Bursts of recent transposable element activity and signatures of element removal in the monarch genome, *Danaus plexippus*

Tobias Baril[1,2]* and Alexander Hayward[1,3]*

[1] Centre for Ecology and Conservation, University of Exeter, Penryn Campus, Cornwall, TR10 9FE, UK

[2] ORCID: 0000-0002-5936-7531

[3] ORCID: 0000-0001-7413-718X

* Corresponding Author information:

Tobias Baril: tb529@exeter.ac.uk

Dr Alexander Hayward: alex.hayward@exeter.ac.uk

Abstract

Background

Lepidoptera (butterflies and moths) are an important model system in ecology and evolution. A high-quality chromosomal genome assembly of the monarch butterfly (*Danaus plexippus*), famous for its North American migration, is available but lacks an in-depth transposable element (TE) annotation. This provides an opportunity to explore host-TE interactions, and the impact TEs have in shaping the monarch genome.

Results

6.47% of the monarch genome is comprised of TEs, a reduction of 6.59% compared to the original TE annotation performed on the draft genome assembly. TE content is low compared to two closely related species, *Danaus chrysippus* (26.70%) and *Danaus melanippus* (11.87%). The biggest contributors to genome size in the monarch are LINEs and *Penelope*-like elements, and 37.7% of TE content is contributed by five newly identified TE families (two LINE, two *Penelope*-like, and one SINE). Some young DNA TE families show similar activity profiles to these LINEs, with their success putatively due to horizontal transposon transfer from species sharing the same environment. There are several recent peaks of TE activity in the monarch, with little evidence for peaks of activity more anciently. LINE fragments demonstrate signatures of genomic deletions as reported by studies on *Heliconius* butterflies, indicating a high rate of TE turnover. Given previous associations in other species, we investigated the association of TEs with wing colouration and immune genes. We find a single unclassified element 7kb upstream of the myosin gene locus, associated with wing colouration, and 49 immune genes with TEs within 5kb upstream of the transcription start site, presenting the potential for the involvement of TEs in regulatory functions.

Conclusions

We provide an in-depth TE annotation and analysis of TE diversity and evolution for the monarch genome. We identify highly successful novel DNA TE families, mirroring the activity profile of the most successful LINEs. We also find evidence of ongoing TE expansion and removal in the monarch, highlighting the dynamic nature of repeat content in genomes over time. Further in-depth comparative studies across closely related species will be beneficial to our understanding of the evolutionary dynamics of TEs and the processes leading to their contrasting distributions.

Keywords

Transposon; TE Annotation; Lepidoptera; *Danaus plexippus*; Butterfly; Genomic Deletion; TE-Host Dynamics; Genome Evolution

Background

Transposable elements (TEs) are autonomous DNA sequences that can move within the genome. TEs are present in nearly all eukaryotes, and are important in shaping their genomes [1–8]. TEs also represent a large reservoir of sequences with the potential to contribute to host genomic novelty, and they have been implicated in the evolution of regulatory networks [3,9–13], chromosomal rearrangements [3,14–19], exon shuffling [20–22], and donation of coding sequence [20,22–27].

Due to the great diversity and dynamic nature of TEs, species-specific TE libraries and accompanying TE annotations are required to provide an understanding of their evolution and impact. By examining differences in the diversity and abundance of TEs within a genome, as well as their integration histories, proliferation frequencies, and rate of removal, we can begin to unravel the factors that govern host-TE evolutionary dynamics. Curating TEs in a wide range of species is also an important step more generally, to expand

understanding of the interplay between TEs and host genomes, and consequences for host evolution [28]. Meanwhile, from a practical perspective, accurate and reliable repeat annotation is essential to provide quality genome annotations, ensuring that repeat sequences are not incorrectly annotated as host genes and vice versa, especially given the vast numbers of genomes currently being sequenced across the tree of life.

Lepidoptera (butterflies and moths) are an important model system in ecology and evolution. Lepidopterans have captivated researchers for decades, proving a fascinating and practical system for research into topics such as development, physiology, host-plant interactions, coevolution, phylogenetics, mimicry, and speciation (e.g [29–33]). Over recent decades, Lepidoptera has become an important model system in genomics especially. For example, Lepidoptera genomes have been pivotal in investigations into the genomic basis of migration [34–36], warning colouration [35], adaptive radiations [37], and evolutionary studies into TEs, including TE horizontal transfer [38–40], the influence of TEs on genome size [41], and the role of a TE in the evolution of industrial melanism in the peppered moth [42].

The monarch butterfly, *Danaus plexippus*, is famous for its incredible North American migration, global dispersal, and orange warning colouration [35]. The monarch has been the focus of much research and broad public interest for decades, with a draft genome released in 2011, and work during the last decade revealing the genetic bases of several key traits [35,36]. Most recently, a high-quality chromosomal assembly of the *D. plexippus* genome has been generated [43], providing a powerful opportunity for a detailed analysis of TEs in the monarch genome. Despite the monarch being an important research focus, we lack a detailed TE annotation for the species, and there is little understanding of host-TE interactions, or the impact that TEs have had on shaping the monarch genome more generally.

Here, we provide a detailed annotation of TEs within the monarch genome, including analyses of TE diversity and abundance, interactions with host genes, genomic distribution

including hot and coldspots, and TE evolution. We also compare the efficacy of a manual and automated *de novo* TE library curation method for repeat annotation. We find evidence of considerable recent TE activity in the monarch, with two novel Tc1-Mariner element families proliferating at a high rate over recent timescales (*danPle_TcM-1* and *danPle_TcM-2*). In contrast, we find a pattern of high proliferation rates of LINE and *Penelope* elements was historically dominant, and these elements still characterise the monarch TE landscape. We also reveal the role of large genomic deletions in protecting the host genome from costly LINE TEs, which show signatures of removal. Overall, we provide an overview of the broadscale dynamics that shape the TE landscape of the monarch genome, and provide a high-quality, curated, and species-specific TE annotation library for use by the scientific community.

Results

Performance of automated and manual curation

We compared a basic manual annotation method with an iterative automated curation method to assess their relative performance. Repetitive element quantifications using both the manual and automatically curated libraries are presented in Table 1.

Repeat Class	Automated *de novo* Library (+ RepBase and Dfam)			Manual *de novo* Library (+ RepBase and Dfam)			Delta (Automated - Manual)		
	No. elements	Total Length (Mb)	Percentage sequence (%)	No. elements	Total Length (Mb)	Percentage sequence (%)	No. elements	Total Length (Mb)	Percentage sequence (%)
Retroelement	31716	9.90	3.98%	21673	7.93	3.19 %	10043	1.97	0.79%
SINE	3977	0.90	0.36%	3338	0.85	0.34 %	639	0.05	0.02%
LINE	12165	5.20	2.09%	10011	4.77	1.92	2154	0.43	0.17%

						%			
Penelope	14532	3.14	1.26%	7770	1.86	0.75 %	6762	1.27	0.51%
LTR element	1042	0.66	0.27%	554	0.45	0.18 %	488	0.22	0.09%
DNA transposon	4497	2.37	0.95%	3443	2.11	0.85 %	1054	0.25	0.10%
Rolling-circle	5454	1.79	0.72%	2928	0.90	0.36 %	2526	0.88	0.36%
Unclassified	6658	1.86	0.75%	11564	3.34	1.34 %	-4906	-1.47	-0.59%
Other	555	0.17	0.07%	359	0.15	0.06 %	196	0.02	0.01%
Total repeats	48880	16.08	6.47%	39967	14.43	5.80 %	8913	1.65	0.66%

Table 1. Repeat content in the monarch using the manual and automated *de novo* library curation methods. “+ RepBase and Dfam” indicates annotations performed with the *de novo* library combined with the *Arthropoda* RepBase library and Dfam.

The automated *de novo* curated library contained 115 consensus sequences (59% classified), whilst the manual *de novo* curated library contained 75 consensus sequences (35% classified) (Datasets S1 and S2; Additional Files 1 and 2). A similar number of identified TE families remained unclassified during the RepeatMasker step using both methods, with 47 (41%) and 49 (65%) unclassified putative TE families being curated under the automated and manual methods, respectively.

In general, consensus sequences in the manually-curated library were shorter than their counterparts in the automated library (mean TE consensus length in automated library = 3,534bp; mean TE consensus length in manual library = 1,236bp). For sequences initially identified as SINEs by RepeatModeler, an automated version of the BLAST, Extract, Extend

process [28] resulted in reclassification of these as LINEs (Table S1; Additional File 3), whilst they remained classified as SINEs following manual curation. These sequences were further interrogated, and it was revealed that these novel families belonged to the HaSE2 SINE family, which were misclassified due to their significant homology with RTE LINEs [44]. The misclassification of these elements by RepeatMasker highlights one of the main issues with homology-based TE detection methods, which often miss, or misclassify, SINEs due to high sequence variation and lack of terminal repeats [45]. We find that further issues remain with all-by-all BLAST methods, as used by RepeatMasker and RepeatModeler, for the identification of putative TE families. When comparing the manual and automated *de novo* libraries, certain consensus sequences shared extremely high levels of similarity. For example, in one case, the "*rnd-1_family-48#LINE/Penelope*" family from the automated library matched both the "*rnd-1_family-39#Unknown"* and the "*rnd-3_family-30#Unknown*" families in the manual library with the same percentage identity and match length (Table S2; Additional File 3). The key difference between these two matches was that one matched in the forward direction and the other in the reverse. This situation was uncovered for two cases where a single sequence from the automated library matched two sequences in the manually curated library (Table S2; Additional File 3). To resolve this issue, cd-hit-est [46,47] was used to cluster matching sequences from opposite strands (For discussion, see Additional File 18).

When using the different *de novo* libraries, there was no significant difference in TE coverage (One-Way ANOVA, F = 0.067, p = 0.799). (Figure S1; Additional File 4). In total, 1.65Mb more sequence was annotated as TE sequence using the automated library than the manual library (Table 1). 807.40kb (5.02% of total TE content) of TE loci annotated with the automated library lacked annotations using the manual library (Tables S3 and S4; Additional File 3), with the largest contribution from rolling circle elements (Tables S3 and S4; Additional File 3). A total of 176.16kb (1.22% of total TE content) of TE loci annotated with the manual library lacked annotations using the automated library, with the largest

contribution from unclassified elements (Table S3 and S4; Additional File 3). Whilst the difference in TE annotation coverage was not significant, the increase in TE coverage when using the automated library is of a scale that could be consequential when assessing the impact of TE insertions on host genomes. Given the similarity in annotation results between the two libraries, and the generally longer consensus sequences achieved using the automated *de novo* library (Figure S2; Additional File 5), the annotation results from the automated *de novo* curation pipeline were used for all downstream analyses in this study.

Transposable Element Landscape

Repeats are a major determinant of genome size [48–53]. However, in the monarch, we find that repeats comprise just 6.47% of total genome size (Assembly Scaffold N50 = 9,210kb, 97.5% Complete Busco orthologs) (Figure 1, Table 1). This differs considerably in comparison to the TE annotation performed on the original monarch genome assembly, where repeats were estimated to cover 13.06% of the genome (Table S5; Additional File 3) [36]. When examining the source of this variation, the major difference in coverage can be attributed to unclassified elements, where we find that coverage of unclassified elements reduced by 16.47Mb (Table S5; Additional File 3). This is likely due to improvements in TE databases and annotation tools over the last decade, particularly given that the original annotation used similar tools and databases to those applied in this study [36], in addition to improved repeat detection due to increased genome assembly quality. Consistent with this, a standard ‘no frills’ RepeatMasker run using the *Arthropoda* database and Dfam identified a repeat content of 5.60% (Table S6; Additional File 3). Thus, the sequences annotated as unclassified in the original annotation were most likely not TE sequences and are no longer annotated as such by current annotation tools. Nevertheless, we are unable to confirm this for certain, as the original *de novo* repeat library generated for the draft genome is unfortunately unavailable.

