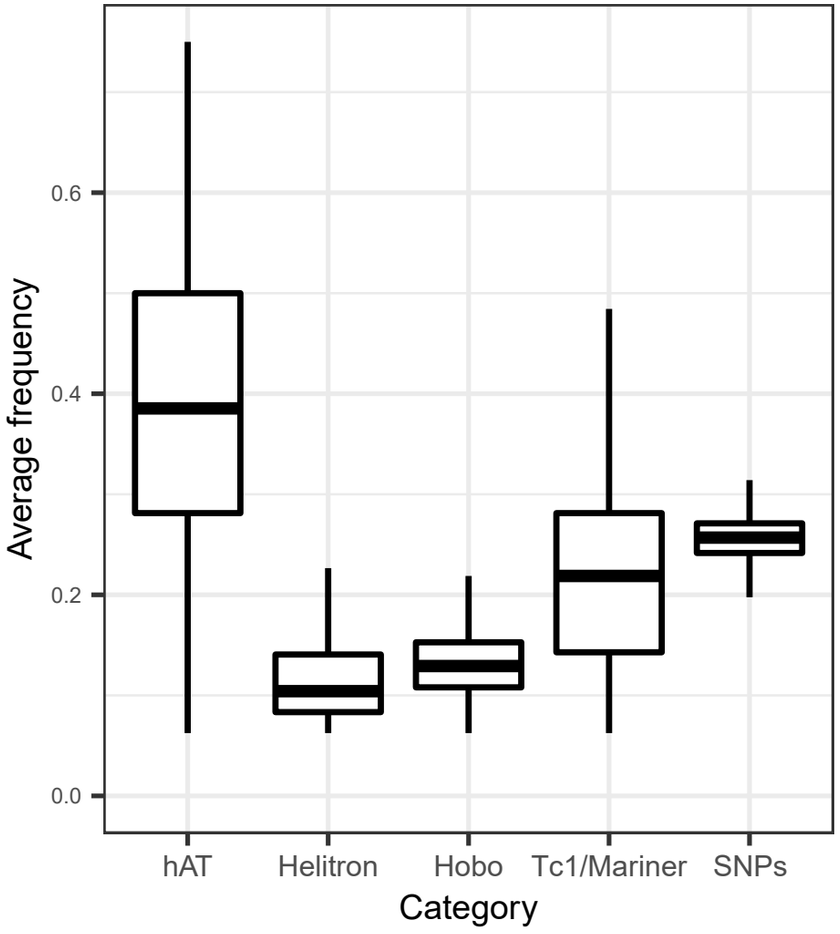

## SINEs (NEF)

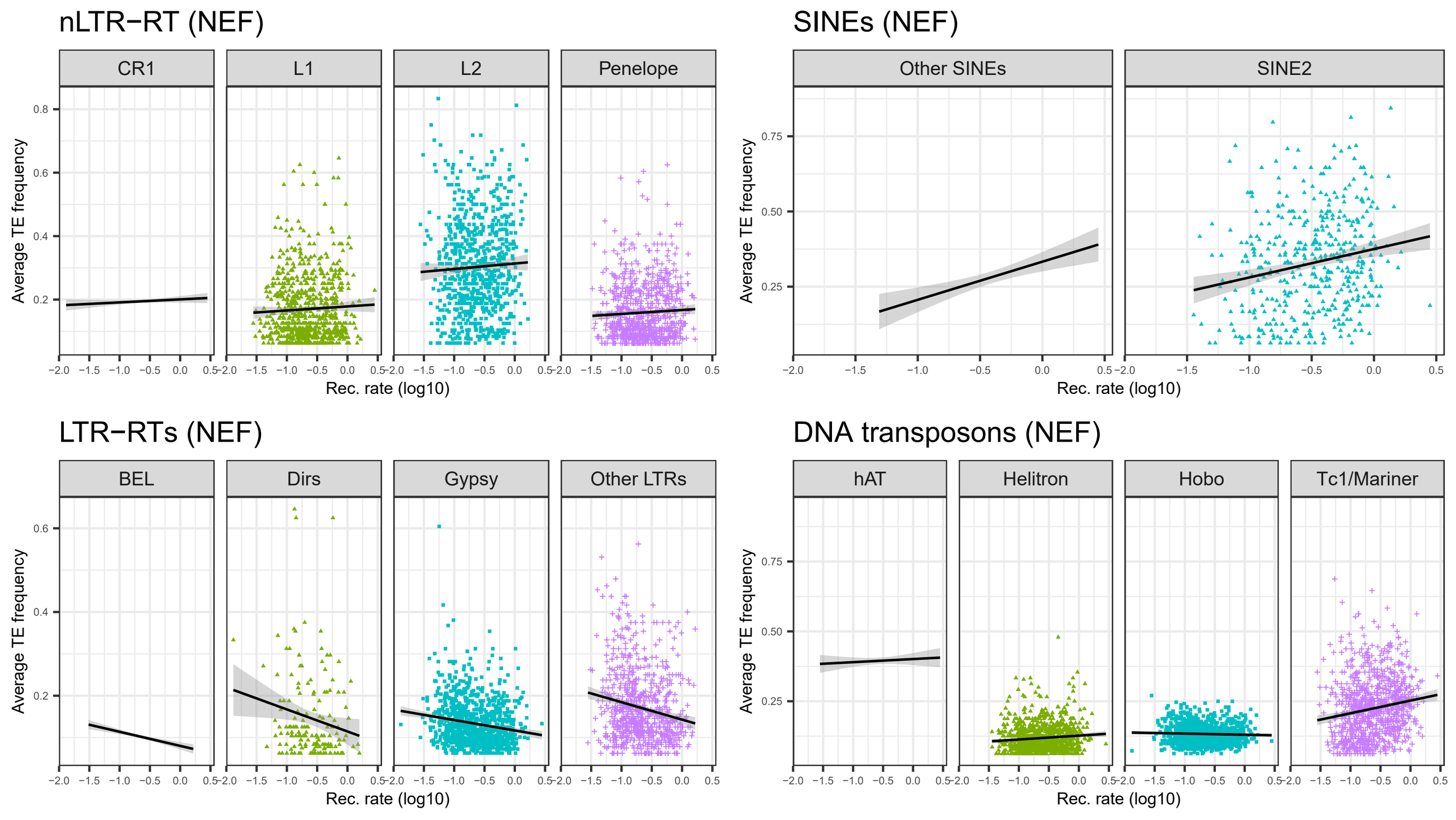


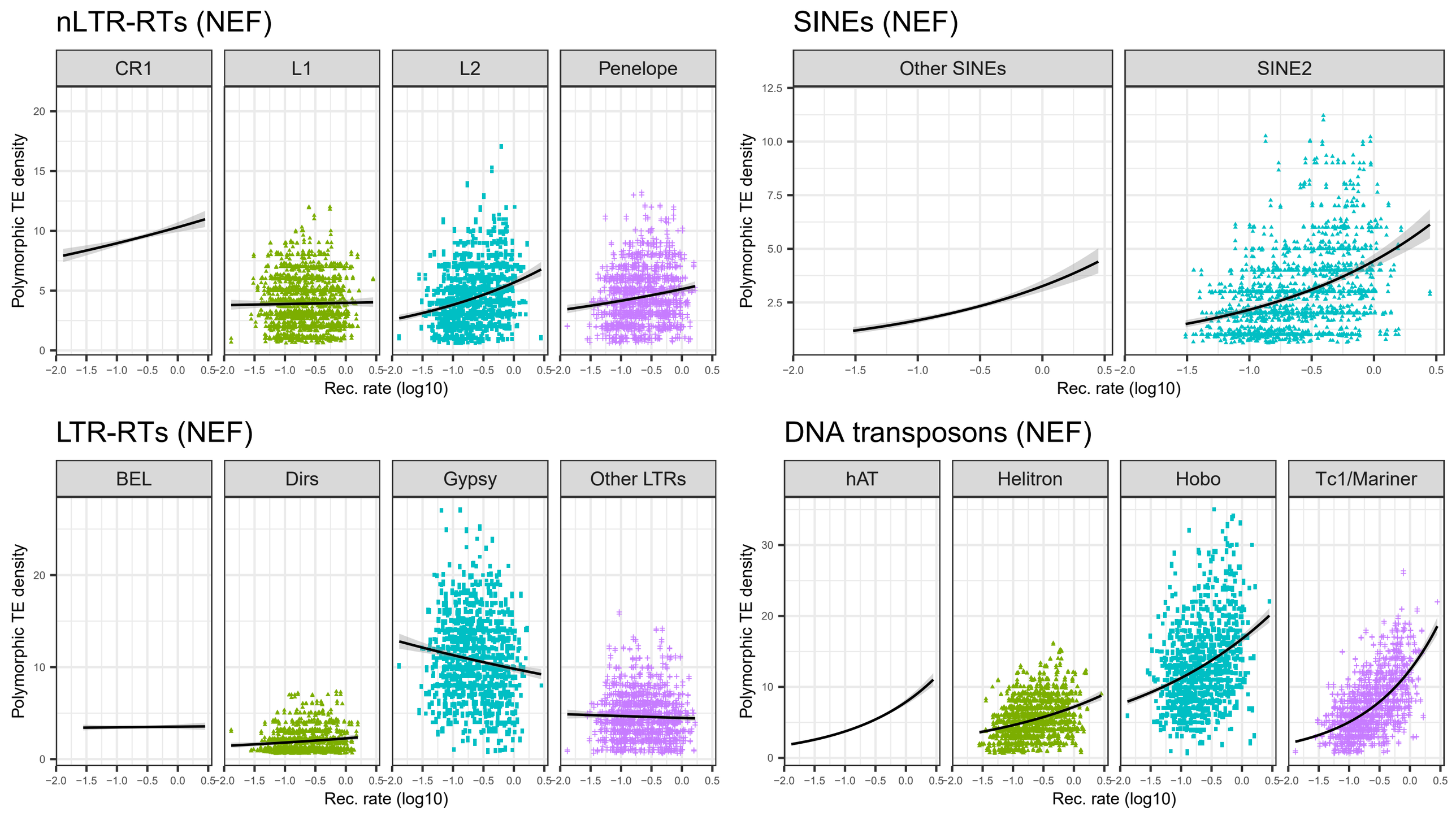

## LTR−RT and DIRS (NEF)

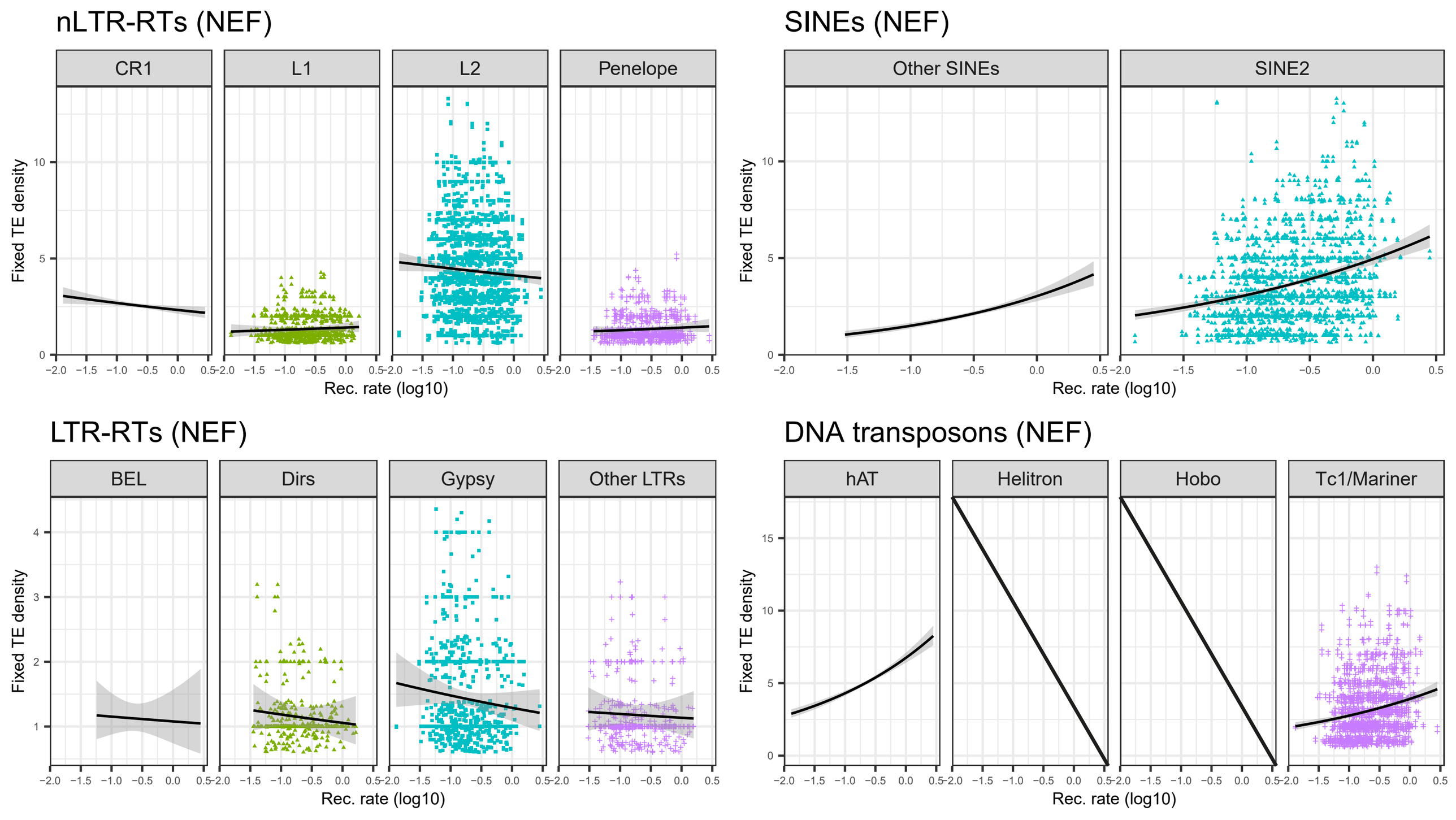



## DNA transposons (NEF)

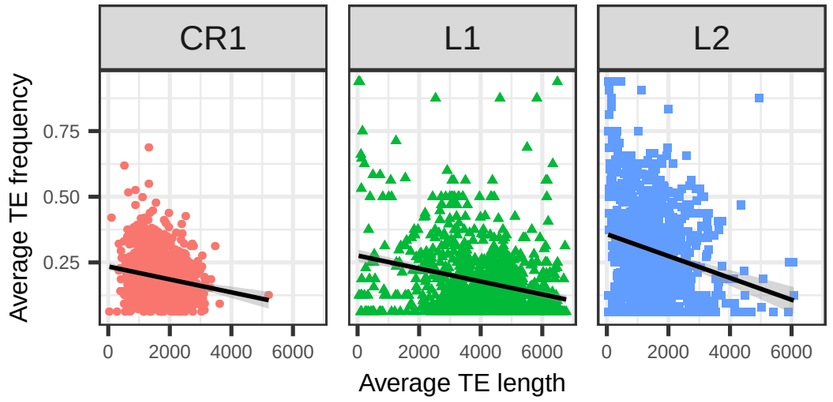



nLTR−RT (NEF)
CR1
L1
L2
Penelope
SINEs (NEF)
Other SINEs
SINE2
LTR−RTs (NEF)
BEL
Dirs
Gypsy
Other LTRs
DNA transposons (NEF)
hAT
Helitron
Hobo
Tc1/Mariner
Average TE frequency