We annotated the two additional genomes available for species in the *Danaus* genus using the same automated TE annotation method that we applied to the monarch. In comparison, the monarch exhibits a lower TE content, with repetitive elements covering 26.70% of the *D. chrysippus* genome (Scaffold N50 = 1,465kb, 93.5% Complete Busco orthologs), and 11.87% of the *D. melanippus* genome (Scaffold N50 = 890kb, 86.0% Complete Busco orthologs) (Figure S3, Table S5; Additional Files 6 and 3). We suggest that the reduced TE load observed in the monarch may have arisen through one of two mechanisms: (i) loss of repeat content in the monarch, or, (ii) gains in repeat content in both *D. chrysippus* and *D. melanippus* (or in their ancestral lineages). Relatively recent TE activity is observed in all three *Danaus* species, indicated by repeats exhibiting low genetic distance from consensus (Figure 2 and Figure S3; Additional File 6). The genome sizes of both *D. chrysippus* and *D. melanippus* are also very similar, at 322Mb and 354Mb, whilst the monarch is only 248Mb. Increases in genome coverage are observed for all TE classifications for both *D. chrysippus* and *D. melanippus* in comparison to the monarch, except for a decrease in *Penelope* coverage in *D. melanippus* (Table S5; Additional File 3). Therefore, we suggest that alternative (ii) is more likely, whereby there have been independent increases in TE content which have led to increased genome sizes for both *D. chrysippus* and *D. melanippus*. The main TE contributors to increases in genome size are LINEs in both *D. chrysippus* (7.7% of genome, 24.7Mb) and *D. melanippus* (4.6% of genome, 16.2Mb) (Table S5; Additional File 3).

TEs are unevenly distributed across the monarch genome and between different chromosomes, indicated by a high standard deviation in TE base pairs per 100kb (mean = 6284bp/100kb, SD = 4183bp/100kb) (Figure 1, Figure 3, Figure 4, Table S14; Additional File 3). A larger standard deviation indicates a less even spread of TEs in a genome, where a standard deviation of 0 indicates TEs are evenly distributed across the genome. The largest density of TEs is found on chromosome 28 (10,128bp/100kb), whilst chromosome 9 exhibits

the lowest TE load (4604bp/100kb) (Table S14; Additional File 3). Across all regions, the biggest contributors to genome size are LINEs and *Penelope* elements (Figure 1).

TEs are known to be unevenly distributed between different genomic compartments, with TE densities varying between genic and intergenic space [1]. LINEs make up 2.38% of the genome, which comprises: 2.80% of gene flanking regions, 1.50% of intronic sequence, 0.93% of exonic sequence, and 3.52% of intergenic sequence. Meanwhile *Penelope* elements make up 1.18% of the genome as a whole: 1.43% of gene flanking regions ,1.13% of intronic sequence, 0.17% of exonic sequence, and 1.88% of intergenic sequence. Collectively, the greatest proportion of TEs is found in intergenic regions (total 3.86%: intergenic space 1.16%, and gene flanking regions 2.70%), in comparison to gene bodies (total 2.49%: introns 2.29%, and exons 0.20%) (Figure 1). As expected, exons are largely devoid of TE sequence, with only 1.72% of exonic sequence consisting of TEs [1,54] (Figure 1). Conversely, intronic regions have accumulated a considerably higher proportion of TE sequences, with 5.13% of total intronic sequence derived from TEs (Figure 1).

Overall, we observed a weak negative correlation between gene and TE density in the monarch genome (Spearman's rank, rho = -0.369, $p < 0.01$), suggesting regions that contain a high density of host genes are less likely to contain TEs. We identified 23 hotspots of TE accumulation in the monarch genome, defined as 100kb windows with a TE content above a 99% AUC cutoff (see Methods). This equates to windows where TE density is at least 3.09 times higher than expected (expected = 6.25kb, lowest hotspot density = 19.34kb) (Figure 4). These windows represent regions of the genome able to tolerate large numbers of TE insertions. Conversely, the presence of 38 cold spots, defined as the 100kb windows with TE content below the 1% AUC cutoff, where TE density is at least 12.56 times lower than expected (expected = 6.25kb, highest coldspot density = 498bp), highlights regions of the genome where TE insertions are likely not tolerated [54]. Of note, we detected 6 windows labelled as TE coldspots but gene hotspots (Figure 4). Genes present in hotspots that are TE coldspots are involved in RNA binding, splicing, polyadenylation, mRNA stability,

and translation initiation [55]; cell polarisation [56]; neuronal development; antenna and eye development [57–61]; and olfactory processes, which are hypothesised to be involved in social behaviour [36] (Table S12; Additional File 3). Given the importance of these functions, it is likely that TE insertion in these hotspot regions would be severely detrimental to host fitness, thus individuals where TEs insert into these regions may be non-viable and purged early in development.

Transposable Element Activity

The genome of *D. plexippus* is dominated by five relatively young and very successful novel TE families, which collectively comprise 37.7% of total TE content (Table 2, Tables S7 and S10, Figure S4; Additional Files 3 and 7).

TE Family	TE Classification	Weighted Average Age (Mya)	Copy Number
danPle_penelope-1	Penelope-like Element	10.25	7814
danPle_penelope-2	Penelope-like Element	12.84	2033
danPle_penelope-3	Penelope-like Element	10.17	2504
danPle_HaSE2-1	SINE/HaSE2	6.67	1829
danPle_r2hero-1	LINE/R2-Hero	9.91	5405

Table 2. Age and Copy number of the five TE families responsible for 37.7% of total TE content in the monarch.

In total, four families are estimated to have inserted <1Mya (Figure S4, Table S9; Additional Files 8 and 3). The low estimated age of these insertions is a sign of recent host colonisation, potentially via horizontal transposon transfer (HTT) from another host genome [1,62–65]. Peaks of low genetic distance to TE consensus sequence indicate an abundance

of closely related elements and thus recent TE proliferation (Figure 2, Table S8; Additional File 3). There has been a recent burst of activity in two Tc1-Mariner elements: *danPle_TcM-1* (499 new copies, from ~0.68Mya), and *danPle_TcM-2* (275 new copies, from ~0.34Mya), with insertion rates of 733 insertions/My and 798 insertions/My, respectively (Tables S7 and S9; Additional File 3). The identified Tc1-Mariner elements are closely related and located in Clade 130 in a recent phylogenetic study of Tc1/mariner diversity [66], which contains elements from diverse insect species, including butterflies, aphids and ants (Figure S5; Additional File 8). Ants and aphids are known to share environments with monarch larvae on milkweed plants [67], increasing the potential for such HTT events [64,68]. The insertion rates of these Tc1-mariner elements are remarkably similar to the estimated insertion rate of the most successful *Penelope* family, at 759 insertions/My, highlighting their great success within the monarch genome when compared to the average TE insertion rate of just 8 insertions/My (Figure S6; Additional File 9). This recent expansion of DNA TEs is interesting given the historic dominance of LINEs in the monarch genome (2.38% of the genome, 38% of total TE content). Whilst there are more DNA TE families than LINE families in the monarch genome (257 vs 207), LINEs have been more successful in proliferating and colonising the genome than their DNA TE counterparts to date, although the reasons for this are unknown. However, it appears that these new DNA TE families are challenging this historic trend (Figure 5, Figure S4, Figure S7; Additional Files 7 and 10).

The oldest TE insertions were estimated to be ~86 Mya, predating the emergence of the monarch ~32Mya [69], with the first invaders being rolling circle, SINE and LTR elements (Figure 2, Figure 5, Figure S4, Table S8; Additional Files 7 and 3). These TEs likely proliferated around the time that the Nymphalidae family emerged, ~91Mya (CI: 71-112Mya) [69]. However, the majority of TE families are estimated to have inserted less than ~45Mya (mean TE family age = 25.41Mya), after the speciation of the monarch (~32Mya [69]) (Figure 2). In total, 246 families (35.4%) are estimated to pre-date the emergence of the monarch, whilst 448 families (64.6%) have emerged on the monarch branch (Figure 5). TEs estimated

to have inserted <25Mya are predominantly DNA, Rolling Circle, and LINE elements (Figure 2).

Turnover of LINE TEs

LINEs are reverse transcribed from their 3' end during mobilisation, and are often incomplete at the 5' end as a consequence of the premature dissociation of reverse transcriptase (RT) or the activity of cellular RNases [70,71]. In contrast, the absence of a complete 3' region is unlikely to arise due to either process, and is evidence of a genomic deletion, presumably as a consequence of ectopic recombination between homologous LINE sequences [70,71].

Consistent with observations in *H. melpomene* [71], we find LINE fragments across all families exhibiting patterns suggestive of genomic deletions acting to remove these elements from the genome (Table 3, Figure 6, Figure S8, Figure S9; Additional Files 11 and 12). In total, 68.90% of LINE fragments contain an intact 3' region, while 31.10% of LINE fragments are truncated at the 3' end. For the *CR1*, *I-Jockey*, *L2*, and *RTE-BovB* families, most LINE fragments display patterns suggestive of genomic deletions (i.e. truncated at the 3' end), whilst the majority of fragments of the *R1*, *R2-Hero*, and *RTE-RTE* families display patterns indicative of processes associated with degradation of LINEs, such as premature disassociation of RT, or breakdown via cellular RNAse activity (Table 3).

Swathes of LINE fragments across all LINE families (mean LINE length = 411bp), suggest a high rate of turnover of LINEs within the monarch genome, in contrast to patterns observed in mammals, where low rates of DNA loss are observed [71,72]. This turnover could occur through either: (i) some form of host mechanism resulting in LINE removal, particularly if LINEs are associated with a high fitness cost, or (ii) removal and degradation of LINEs due to their particularly repetitive nature, which provides loci for ectopic recombination resulting in their partial removal from the host genome [1,19].

LINE Classification	Putative Removal Process	Fragment Types	Count	Proportion of Classification Type
CR1	Premature RT Dissociation/Activity of Cellular RNAses	Lone 3', Internal + 3'	16	0.36
	Putative Genomic Deletion	Lone 5', 5' + Internal, Internal Lacking 3'	29	0.64
I-Jockey	Premature RT Dissociation/Activity of Cellular RNAses	Lone 3', Internal + 3'	42	0.20
	Putative Genomic Deletion	Lone 5', 5' + Internal, Internal Lacking 3'	164	0.80
L2	Premature RT Dissociation/Activity of Cellular RNAses	Lone 3', Internal + 3'	61	0.20
	Putative Genomic Deletion	Lone 5', 5' + Internal, Internal Lacking 3'	244	0.80
R1	Premature RT Dissociation/Activity of Cellular RNAses	Lone 3', Internal + 3'	223	0.55
	Putative Genomic Deletion	Lone 5', 5' + Internal, Internal Lacking 3'	181	0.45
R2-Hero	Premature RT Dissociation/Activity of Cellular RNAses	Lone 3', Internal + 3'	2034	0.88
	Putative Genomic Deletion	Lone 5', 5' + Internal, Internal Lacking 3'	282	0.12
RTE-BovB	Premature RT Dissociation/Activity of Cellular RNAses	Lone 3', Internal + 3'	225	0.46
	Putative Genomic Deletion	Lone 5', 5' + Internal, Internal Lacking 3'	268	0.54
RTE-RTE	Premature RT Dissociation/Activity of Cellular RNAses	Lone 3', Internal + 3'	137	0.67
	Putative Genomic Deletion	Lone 5', 5' + Internal, Internal Lacking 3'	68	0.33

Table 3. Number of LINEs classified by fragment type, and putative removal process.

Despite evidence of removal, LINE sequence continues to dominate the TE landscape of the monarch. To be active, an intact LINE will contain start and stop codons, and an RT domain [73]. In total, we identified five intact LINEs in the monarch (Table S13,

Additional File 3), including *danPle_r2hero-1*, which is the second most populous TE in the monarch genome (Table 2). We also observe a significant accumulation of certain LINE families, with ~20% maintaining a length above 1kb (Table S11; Additional File 3). Short elements are more likely to persist in the genome than their longer counterparts, as they are less prone to recombination [19]. This, combined with the high activity level of LINEs before and throughout the evolution of the monarch (Figure 2), explains the dominance of LINE fragments across the genomic landscape, despite their high turnover [19,71].

Association of TEs with Monarch Colouration and Immune Genes

Given a previous role for TEs in wing colouration in the peppered moth (*Biston betularia*) [42] and the clouded yellow (*Colias croceus*) [74], we investigated the possibility of TE associations with wing colouration in the monarch. The myosin gene is associated with wing colouration in the monarch [75]. We identified a single unclassified element 7kb upstream of this gene (Figure S10; Additional File 12), contrasting the location of TEs implicated in colouration in the peppered moth and clouded yellow, where the TE insertions were found within an intron and 6kb downstream of the gene body, respectively [42,74]. This is not entirely unexpected, as the colouration differences seen in the peppered moth and clouded yellow are due to altered gene expression, whilst the white *nivosus* form of the monarch is due to altered pigment transport [35,42,74].

Immune genes are some of the fastest evolving genes across metazoans [76–79], and so are strong candidates for exploring the role of TEs in host evolution. Across the monarch genome, TEs of all major types were found within immune gene flanks (20kb). When normalised to compare TE enrichment around different immune gene types, we find no significant difference in TE base coverage (Kruskal-Wallis $\chi^2 = 47.17$, $p = 0.1221$), or in TE count (Kruskal-Wallis $\chi^2 = 40.351$, $p = 0.3244$) between different types of immune genes, suggesting no enrichment around specific classes of immune genes (Figure S11;

Additional File 13). We also find no significant difference in coverage or count of TE classifications around immune genes (TE count by classification: Kruskal-Wallis χ^2 = 14.46, p = 0.8493; TE coverage by classification: Kruskal-Wallis χ^2 = 125.9, p = 0.5854). However, there are TEs within 5kb of the 5' end in 49/103 identified immune genes. These proximal TEs are of interest given the short distance between them and the immune gene transcription start sites (TSS) , making regulatory functions a possibility [80].

Discussion

Here, we provide a detailed annotation of TEs in the monarch genome and an analysis of their diversity and evolution. We found a large reduction in TE sequence compared to the original annotation performed on the draft monarch genome. We suggest that this is due to unclassified TE sequences in the original annotation no longer being recognised as TEs with updated TE databases and annotation tools. However, we are unable to confirm this, due to the original *de novo* TE library and annotation results being unavailable. This raises an important issue in genomics currently. Given TE annotation is a key step in genome assembly and annotation, we suggest that it should be required to include TE libraries alongside the publication of genome projects, ideally in a specific repository, to: (i) reduce duplication of efforts required for reannotation; (ii) improve reproducibility in annotation; and, (iii) improve TE consensus libraries to reduce cases of a single TE family being given different names in multiple species. As the number sequenced genomes increases, the wider availability of TE libraries will benefit a range of studies, where gene annotations can be improved by accurate TE annotation, and the availability of large numbers of well-curated TE sequences can facilitate large-scale studies to study TE evolutionary dynamics.

We identified an uneven distribution of TEs across the genome in line with expectations of TE distribution in host genomes [1] (Figures 3 and 4). TE insertions in non-

coding regions are less likely to have a detrimental impact on host genome function than TE insertions into coding regions [1,21,54,81–83], and so are less stringently removed by selection, enabling TE accumulation in non-coding regions. Indeed, some TEs actively avoid coding regions by targeting regions of the genome unlikely to impact host fitness [54,84], with this process potentially occurring in the monarch, where we find overlapping gene hotspots and TE coldspots.

The most populous regions of the monarch genome for TEs are intergenic regions (Figure 1). The presence of large numbers of TEs in host gene flanks is interesting, given the implication of TEs in the evolution of gene regulatory networks [10–13,85], although this could also be due to the relatively compact nature of the monarch genome leaving TEs with few other loci to integrate within. TEs are implicated in rapidly generating extensive intron gain in eukaryotic genomes, where short non-autonomous DNA transposons independently generated hundreds to thousands of introns across the genomes of eukaryotes [1,54,86]. It is also hypothesised that transposon insertion-associated introns are responsible for the prevalence of nucleosome-sized exons observed in eukaryotes [86]. The presence of TEs in intronic sequence of the monarch could therefore indicate genomic refugia, where TE insertions are less likely to be detrimental to the host, but they could also indicate ongoing intron generation, where new insertions might lead to the generation of new, or longer, introns over evolutionary time.

The presence of young and highly populous TE families in the monarch genome demonstrates the success of TEs in naïve host genomes [84,87–90]. The putative recent arrival of several young DNA TEs, namely two novel Tc1-Mariner families (*danPle_TcM-1* and *danPle_TcM-2*), has led to another wave of TE proliferation. Over time, it is hypothesised that the activity of these families will decrease as the host develops mechanisms to suppress activity, either through capture of TE sequences in piRNA regions, or through other defensive mechanisms such as methylation [90–92]. A reduction in transposition activity of DDE transposase-containing DNA transposons can also be

attributed to increases in non-autonomous copies [66]. In the meantime, however, these TEs are proliferating and dynamically altering the genomic landscape of *D. plexippus*, with potential connotations for ongoing genome structural and functional evolution [1–3,19].

While factors leading to the relatively low TE content observed in the monarch compared to other *Danaus* species are currently unknown, it is clear TEs have been unable to accumulate to the same level as other members of the genus. The maintenance of a small genome in the monarch can likely be partly attributed to high turnover of TEs, evidenced by the lack of ancient peaks in repeat landscapes and the presence of few intact LINEs, many of which show hallmarks of genomic deletions. To remove LINEs, ectopic recombination between similar elements is likely to play a role, and it has been postulated that an increased rate of ectopic recombination could be responsible for the differences in accumulation of TEs in insects and mammals [93]. A lower ectopic recombination rate in mammals, despite frequent large TE loads, enables older elements to accumulate, whereas high ectopic recombination rates in insects can lead to removal of TEs before older copies accumulate in large numbers [93]. Given our results, it is likely that ectopic recombination is occurring in the monarch to remove LINEs. Indeed, the only families with an abundance of 3' fragments are, *R1*, *R2-Hero*, and *RTE-RTE* (mean insertion times ~16.02Mya, ~9.9Mya, and ~27.38Mya). It is tempting to suggest that *R1*, *R2-Hero*, and *RTE-RTE* insertions have a lower impact on host fitness, and so are "allowed" to degrade over time, rather than being removed *en masse* through genomic mechanisms. However, it must also be noted that ectopic recombination is a passive process. Whilst it is more likely that LINEs are being fragmented and removed through ectopic recombination between copies, this does not account for inter-family differences in the ratio of fragments undergoing degradation through ectopic recombination versus degradation via the activity of RT or cellular RNAses. In the case of the most populous family, *R2-Hero*, it is likely that these elements have so far avoided ectopic recombination, given these elements are young compared to other LINEs (insertion ~9.9Mya) (Figure 5, Table S8; Additional File 3)). Our results are consistent with

those of previous findings in *Heliconius* species [94], where a high TE turnover and lack of ancient TEs was observed, suggesting these characteristics might be widespread among lepidopterans, although further comparative studies will be required to confirm this.

In addition to TE turnover, genome size is likely also maintained through sequestering TE activity via epigenetic mechanisms and piRNAs [1,91,95,96]. Further research to characterise piRNA clusters in the monarch, and interrogation of its transcriptome, will further improve our understanding of how these processes have shaped the TE landscape within the monarch genome.

The lack of closer or more associations of TEs with the myosin gene associated with warning colouration in the monarch is perhaps not unsurprising, as it has been postulated that the white *nivosus* form of the monarch arises from altered pigment transport [35], rather than altered gene expression, as is the case with the peppered moth and clouded yellow [42,74]. Whilst TEs were not significantly enriched around immune genes, the prominence of TEs around immune genes is nevertheless intriguing, as is the presence of TEs <5kb from the TSS of 49 immune genes, which opens up the possibility for TEs to alter the regulation of these genes. To further understand the impact of TEs in these regions, TE knockout studies in combination with transcriptomic analyses may be able to elucidate any impacts TEs might be having on immune genes in the monarch.

Conclusions

In-depth studies into TE landscapes and dynamics within host species of broad scientific and public interest remain relatively few, even with widespread acknowledgement of the key role that TEs have played in eukaryotic evolution [1,48,97]. We provide a detailed annotation of TEs in the monarch genome and an analysis of TE diversity and evolution, including the identification of highly successful young DNA TE families, which mirror the activity profile of the most successful LINEs found in the monarch, and evidence of ongoing

TE expansion and removal in *D. plexippus*, highlighting the dynamic nature of repeat content within genomes over time. Further large-scale comparative studies will be greatly beneficial to further our understanding of the evolutionary dynamics of TEs, and the drivers leading to diverse TE landscapes across closely related species.

Materials and Methods

Transposable Element Annotation

The monarch chromosomal assembly, Dplex_v4, was downloaded from NCBI (Refseq GCF_009731565.1), along with corresponding genome annotations in GFF format [43,98]. The monarch genome assembly is 248.676Mb, with a GC content of 32.1%, and comprises 4,115 scaffolds with a scaffold N50 of 9.21Mb.

The *D. plexippus* genome was annotated using a bespoke automated in-house TE annotation pipeline, as previously described [99–104]. For this, known repeats were first identified and masked using RepeatMasker (version 4.1.0) [105] using the *Arthropoda* library from RepBase (version 23.08) and Dfam (release 3.3) [106,107]. Low complexity repeats and RNA were not masked (*-nolow and -norna*) and a sensitive search was performed (*-s*). Following this, a *de novo* repeat library was constructed using RepeatModeler (version 1.0.11), with RECON (version 1.08) and RepeatScout (version 1.0.5) [108–110]. To generate maximum-length consensus sequences for the *de novo* repeat library, novel repeats identified by RepeatModeler were curated using an automated version of the 'BLAST, Extract, Extend' process [28]. A BLASTn search was performed on the *D. plexippus* genome to obtain up to the top 40 hits for each TE family identified by RepeatModeler [111]. Sequences were extracted with 1,000 base pairs of flanking sequence at the 5' and 3' ends. Each set of family sequences were aligned using MAFFT (version 7.453) [112]. Alignments were trimmed with trimAl (version 1.4.rev22) to retain high-quality positions in the alignment (*-gt* 0.6 *-cons* 60) [113]. Updated consensus sequences were computed with EMBOSS cons (*-plurality* 3) to generate a new TE library featuring consensus sequences with extended flanks [114]. This process was repeated through 5 iterations to obtain maximum-length consensus sequences. The resulting *de novo* repeat library was combined with the RepBase *Arthropoda* library and utilised to annotate repetitive elements using RepeatMasker with a score threshold at the relatively conservative level of 400 (*-cutoff* 400), to exclude poor

matches unlikely to be true TE sequences. Following this, the *D. plexippus* genome was analysed with LTR_Finder (version 1.07), using the LTR_Finder parallel wrapper [115,116] to identify full-length LTR elements. Final TE annotations were defragmented and refined using the loose merge (*-loose*) command in RepeatCraft [117]. Overlapping annotations were removed using MGKit (version 0.4.1) filter-gff, with the longest element of overlapping annotations retained (*-c* length *-a* length) [118]. Finally, all repeats less than 100bp in length were removed before the final TE quantification to decrease spurious hits. Final repeat annotation file is provided in dataset S3 (Additional File 17).

To assess the performance of the automated annotation method, a manual curation of the *de novo* library was also performed. Briefly, for this, the 'BLAST, Extract, Extend' process was performed manually, rather than using the automated pipeline. For each iteration, up to the top 40 hits for each TE family were extracted from the *D. plexippus* genome using BLASTn [111]. Sequences were extracted with 1,000 base pairs of flanking sequence at the 5' and 3' ends, before each set of family sequences was aligned using MAFFT [112]. Alignments were then visualised using AliView, and poorly-aligned positions and ends were trimmed where single-copy DNA was present [119]. If either the 5' or 3' end of the alignment terminated within the repetitive region, the BLASTn and extension process was repeated until single-copy DNA was observed at both the 5' and 3' ends of the alignment for each family, signalling the end of the repeat and the transition to host genome sequence. New consensus sequences were generated using EMBOSS cons to create a new TE library of full-length elements. The presence of duplicate consensus sequences was checked with *cd-hit-est* using stringent parameters (*-n* 10 *-c* 0.95) and any duplicates were removed, retaining the longest consensus sequence only [46,47]. This manually curated *de novo* library was combined with the RepBase *Arthropoda* library and utilised to provide a separate TE annotation for the monarch genome. Full-length LTR elements were again annotated using LTRFinder in the LTRFinder parallel wrapper and the final TE annotations

were defragmented and refined using RepeatCraft loose merge. The same post-processing was applied to the annotations as with the automated curation.

Comparison of results from automated and manually curated libraries

The automated and manually curated libraries were compared to assess the performance of the automated curation process. Final annotation quantifications were compared using a One-way ANOVA, as observations were normally distributed for both groups (Automated annotation: Shapiro-Wilk, $W = 0.88401$, $p > 0.05$; Manual annotation: Shapiro-Wilk, $W = 0.90517$, $p > 0.05$). Comparisons were performed in R (version 3.6.3) using the Rstudio IDE [120,121]. As differences in quantification between the two methods were not statistically significant (see Results), the annotations from the automated process were used for downstream analyses to maximise the reproducibility of this study. For discussion, see [Additional File 18].

Identifying Transposable Element Association with Genomic Compartments

The *D. plexippus* gene annotation file was processed to generate a GFF file containing coordinates of exons, introns, and 5' and 3' flanking regions. Gene flanking regions are defined here as 20kb directly upstream or downstream of the gene body. These flanking regions are determined to identify TEs at a distance that could be described as the proximate promoter region, rather than just accounting for the core promoter region, to include both promoter and more distal enhancer regions for genes [80]. Bedtools intersect (version 2.27.1) was used to identify the number of bases of overlap (*-wao*) between TEs and different genomic compartments [122]. Following this, quantification was performed in R.