Rec. rate (log10)

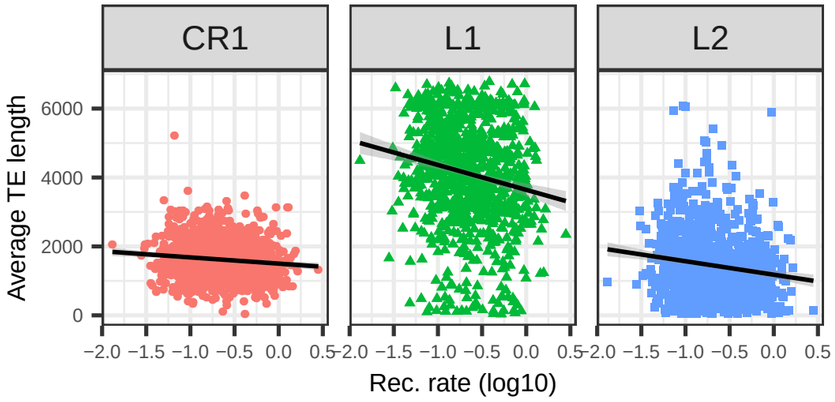

nLTR-RTs (NEF)
CR1
L1
L2
Penelope
SINEs (NEF)
Other SINEs
SINE2
LTR-RTs (NEF)
BEL
Dirs
Gypsy
Other LTRs
DNA transposons (NEF)
hAT
Helitron
Hobo
Tc1/Mariner
Polymorphic TE density
Rec. rate (log10)


nLTR-RTs (NEF)
CR1
L1
L2
Penelope
SINEs (NEF)
Other SINEs
SINE2
LTR-RTs (NEF)
BEL
Dirs
Gypsy
Other LTRs
DNA transposons (NEF)
hAT
Helitron
Hobo
Tc1/Mariner
Fixed TE density
Rec. rate (log10)