TE Hotspots and Coldspots

To identify TE and gene hotspots and coldspots, chromosomal scaffolds were split into 100kb windows, which were defined using the frequency distribution of observations. A polynomial model was fitted to the frequency distribution, to generate a smooth curve through the data points using the lm function in R. Using these models, the total area under the curve (AUC) was calculated using the area_under_curve function of the bayestestR package [123]. From this, the 1% and 99% AUC cutoff values were determined. Starting with the first data point and iteratively adding subsequent datapoints with each round, the AUC for each collection of points was determined. The maximum AUC value not exceeding 99% of the total AUC was taken as the cutoff where windows containing higher TE or gene coverage were considered to be hotspots. The minimum AUC value not exceeding 1% of the total AUC was taken as the cutoff where windows containing lower TE or gene coverage were considered to be coldspots (Figure S12; Additional File 15). Karyoplots of the *D. plexippus* genome, illustrating gene and TE locations, highlighting TE hotspots and coldspots were generated using the KaryoploteR package in R [124].

Estimating TE Age and Activity

TE activity dates were estimated using Kimura2-parameter genetic distances (including CpG sites) calculated between individual insertions and the consensus sequence of each TE family [125]. A neutral mutation rate of 0.0116 substitutions per site per million years was used for *D. plexippus*, as previously described for monarch [126].

Turnover of LINEs

To interrogate LINE decay patterns, each LINE sequence was extracted from the genome using bedtools getfasta [122]. LINE sequences were used to query the repeat library using BLASTn with tabular output specified (*-outfmt* 6). In the event of multiple matches, only the top match was retained (defined by percent identity). Coordinates of each

match were scaled to normalise the length of each LINE consensus sequence. Mapped LINE fragments were plotted using ggplot2 of the tidyverse package in R [127,128].

Identification of Intact LINEs

To identify full-length LINEs, all elements longer than 2,700bp were extracted from the monarch genome using bedtools getFasta [122]. Open reading frames (ORFs) were detected using EMBOSS getORF [114]. ORF sequences were filtered to retain all sequences > 600aa. RPSBLAST+ (v2.9.0+) [111] was used to query the ORF sequences against the CDD database [129] to identify conserved protein domains. Sequences were classified as intact LINEs based on the presence of relevant protein domains as previously described [130] (Table S13, Additional File 3).

Classification of Novel DNA TEs

To further classify the novel DNA elements in the monarch beyond homology to known elements, as used by RepeatModeler. DDE transposases were compared with curated alignments for all DDE TE superfamilies [66]. Specifically, dashes in curated alignments of all known DDE TE sequences for each of the 19 families were removed, and the resultant fasta file was used as the subject in a blastx search against the *de novo* monarch TE library, with an e-value threshold of $1x10^{-20}$ and tabular output specified (-*evalue* 1e-20 -*outfmt* 6) [111]. The best match for each *de novo* query sequence was retained, with a minimum 50% identity over at least 50% of the total length of the curated DDE subject sequence. Following this, novel DNA elements were only detected for the Tc1-Mariner and Sola1 families (Figures S5 and S13; Additional Files 8 and 16). Transposase sequences for the *de novo* monarch sequences were extracted and aligned to the original DDE alignments for their respective families using the profile alignment command in MAFFT [112]. Sequences from related transposon superfamilies were used as outgroups, as

described [66]. To estimate phylogenies for the two DDE superfamilies, we performed 1,000 ultra-fast bootstrap replicates in IQ-TREE (v1.6.12) [131] specifying the best fit amino acid model for each family as indicated by ModelFinder [132]. Phylogenetic trees were visualised in figtree (v1.4.4) (http://tree.bio.ed.ac.uk/software/figtree/).

Association of TEs with Monarch Colouration and Immune Genes

We obtained the transcript sequence in fasta format for the myosin gene responsible for wing colouration from MonarchBase, under accession DPOGS206617 [34]. To identify the matching region in the v4 monarch assembly, a blastn search was performed to query the monarch assembly [111]. The coordinates of the myosin gene were obtained and used in a bedtools window search to identify TEs within 20kb (*-w* 20000)[122]. A karyoplot of the myosin gene region including 20kb flanks was generated in R using the karyoploteR package [124].

To investigate associations between monarch immunity genes and TEs, nucleotide and protein sequences of monarch immune genes were obtained in fasta format from MonarchBase [34]. To identify corresponding regions in the v4 monarch assembly, blastn and tblastn searches were used to query the monarch genome assembly with the monarch immune gene nucleotide and protein sequences [111]. Matches were filtered by percentage identity and query match coordinates to identify the locus of each immune gene. A bed file was generated, containing the name and location of each immune gene, and a bedtools window search was performed to identify TEs within 20kb of each immune gene(*-w* 20000)[122]. Karyoplots of each immune gene region including 20kb flanks were generated in R using the karyoploteR package [124]. TE bases surrounding each immune gene were quantified in R [120], and normalised to account for variation in the number of each gene type. As TE coverage and count were not normally distributed, a Kruskal-Wallis test was used to compare TE coverage and count between different immune gene types. Unless

otherwise stated, all plots were generated using the ggplot2 package of the tidyverse suite in R [127,128].

Figure Legends

Figure 1. TE content and location in the monarch. Major TE types are represented by different colours indicated in the key (A) Pie chart illustrating proportions of the monarch genome comprised of the main TE classifications. (B) Bars illustrating the number of base pairs of the monarch genome found in different genomic compartments. (C) Bars illustrating the proportion of each genomic compartment derived from bases annotated as TE sequence.

Figure 2. Repeat landscape of the monarch. The x axis indicates the level of Kimura2-parameter genetic distance observed between TE insertions and their respective consensus sequences in percent. More recent elements are located to the right of the x axis. The y axis indicates the percentage of the genome occupied by TE insertions of each genetic distance class.

Figure 3. Violin plot showing kilobases of TE sequence per 100kb window across the genome for each of the main TE classifications. Red points show the mean kilobases of TE sequence per 100kb.

Figure 4. Karyoplots illustrating monarch chromosomes, with genes represented as black bars. (A) Above each chromosome, TE density is shown for 100kb windows, with major TE types represented by different colours indicated in the key. (B) Above each chromosome, TE

hotspots and coldspots are shown, with severity represented by the different colours indicated in the key,

Figure 5. Violin plot illustrating the estimated insertion time (Mya) for each of the major TE types, represented by different colours indicated in the key. The black bar shows the estimated divergence time of the monarch, approximately 32Mya [69].

Figure 6. Plot illustrating LINE fragments annotated in the monarch genome. The x axis indicates normalised start and end coordinates for each LINE element. Elements are organised from longest to shortest, with short elements found towards the top of the plot. LINE families are represented by different colours indicated by the key.

Additional Files

Additional File 1: Dataset S1. *De novo* transposable element consensus sequences identified using the automated curation method, in FASTA format.

Additional File 2: Dataset S2. *De novo* transposable element consensus sequences identified using the manual curation method, in FASTA format.

Additional File 3: Supplementary Tables, including contents page with table legends. Tables S1-S13.

Additional File 4: Figure S1. Set of plots comparing the final TE quantifications between the manual and automated *de novo* libraries, both in conjunction with RepBase *Arthropoda* library and Dfam. Major TE types represented by different colours, indicated in the key. Top row: Stacked bar showing number of elements identified with each library. Scatter plot showing number of elements identified with each library, lines indicate matching TE classifications. Box plot showing number of elements identified with each library. Middle row: Stacked bar showing base pair coverage for TEs annotated with each library. Scatter plot showing base pair coverage for TEs annotated with each library, lines indicate matching TE classifications. Box plot showing base pair coverage for TEs annotated with each library. Bottom row: Stacked bar showing proportion of monarch genome comprised of TEs annotated with each library. Scatter plot showing proportion of monarch genome comprised of TEs annotated with each library, lines indicate matching TE classifications. Box plot showing proportion of monarch genome comprised of TEs annotated with each library.

Additional File 5: Figure S2. Scatter plots showing the length of consensus sequences identified in the manual and automated curation methods. Lines join consensus sequences identical prior to BLAST, Extract, Extend process. Major TE types are represented by different colours indicated in the key. Orphaned points represent consensus sequences only found in one of the two libraries at the end of the BLAST, Extract, Extend process.

Additional File 6: Figure S3. Repeat landscape for *Danaus chrysippus* and *Danaus melanippus*. The x axis indicates the level of Kimura2-parameter genetic distance observed between TE insertions and their respective consensus sequences in percent. More recent elements are located to the right of the x axis. The y axis indicates the percentage of the genome occupied by TE insertions of each genetic distance class.

Additional File 7: Figure S4. Scatter plot illustrating estimated TE age for each TE family. Major TE types are represented by different colours indicated in the key. Elements are ordered from oldest to youngest, within each major type.

Additional File 8: Figure S5. Phylogenetic tree of Tc1-Mariner DDE transposases, with novel monarch families added (highlighted in red). Branch support from 1000 ultra-fast bootstrap repetitions. Novel Tc1-Mariner families highlighted in teal are the young TEs found with extremely high copy number.

Additional File 9: Figure S6. Scatter plots illustrating the number of insertions per million years for each TE family. Major TE types are represented by different colours indicated in the key.

Additional File 10: Figure S7. Scatter plots illustrating TE age against copy number. Major TE types are represented by different colours indicated in the key. X axis shows estimated age of each TE family, Y axis shows TE copy number identified in the monarch.

Additional File 11: Figure S8. Plots illustrating LINE fragments annotated in the monarch genome, separated by LINE family. The x axis indicates normalised start and end coordinates for each LINE element. Elements are organised from longest to shortest, with short elements found towards the top of the plot. LINE families are represented by different colours indicated by the key.

Additional File 12: Figure S9. Bar plots illustrating the abundance of LINE fragments identified in the monarch. LINE classifications are represented by different colours as indicated in the key.

Additional File 13: Figure S10. Karyoplot of the monarch gene association with colouration, myosin, obtained from MonarchBase under accession DPOGS206617. Main TE types within 20kb are represented by different colours indicated in the key.

Additional File 14: Figure S11. Boxplots illustrating: (A) Normalised TE coverage in 20kb regions surrounding each immune gene type. (B) Normalised TE count in 20kb regions surrounding each immune gene type. Main TE types are represented by different colours indicated in the key.

Additional File 15: Figure S12. Plots illustrating: (A) Distribution of genome windows with TE coverage, with lines showing 1% and 99% AUC cutoffs used to define hotspots and coldspots. (B) Distribution of genome windows with host gene coverage, with lines showing 1% and 99% AUC cutoffs used to define hotspots and coldspots.

Additional File 16: Figure S13. Phylogenetic tree of Sola1 DDE transposases, with novel monarch families added (highlighted in red). Branch support from 1000 ultra-fast bootstrap repetitions.

Additional File 17: Dataset S3. Coordinates and classification of all transposable element sequences identified in the monarch genome using the automated *de novo* library in conjunction with RepBase *Arthropoda* library and Dfam, in bed format.

Additional File 18: Discussion on the comparison between our automated annotation method and a basic manual curation.

Abbreviations

TE: Transposable element

Non-LTR: Non-long terminal repeat

LTR: Long terminal repeat

LINEs: Long INterspersed Elements

SINEs: Short INterspersed Elements

TSS: Transcription Start Site

References

[1] Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biol 2018;19:199. https://doi.org/10.1186/s13059-018-1577-z.

[2] Klein SJ, O'Neill RJ. Transposable elements: genome innovation, chromosome diversity, and centromere conflict. Chromosom Res 2018;26:5–23. https://doi.org/10.1007/s10577-017-9569-5.

[3] Feschotte C, Pritham EJ. DNA Transposons and the Evolution of Eukaryotic Genomes. Annu Rev Genet 2007;41:331–68. https://doi.org/10.1146/annurev.genet.40.110405.090448.

[4] Belyayev A. Bursts of transposable elements as an evolutionary driving force. J Evol Biol 2014;27:2573–84. https://doi.org/10.1111/jeb.12513.

[5] Schrader L, Schmitz J. The impact of transposable elements in adaptive evolution. Mol Ecol 2018. https://doi.org/10.1111/mec.14794.

[6] Casola C, Lawing AM, Betrán E, Feschotte C. PIF-like transposons are common in Drosophila and have been repeatedly domesticated to generate new host genes. Mol Biol Evol 2007;24:1872–88. https://doi.org/10.1093/molbev/msm116.

[7] Jangam D, Feschotte C, Betrán E. Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. Trends Genet 2017;33:817–31. https://doi.org/10.1016/j.tig.2017.07.011.

[8] Volff JN. Turning junk into gold: Domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays 2006. https://doi.org/10.1002/bies.20452.

[9] Mita P, Boeke JD. How retrotransposons shape genome regulation. Curr Opin Genet Dev 2016;37:90–100. https://doi.org/10.1016/j.gde.2016.01.001.

[10] Bourque G. Transposable elements in gene regulation and in the evolution of

vertebrate genomes. Curr Opin Genet Dev 2009;19:607–12. https://doi.org/10.1016/j.gde.2009.10.013.

[11] Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet 2008;9:397–405. https://doi.org/10.1038/nrg2337.

[12] Rebollo R, Romanish MT, Mager DL. Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. Annu Rev Genet 2012;46:21–42. https://doi.org/10.1146/annurev-genet-110711-155621.

[13] Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: From conflicts to benefits. Nat Rev Genet 2017;18:71–86. https://doi.org/10.1038/nrg.2016.139.

[14] Bennetzen JL. Transposable element contributions to plant genome evolution. Plant Mol Biol 2000;42:251–69. https://doi.org/10.1023/A:1006344508454.

[15] Casals F, Cáceres M, Ruiz A. The Foldback-like transposon Galileo is involved in the generation of two different natural chromosomal inversions of Drosophila buzzatii. Mol Biol Evol 2003;20:674–85. https://doi.org/10.1093/molbev/msg070.

[16] Campos-Sánchez R, Cremona MA, Pini A, Chiaromonte F, Makova KD. Integration and Fixation Preferences of Human and Mouse Endogenous Retroviruses Uncovered with Functional Data Analysis. PLoS Comput Biol 2016;12:1–41. https://doi.org/10.1371/journal.pcbi.1004956.

[17] Oliver KR, Greene WK. Transposable elements: Powerful facilitators of evolution. BioEssays 2009;31:703–14. https://doi.org/10.1002/bies.200800219.

[18] Hughes JF, Coffin JM. Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome. Genetics 2005;171:1183–94. https://doi.org/10.1534/genetics.105.043976.

[19] Kent T V., Uzunović J, Wright SI. Coevolution between transposable elements and

recombination. Philos Trans R Soc B Biol Sci 2017;372. https://doi.org/10.1098/rstb.2016.0458.

[20] Coates BS, Hellmich RL, Grant DM, Abel CA. Mobilizing the genome of lepidoptera through novel sequence gains and end creation by non-autonomous Lep1 Helitrons. DNA Res 2012;19:11–21. https://doi.org/10.1093/dnares/dsr038.

[21] Biémont C, Vieira C. What transposable elements tell us about genome organization and evolution: The case of Drosophila. Cytogenet Genome Res 2005;110:25–34. https://doi.org/10.1159/000084935.

[22] Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet 2005;37:997–1002. https://doi.org/10.1038/ng1615.

[23] Philippsen GS, Avaca-Crusca JS, Araujo APU, DeMarco R. Distribution patterns and impact of transposable elements in genes of green algae. Gene 2016;594:151–9. https://doi.org/10.1016/j.gene.2016.09.012.

[24] Wu M, Li L, Sun Z. Transposable element fragments in protein-coding regions and their contributions to human functional proteins. Gene 2007;401:165–71. https://doi.org/10.1016/j.gene.2007.07.012.

[25] Daniel C, Behm M, Öhman M. The role of Alu elements in the cis-regulation of RNA processing. Cell Mol Life Sci 2015;72:4063–76. https://doi.org/10.1007/s00018-015-1990-3.

[26] Vorechovsky I. Transposable elements in disease-associated cryptic exons. Hum Genet 2010;127:135–54. https://doi.org/10.1007/s00439-009-0752-4.

[27] Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, et al. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 2006;441:87–90. https://doi.org/10.1038/nature04696.

[28] Platt RN, Blanco-Berdugo L, Ray DA. Accurate transposable element annotation is vital when analyzing new genome assemblies. Genome Biol Evol 2016;8:403–10. https://doi.org/10.1093/gbe/evw009.

[29] Bates HW. Contributions to an Insect Fauna of the Amazon Valley.—Lepidoptera:—Heliconinæ. J Proc Linn Soc London Zool 1862;6:73–7. https://doi.org/10.1111/J.1096-3642.1862.TB00932.X.

[30] Müller F. Ituna and Thyridia: a remarkable case of mimicry in butterflies. Trans Entomol Soc L 1879;1879:20–9.

[31] Ehrlich PR, Raven PH. Butterflies and Plants: A Study in Coevolution. Evolution (N Y) 1964;18:586. https://doi.org/10.2307/2406212.

[32] Kettlewell, B. Evolution of melanism: the study of a recurring necessity 1973.

[33] Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, Nadeau NJ, et al. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 2012;487:94–8. https://doi.org/10.1038/nature11041.

[34] Zhan S, Reppert SM. MonarchBase: the monarch butterfly genome database. Nucleic Acids Res 2013;41:D758–63. https://doi.org/10.1093/NAR/GKS1057.

[35] Zhan S, Zhang W, Niitepõld K, Hsu J, Haeger JF, Zalucki MP, et al. The genetics of monarch butterfly migration and warning colouration. Nature 2014;514:317–21. https://doi.org/10.1038/nature13812.

[36] Zhan S, Merlin C, Boore JL, Reppert SM. The monarch butterfly genome yields insights into long-distance migration. Cell 2011;147:1171–85. https://doi.org/10.1016/j.cell.2011.09.052.

[37] Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, Nadeau NJ, et al. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 2012;487:94–8. https://doi.org/10.1038/nature11041.

[38] Wu M, Sun Z, Luo G, Hu C, Zhang W, Han Z. Cloning and characterization of piggyBac-like elements in lepidopteran insects. Genetica 2011;139:149–54. https://doi.org/10.1007/s10709-010-9542-0.

[39] Sormacheva I, Smyshlyaev G, Mayorov V, Blinov A, Novikov A, Novikova O. Vertical evolution and horizontal transfer of CR1 non-LTR retrotransposons and Tc1/mariner DNA transposons in Lepidoptera species. Mol Biol Evol 2012;29:3685–702. https://doi.org/10.1093/molbev/mss181.

[40] Reiss D, Mialdea G, Miele V, de Vienne D, Peccoud J, Gilbert C, et al. Global survey of mobile DNA horizontal transfer in arthropods reveals Lepidoptera as a prime hotspot. PLOS Genet 2019;15:e1007965. https://doi.org/10.1371/journal.pgen.1007965.

[41] Talla V, Suh A, Kalsoom F, Dinca V, Vila R, Friberg M, et al. Rapid increase in genome size as a consequence of transposable element hyperactivity in wood-white (leptidea) butterflies. Genome Biol Evol 2017;9:2491–505. https://doi.org/10.1093/gbe/evx163.

[42] Hof AEV t., Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, et al. The industrial melanism mutation in British peppered moths is a transposable element. Nature 2016;534:102–5. https://doi.org/10.1038/nature17951.

[43] Gu L, Reilly PF, Lewis JJ, Reed RD, Andolfatto P, Walters JR. Dichotomy of dosage compensation along the neo Z chromosome of the Monarch butterfly. Curr Biol 2019;29:4071–7.

[44] Wang J, Wang A, Han Z, Zhang Z, Li F, Li X. Characterization of Three Novel SINE Families with Unusual Features in Helicoverpa armigera. PLoS One 2012;7:e31355. https://doi.org/10.1371/JOURNAL.PONE.0031355.

[45] Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking

transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol 2019;20:1–45. https://doi.org/10.1186/s13059-019-1905-y.

[46] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–9.

[47] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28:3150–2.

[48] Chénais B, Caruso A, Hiard S, Casse N. The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. Gene 2012;509:7–15. https://doi.org/10.1016/j.gene.2012.07.042.

[49] Osanai-Futahashi M, Suetsugu Y, Mita K, Fujiwara H. Genome-wide screening and characterization of transposable elements and their distribution analysis in the silkworm, Bombyx mori. Insect Biochem Mol Biol 2008;38:1046–57. https://doi.org/10.1016/j.ibmb.2008.05.012.

[50] Kapusta A, Suh A, Feschotte C. Dynamics of genome size evolution in birds and mammals. Proc Natl Acad Sci U S A 2017;114:E1460–9. https://doi.org/10.1073/pnas.1616702114.

[51] Talla V, Suh A, Kalsoom F, Dincă V, Vila R, Friberg M, et al. Rapid increase in genome size as a consequence of transposable element hyperactivity in wood-white (leptidea) butterflies. Genome Biol Evol 2017;9:2491–505. https://doi.org/10.1093/gbe/evx163.

[52] Kidwell MG. Transposable elements and the evolution of genome size in eukaryotes. Genetica 2002;115:49–63. https://doi.org/10.1023/A:1016072014259.

[53] Stritt C, Gordon SP, Wicker T, Vogel JP, Roulin AC. Recent activity in expanding populations and purifying selection have shaped transposable element landscapes

across natural accessions of themediterranean grass brachypodium distachyon. Genome Biol Evol 2018;10:304–18. https://doi.org/10.1093/gbe/evx276.

[54] Sultana T, Zamborlini A, Cristofari G, Lesage P. Integration site selection by retroviruses and transposable elements in eukaryotes. Nat Rev Genet 2017;18:292–308. https://doi.org/10.1038/nrg.2017.7.

[55] Sawicka K, Bushell M, Spriggs KA, Willis AE. Polypyrimidine-tract-binding protein: A multifunctional RNA-binding protein. Biochem Soc Trans 2008;36:641–7. https://doi.org/10.1042/BST0360641.

[56] Tree DRP, Shulman JM, Rousset R, Scott MP, Gubb D, Axelrod JD. Prickle mediates feedback amplification to generate asymmetric planar cell polarity signaling. Cell 2002;109:371–81. https://doi.org/10.1016/S0092-8674(02)00715-8.

[57] E L, Z Y, L F, S K, C V. A critical role of teashirt for patterning the ventral epidermis is masked by ectopic expression of tiptop, a paralog of teashirt in Drosophila. Dev Biol 2005;283:446–58. https://doi.org/10.1016/J.YDBIO.2005.05.005.

[58] RR D, JM L, JP K. Restriction of ectopic eye formation by Drosophila teashirt and tiptop to the developing antenna. Dev Dyn 2009;238:2202–10. https://doi.org/10.1002/DVDY.21927.

[59] Contreras EG, Palominos T, Glavic Á, Brand AH, Sierralta J, Oliva C. The transcription factor SoxD controls neuronal guidance in the Drosophila visual system. Sci Reports 2018 81 2018;8:1–12. https://doi.org/10.1038/s41598-018-31654-5.

[60] Matsubara D, Horiuchi SY, Shimono K, Usui T, Uemura T. The seven-pass transmembrane cadherin Flamingo controls dendritic self-avoidance via its binding to a LIM domain protein, Espinas, in Drosophila sensory neurons. Genes Dev 2011;25:1982–96. https://doi.org/10.1101/GAD.16531611.

[61] Bessa J, Gebelein B, Pichaud F, Casares F, Mann RS. Combinatorial control of

Drosophila eye development by Eyeless, Homothorax, and Teashirt. Genes Dev 2002;16:2415–27. https://doi.org/10.1101/GAD.1009002.

[62] Oliveira SG, Bao W, Martins C, Jurka J. Horizontal transfers of Mariner transposons between mammals and insects. Mob DNA 2012;3:1–6. https://doi.org/10.1186/1759-8753-3-14.

[63] Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL. Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. Genome Biol 2018;19:1–13. https://doi.org/10.1186/s13059-018-1456-7.

[64] Gilbert C, Schaack S, Pace JK, Brindley PJ, Feschotte C. A role for host-parasite interactions in the horizontal transfer of transposons across phyla. Nature 2010;464:1347–50. https://doi.org/10.1038/nature08939.

[65] Wallau GL, Ortiz MF, Loreto ELS. Horizontal transposon transfer in eukarya: Detection, bias, and perspectives. Genome Biol Evol 2012;4:689–99. https://doi.org/10.1093/gbe/evs055.

[66] Dupeyron M, Baril T, Hayward A. Broadscale evolutionary analysis of eukaryotic DDE transposons. BioRxiv 2021:2021.09.26.461848. https://doi.org/10.1101/2021.09.26.461848.

[67] Prysby MD, Oberhauser KS. Temporal and Geographic Variation in Monarch Densities: Citizen. Monarch Butterfly Biol Conserv 2004:9.

[68] Tang Z, Zhang HH, Huang K, Zhang XG, Han MJ, Zhang Z. Repeated horizontal transfers of four DNA transposons in invertebrates and bats. Mob DNA 2015;6:1–10. https://doi.org/10.1186/s13100-014-0033-1.