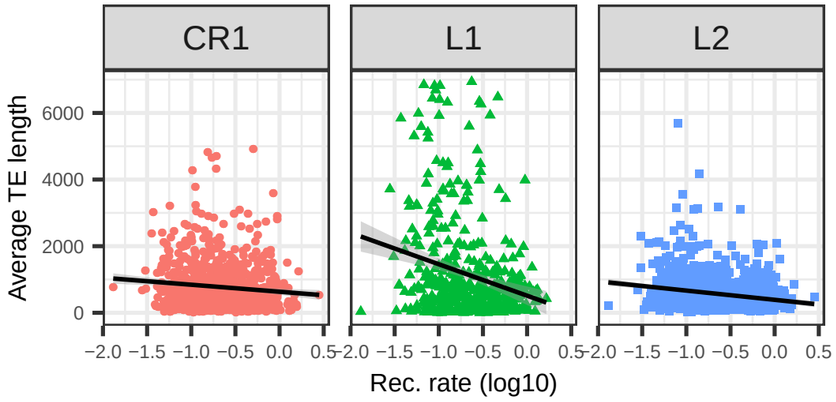
Frequency v LINEs length (EF)



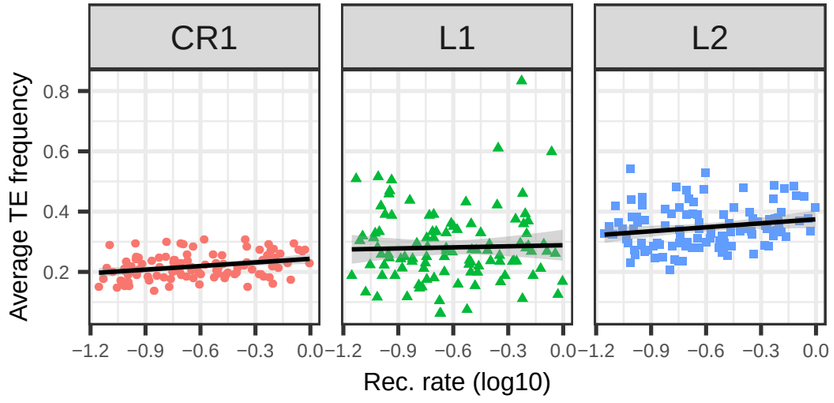

Length of polymorphic LINEs (EF)



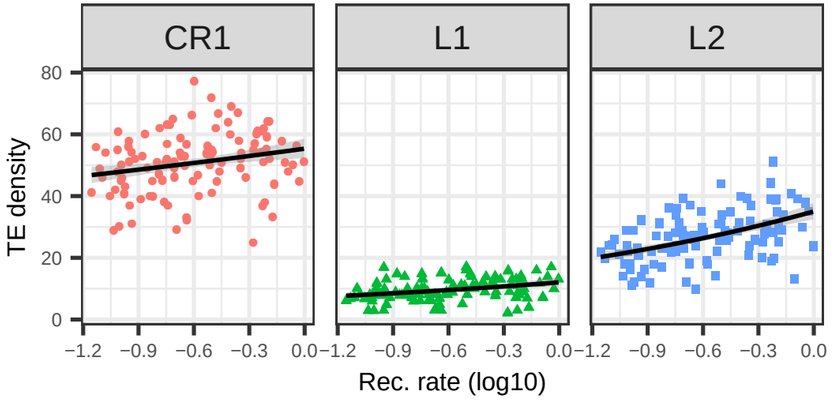

Length of fixed LINEs (EF)



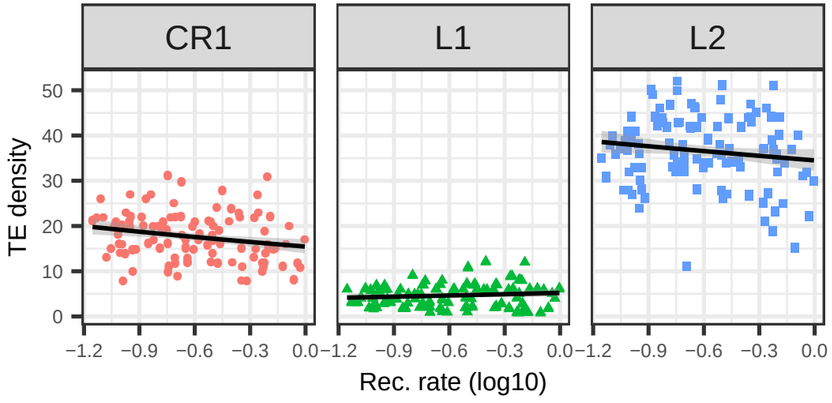

Frequency of short polymorphic LINEs (EF)