[69] Espeland M, Breinholt J, Willmott KR, Warren AD, Vila R, Toussaint EFA, et al. A Comprehensive and Dated Phylogenomic Analysis of Butterflies. Curr Biol 2018;28:770-778.e5. https://doi.org/10.1016/J.CUB.2018.01.061.

[70] Ustyugova S V, Amosova AL, Lebedev YB, Sverdlov ED. Cell line fingerprinting using retroelement insertion polymorphism. Biotechniques 2005;38:561–5.

[71] Lavoie CA, II RNP, Novick PA, Counterman BA, Ray DA. Transposable element evolution in Heliconius suggests genome diversity within Lepidoptera. Mob DNA 2013;21:1–10.

[72] Blass E, Bell M, Boissinot S. Accumulation and rapid decay of non-LTR retrotransposons in the genome of the three-spine stickleback. Genome Biol Evol 2012;4:687–702.

[73] Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL. LINEs between Species: Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life. Genome Biol Evol 2016;8:3301–22. https://doi.org/10.1093/GBE/EVW243.

[74] Woronik A, Tunström K, Perry MW, Neethiraj R, Stefanescu C, Celorio-Mancera M de la P, et al. A transposable element insertion is associated with an alternative life history strategy. Nat Commun 2019 101 2019;10:1–11. https://doi.org/10.1038/s41467-019-13596-2.

[75] Zhan S, Zhang W, Niitepõld K, Hsu J, Haeger JF, Zalucki MP, et al. The genetics of monarch butterfly migration and warning colouration. Nat 2014 5147522 2014;514:317–21. https://doi.org/10.1038/nature13812.

[76] Hill T, Koseva BS, Unckless RL. The Genome of Drosophila innubila Reveals Lineage-Specific Patterns of Selection in Immune Genes. Mol Biol Evol 2019;36:1405–17. https://doi.org/10.1093/MOLBEV/MSZ059.

[77] Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. Dynamic evolution of the innate immune system in Drosophila. Nat Genet 2007 3912 2007;39:1461–8. https://doi.org/10.1038/ng.2007.60.

[78] Shultz AJ, Sackton TB. Immune genes are hotspots of shared positive selection

across birds and mammals. Elife 2019;8. https://doi.org/10.7554/ELIFE.41815.

[79] Kimbrell DA, Beutler B. The evolution and genetics of innate immunity. Nat Rev Genet 2001 24 2001;2:256–67. https://doi.org/10.1038/35066006.

[80] Lenhard B, Sandelin A, Carninci P. Metazoan promoters: Emerging characteristics and insights into transcriptional regulation. Nat Rev Genet 2012;13:233–45. https://doi.org/10.1038/nrg3163.

[81] Jurka J, Kapitonov V V., Kohany O, Jurka M V. Repetitive Sequences in Complex Genomes: Structure and Evolution. Annu Rev Genomics Hum Genet 2007;8:241–59. https://doi.org/10.1146/annurev.genom.8.080706.092416.

[82] Kidwell MG, Lisch DR. Transposable elements and host genome evolution. Trends Ecol Evol 2000;15:95–9. https://doi.org/10.1016/S0169-5347(99)01817-0.

[83] Biémont C, Vieira C. Junk DNA as an evolutionary force. Nature 2006;443:521–4.

[84] Cosby RL, Chang N-C, Feschotte C. Host–transposon interactions: conflict, cooperation, and cooption. Genes Dev 2019;33:1098–116. https://doi.org/10.1101/GAD.327312.119.

[85] Casacuberta E, González J. The impact of transposable elements in environmental adaptation. Mol Ecol 2013;22:1503–17. https://doi.org/10.1111/mec.12170.

[86] Huff JT, Zilberman D, Roy SW. Mechanism for DNA transposons to generate introns on genomic scales. Nature 2016;538:533–6.

[87] Gilbert C, Feschotte C. Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. Curr Opin Genet Dev 2018;49:15–24. https://doi.org/10.1016/j.gde.2018.02.007.

[88] Galbraith JD, Ludington AJ, Sanders KL, Suh A, Adelson DL. Horizontal transfer and subsequent explosive expansion of a DNA transposon in sea kraits (Laticauda). Biol

Lett 2021;17:20210342. https://doi.org/10.1098/RSBL.2021.0342.

[89] Le Rouzic A, Capy P. The First Steps of Transposable Elements InvasionParasitic Strategy vs. Genetic Drift. Genetics 2005;169:1033–43. https://doi.org/10.1534/GENETICS.104.031211.

[90] Kofler R, Senti KA, Nolte V, Tobler R, Schlötterer C. Molecular dissection of a natural transposable element invasion. Genome Res 2018;28:824–35. https://doi.org/10.1101/gr.228627.117.

[91] Kelleher ES, Azevedo RBR, Zheng Y. The evolution of small-RNA-mediated silencing of an invading transposable element. Genome Biol Evol 2018;10:3038–57. https://doi.org/10.1093/gbe/evy218.

[92] Goodier JL. Restricting retrotransposons: A review. Mob DNA 2016;7. https://doi.org/10.1186/s13100-016-0070-z.

[93] Eickbush TH, Furano A V. Fruit flies and humans respond differently to retrotransposons. Curr Opin Genet Dev 2002. https://doi.org/10.1016/S0959-437X(02)00359-3.

[94] Ray DA, Grimshaw JR, Halsey MK, Korstian JM, Osmanski AB, Sullivan KAM, et al. Simultaneous TE Analysis of 19 Heliconiine Butterflies Yields Novel Insights into Rapid TE-Based Genome Diversification and Multiple SINE Births and Deaths. Genome Biol Evol 2019;11:2162–77. https://doi.org/10.1093/gbe/evz125.

[95] Levin HL, Moran J V. Dynamic interactions between transposable elements and their hosts. Nat Rev Genet 2011;12:615–27. https://doi.org/10.1038/nrg3030.

[96] Fontanillas P, Hartl DL, Reuter M. Genome organization and gene expression shape the transposable element distribution in the Drosophila melanogaster euchromatin. PLoS Genet 2007;3:2256–67. https://doi.org/10.1371/journal.pgen.0030210.

[97] Parfrey LW, Lahr DJG, Katz LA. The dynamic nature of eukaryotic genomes. Mol Biol

Evol 2008;25:787–94. https://doi.org/10.1093/molbev/msn032.

[98] Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2014;42:D7-17. https://doi.org/10.1093/nar/gky1069.

[99] Qu Z, Nong W, Yu Y, Baril T, Yip HY, Hayward A, et al. Genome of the four-finger threadfin Eleutheronema tetradactylum (Perciforms: Polynemidae). BMC Genomics 2020;21:1–10.

[100] Qu Z, Nong W, So WL, Barton-Owen T, Li Y, Leung TCN, et al. Millipede genomes reveal unique adaptations during myriapod evolution. PLoS Biol 2020;18:e3000636.

[101] Li Y, Nong W, Baril T, Yip HY, Swale T, Hayward A, et al. Reconstruction of ancient homeobox gene linkages inferred from a new high-quality assembly of the Hong Kong oyster (Magallana hongkongensis) genome. BMC Genomics 2020;21:1–17.

[102] Nong W, Law STS, Wong AYP, Baril T, Swale T, Chu LM, et al. Chromosomal-level reference genome of the incense tree Aquilaria sinensis. Mol Ecol Resour 2020;20:971.

[103] Nong W, Qu Z, Li Y, Barton-Owen T, Wong AYP, Yip HY, et al. Horseshoe crab genomes reveal the evolution of genes and microRNAs after three rounds of whole genome duplication. Commun Biol 2021;4:1–11.

[104] da Fonseca RR, Couto A, Machado AM, Brejova B, Albertin CB, Silva F, et al. A draft genome sequence of the elusive giant squid, Architeuthis dux. Gigascience 2020;9:1–12. https://doi.org/10.1093/gigascience/giz152.

[105] Smit AFA, Hubley RR, Green PR. RepeatMasker Open-4.0. Http://RepeatmaskerOrg 2013.

[106] Jurka J, Kapitonov V V., Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome

Res 2005;110:462–7. https://doi.org/10.1159/000084979.

[107] Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. Nucleic Acids Res 2016;44:D81–9. https://doi.org/10.1093/nar/gkv1272.

[108] Smit A, Hubley R. RepeatModeler Open-1.0. Http://RepeatmaskerOrg 2015.

[109] Bao Z, Eddy SR. Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. Genome Res 2002;12:1269–76. https://doi.org/10.1101/gr.88502.

[110] Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics 2005;21:i351–8.

[111] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture and applications. BMC Bioinformatics 2009;10:1–9. https://doi.org/10.1186/1471-2105-10-421.

[112] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 2013;30:772–80. https://doi.org/10.1093/molbev/mst010.

[113] Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 2009;25:1972–3. https://doi.org/10.1093/bioinformatics/btp348.

[114] Rice P, Longden L, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet 2000;16:276–7. https://doi.org/10.1016/S0168-9525(00)02024-2.

[115] Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 2007;35:W265–8. https://doi.org/10.1093/nar/gkm286.

[116] Ou S, Jiang N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. BioRxiv 2019:2–6.

[117] Wong WY, Simakov O. RepeatCraft: a meta-pipeline for repetitive element de-fragmentation and annotation. Bioinformatics 2018;35:1051–2. https://doi.org/10.1093/bioinformatics/bty745.

[118] Rubino F, Creevey CJ. MGkit: Metagenomic framework for the study of microbial communities 2014.

[119] Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 2014;30:3276–8.

[120] Team RC. R: A language and environment for statistical computing 2013.

[121] Team Rs. RStudio: integrated development for R. RStudio, Inc, Boston, MA URL Http//Www Rstudio Com 2015.

[122] Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–2. https://doi.org/10.1093/bioinformatics/btq033.

[123] Makowski D, Ben-Shachar M, Lüdecke D. bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. J Open Source Softw 2019;4:1541. https://doi.org/10.21105/joss.01541.

[124] Gel B, Serra E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics 2017;33:3088–90. https://doi.org/10.1093/bioinformatics/btx346.

[125] Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 1980;16:111–20.

[126] Pfeiler E, Nazario-Yepiz NO, Pérez-Gálvez F, Chávez-Mora CA, Laclette MRL, Rendón-Salinas E, et al. Population Genetics of Overwintering Monarch Butterflies,

Danaus plexippus (Linnaeus), from Central Mexico Inferred from Mitochondrial DNA and Microsatellite Markers. J Hered 2017;108:163–75. https://doi.org/10.1093/jhered/esw071.

[127] Wickham H. ggplot2: elegant graphics for data analysis. Springer; 2016.

[128] Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the Tidyverse. J Open Source Softw 2019;4:1686.

[129] Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res 2017;45:D200–3. https://doi.org/10.1093/NAR/GKW1129.

[130] Kapitonov V V., Tempel S, Jurka J. Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. Gene 2009;448:207. https://doi.org/10.1016/J.GENE.2009.07.019.

[131] Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol 2015;32:268–74. https://doi.org/10.1093/MOLBEV/MSU300.

[132] Kalyaanamoorthy S, Minh BQ, Wong TKF, Haeseler A von, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 2017 146 2017;14:587–9. https://doi.org/10.1038/nmeth.4285.

[133] Barreiro LB, Quintana-Murci L. From evolutionary genetics to human immunology: How selection shapes host defence genes. Nat Rev Genet 2010;11:17–30. https://doi.org/10.1038/NRG2698.

[134] Brunham RC, Plummer FA, Stephens3 RS. Bacterial Antigenic Variation, Host Immune Response, and Pathogen-Host Coevolution. INFECrION Immun 1993:2273–6.

[135] Vilcinskas A. Coevolution between pathogen-derived proteinases and proteinase

inhibitors of host insects. Virulence 2010;1:206–14. https://doi.org/10.4161/viru.1.3.12072.

[136] Yu P. Some host-parasite genetic interaction models. Theor Popul Biol 1972;3:347. https://doi.org/10.1016/0040-5809(72)90008-1.

[137] Anderson RM, May RM. Coevolution of hosts and parasites. Parasitology 1982;85:411–26. https://doi.org/10.1017/S0031182000055360.

Declarations

Ethics approval and consent to participate: Not applicable

Consent for Publication: Not applicable

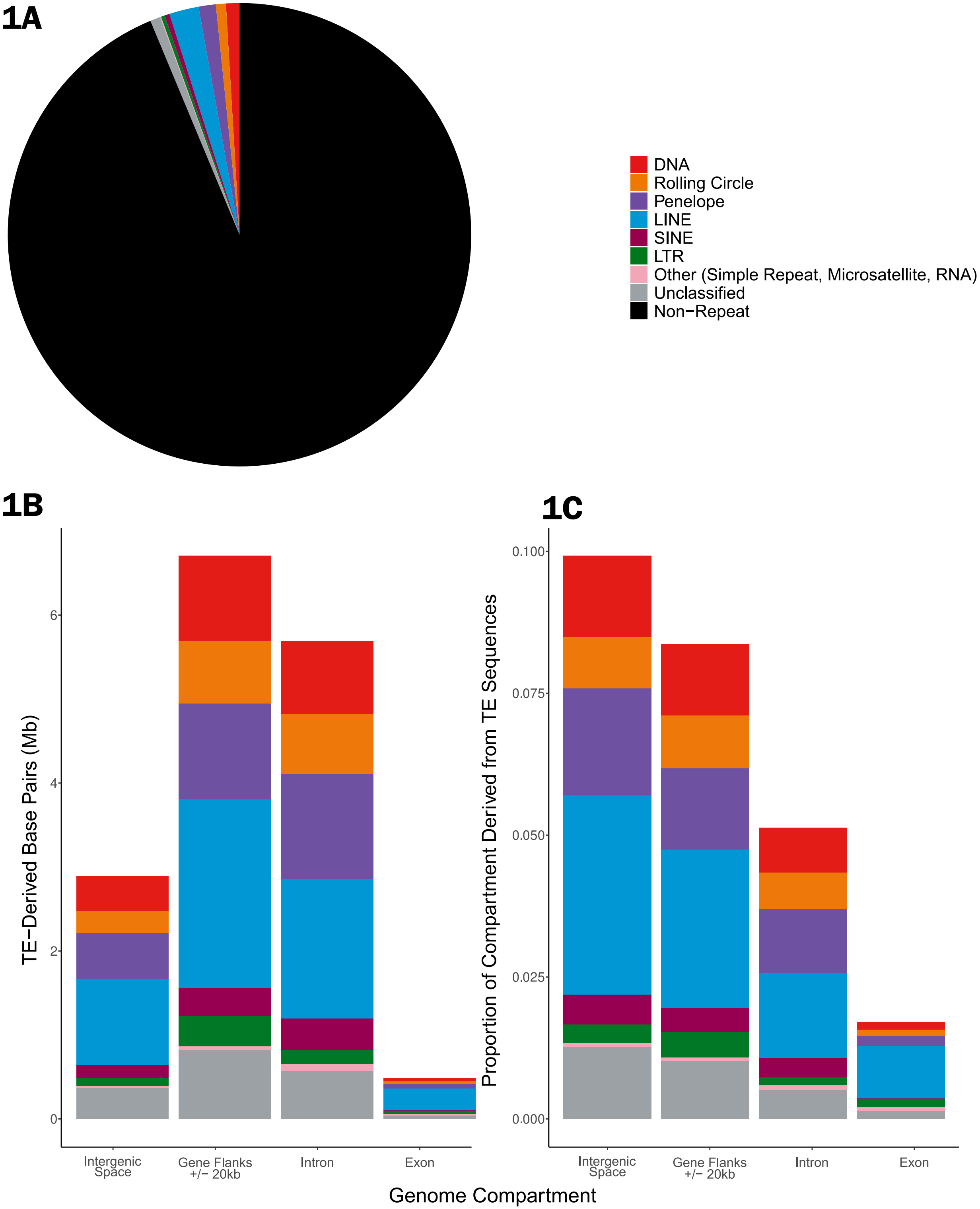
Availability of Data and Materials: All data generated or analysed during this study are included in this article and its supplementary information files.

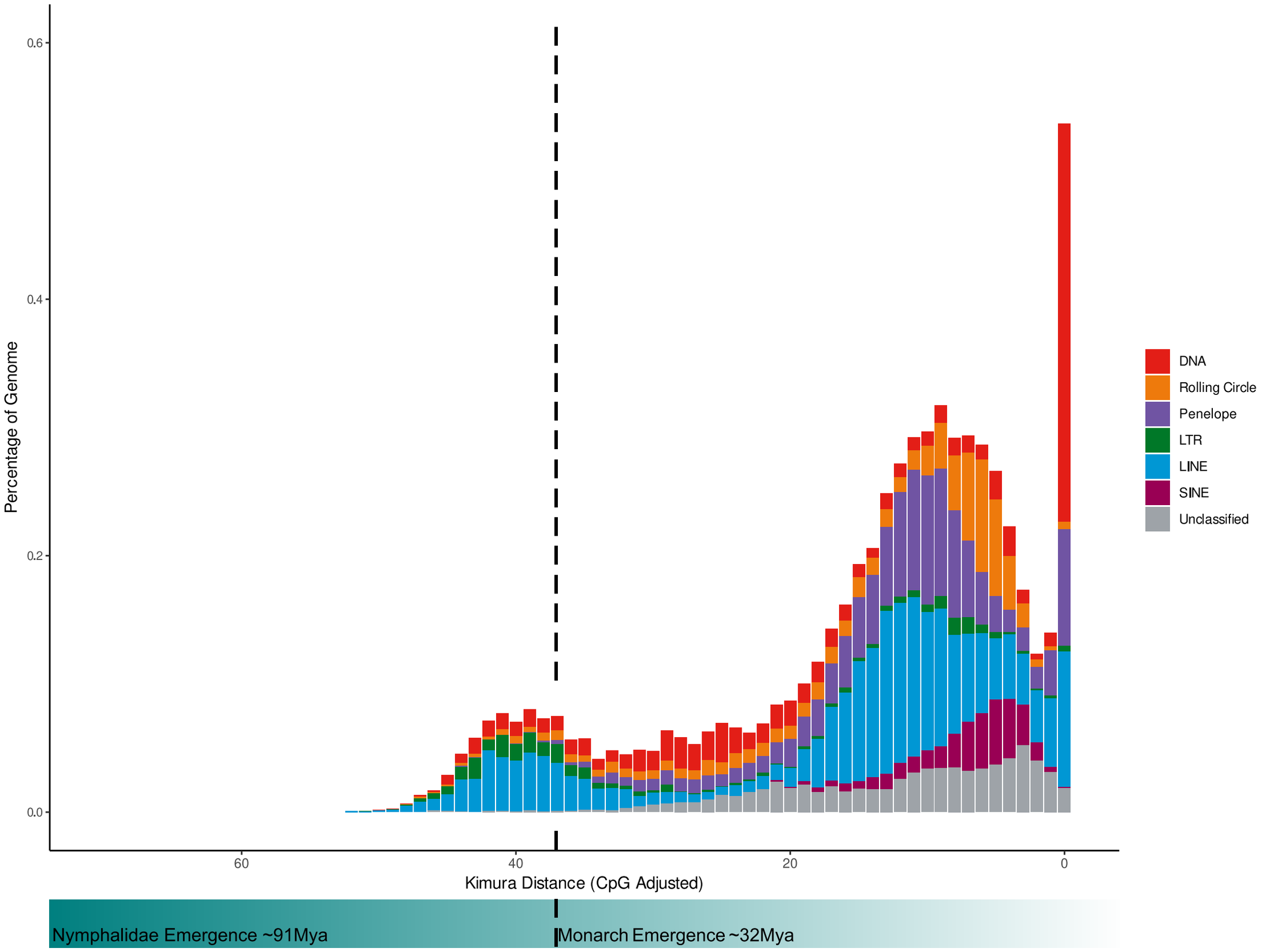
Competing Interests: The authors declare that they have no competing interests.

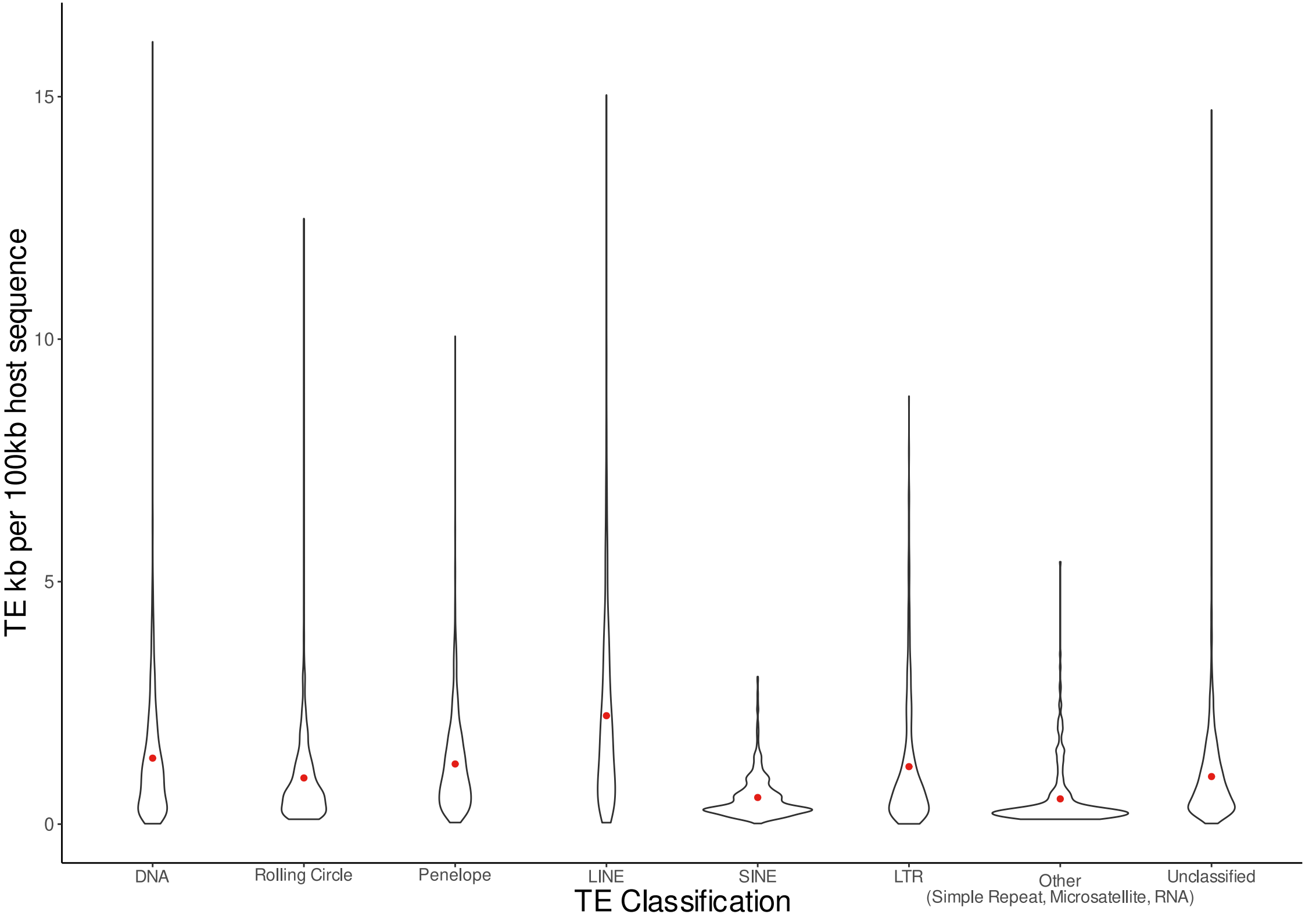
Funding: TB was supported by a studentship from the Biotechnology and Biological Sciences Research Council-funded South West Biosciences Doctoral Training Partnership (BB/M009122/1). AH was supported by a Biotechnology and Biological Sciences Research Council (BBSRC) David Phillips Fellowship (BB/N020146/1).