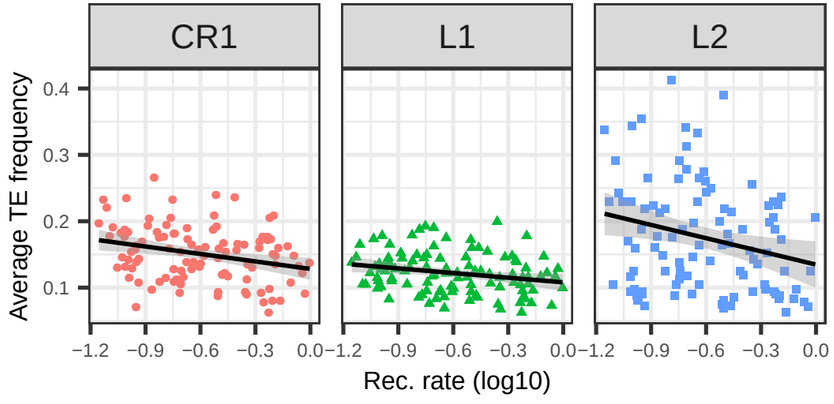

Density of short polymorphic LINEs (EF)



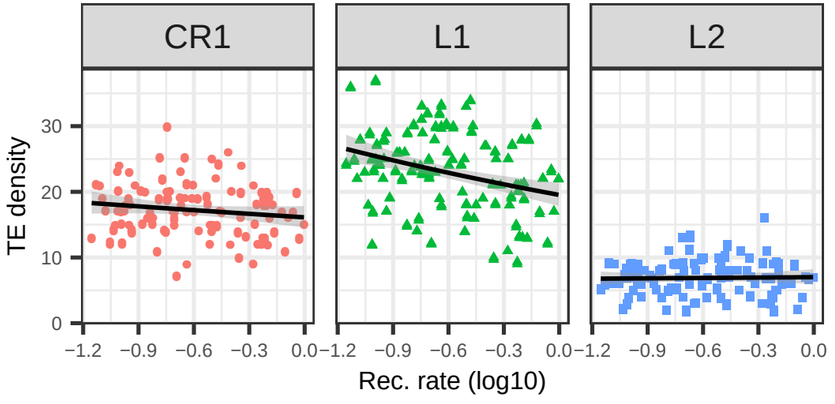

Density of short fixed LINEs (EF)



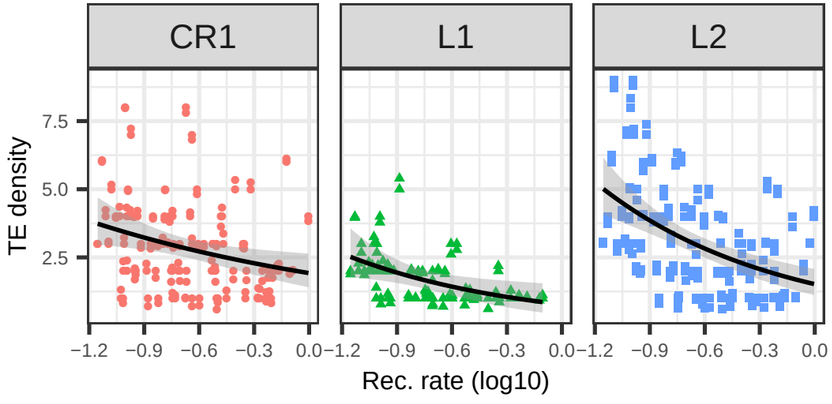

Frequency of long polymorphic LINEs (EF)




Density of long polymorphic LINEs (EF)




Density of long fixed LINEs (EF)



Frequency of polymorphic insertions

Density of polymorphic insertions

Density of fixed insertions

pref:0.5/0.5, constant

pref:0.5/0.5, burst

pref:0.7/0.3, constant

pref:0.7/0.3, burst

A

B

C

D

Average frequency

Density of insertions

Position

1e+06

2e+06

3e+06

4e+06

selection

s(HR)=−0.005, s(LR)=−0.0005

s(HR)=s(LR)=−0.0001