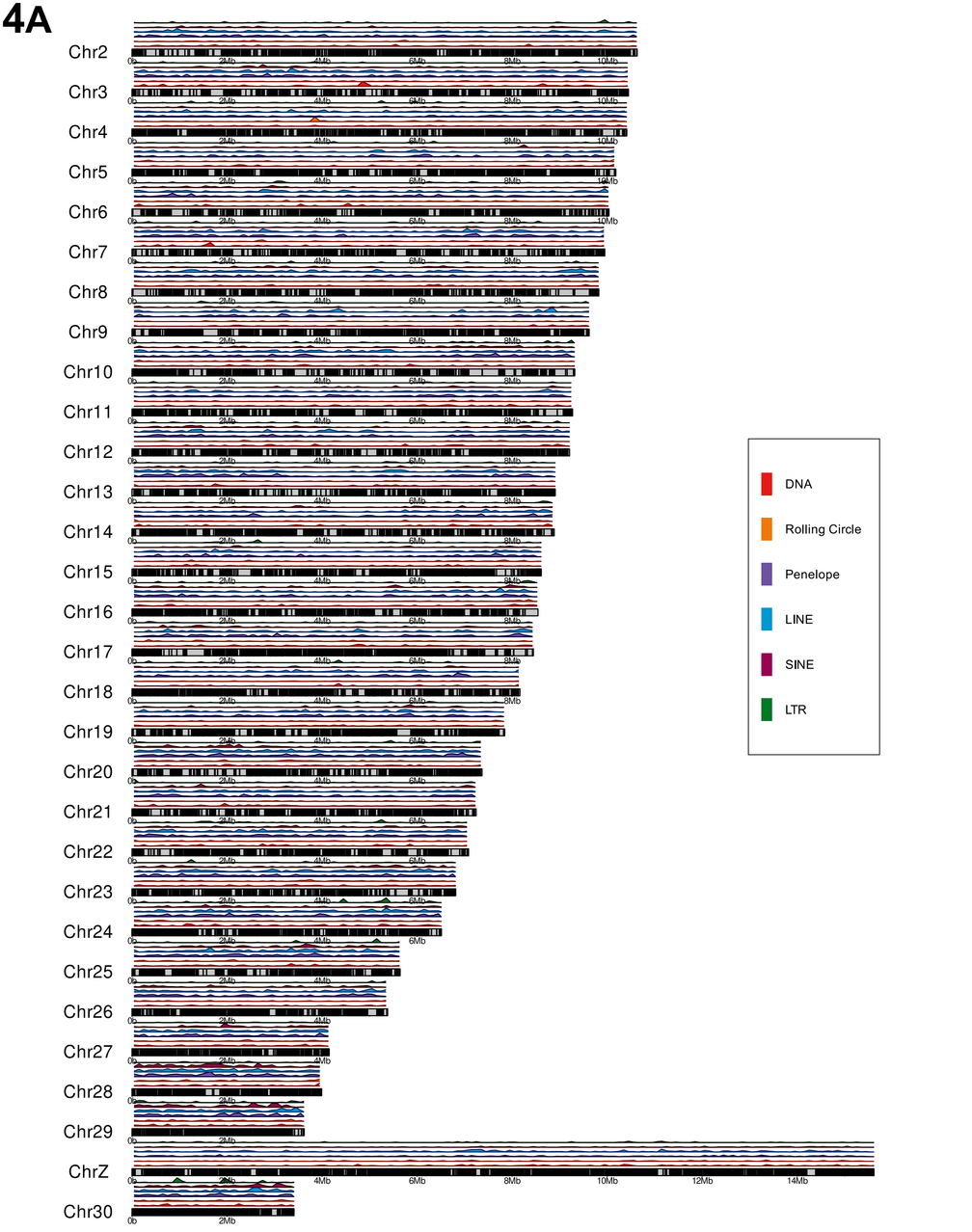
Authors' Contributions: TB developed the pipelines and scripts for automated TE annotation and downstream analyses, performed manual curation of TEs for the monarch, produced the figures, and drafted the manuscript. AH conceived and coordinated the study and participated in writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements: We would like to thank Dr James Galbraith, Ryan Biscocho, and Ryan Imrie for constructive comments on early drafts of this manuscript.

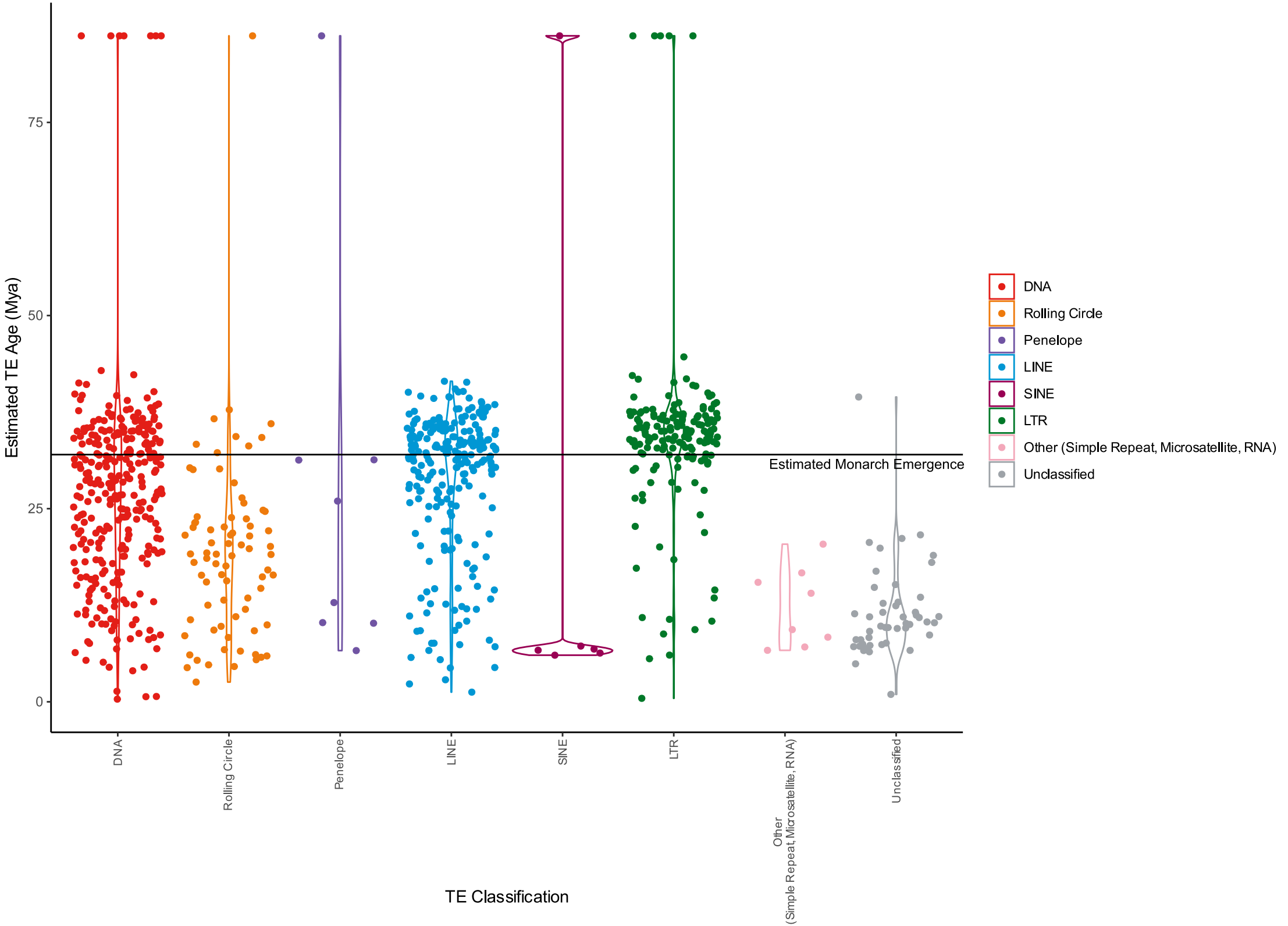

1A
DNA
Rolling Circle
Penelope
LINE
SINE
LTR
Other (Simple Repeat, Microsatellite, RNA)
Unclassified
Non−Repeat
1B
TE−Derived Base Pairs (Mb)
0
2
4
6
1C
Proportion of Compartment Derived from TE Sequences
0.000
0.025
0.050
0.075
0.100
Intergenic Space
Gene Flanks +/− 20kb
Intron
Exon
Genome Compartment

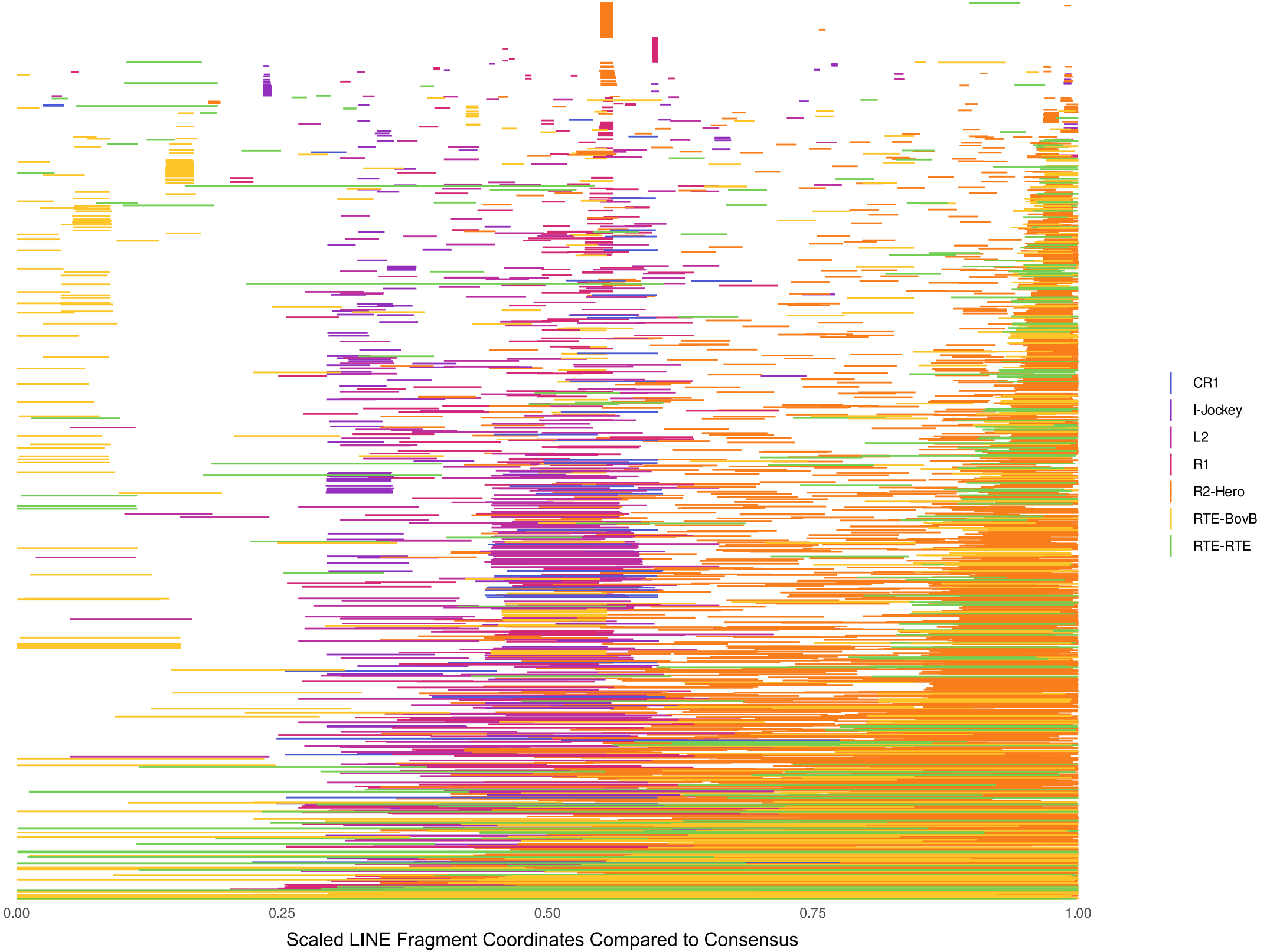

Percentage of Genome
0.6
0.4
0.2
0.0
60
40
20
0
Kimura Distance (CpG Adjusted)
DNA
Rolling Circle
Penelope
LTR
LINE
SINE
Unclassified
Nymphalidae Emergence ~91Mya
Monarch Emergence ~32Mya


TE kb per 100kb host sequence
15
10
5
0
DNA
Rolling Circle
Penelope
LINE
SINE
LTR
Other
(Simple Repeat, Microsatellite, RNA)
Unclassified
TE Classification


4A
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
Chr23
Chr24
Chr25
Chr26
Chr27
Chr28
Chr29
ChrZ
Chr30
0b
2Mb
4Mb
6Mb
8Mb
10Mb
12Mb
14Mb
DNA
Rolling Circle
Penelope
LINE
SINE
LTR

4B
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
Chr23
Chr24
Chr25
Chr26
Chr27
Chr28
Chr29
ChrZ
Chr30
0b
2Mb
4Mb
6Mb
8Mb
10Mb
12Mb
14Mb
Host Genes
Transposable Elements
TE Hotspot
Host Gene Hotspot
TE Coldspot
Host Gene Coldspot


Estimated TE Age (Mya)
75
50
25
0
Estimated Monarch Emergence
DNA
Rolling Circle
Penelope
LINE
SINE
LTR
Other (Simple Repeat, Microsatellite, RNA)
Unclassified
TE Classification


CR1
I-Jockey
L2
R1
R2-Hero
RTE-BovB
RTE-RTE
0.00
0.25
0.50
0.75
1.00
Scaled LINE Fragment Coordinates Compared to Consensus