Title

Effective population size does not explain long-term variation in genome size and transposable element content in animals

Authors

Alba Marino[1@], Gautier Debaecker[2], Anna-Sophie Fiston-Lavier[1,3], Annabelle Haudry[4], Benoit Nabholz[1,3]

Affiliations

[1] ISEM, Univ Montpellier, CNRS, IRD, Montpellier, France

[2] Université Claude Bernard Lyon 1, LEHNA, UMR 5023, CNRS, Villeurbanne, France

[3] Institut Universitaire de France (IUF), Paris, France

[4] Université Claude Bernard Lyon 1, LBBE, UMR 5558, CNRS, Villeurbanne, France

[@] Corresponding author: alba.marino@umontpellier.fr

Running title

Long-term evolution of genome size in animals

Abstract

Animal genomes exhibit a remarkable variation in size, but the evolutionary forces responsible for such variation are still debated. As the effective population size (N_e) reflects the intensity of genetic drift, it is expected to be a key determinant of the fixation rate of nearly-neutral mutations. Accordingly, the Mutational Hazard Hypothesis postulates lineages with low N_e to have bigger genome sizes due to the accumulation of slightly deleterious transposable elements (TEs), and those with high N_e to maintain streamlined genomes as a consequence of a more effective selection against TEs. However, the existence of both empirical confirmation and refutation using different methods and different scales precludes its general validation. Using high-quality public data, we estimated genome size, TE content and rate of non-synonymous to synonymous substitutions (dN/dS) as N_e proxy for 807 species including vertebrates, molluscs and insects. After collecting available life-history traits, we tested the associations among population size proxies, TE content and genome size, while accounting for phylogenetic non-independence. Our results confirm TEs as major drivers of genome size variation, and endorse life-history traits and dN/dS as reliable proxies for N_e. However, we do not find any evidence for increased drift to result in an accumulation of TEs across animals. Within more closely related clades, only a few isolated and weak associations emerge in fishes and birds. Our results outline a scenario where TE dynamics vary according to lineage-specific patterns, lending no support for genetic drift as the predominant force driving long-term genome size evolution in animals.

1. Introduction

The variation in genome size among animals is remarkable, spanning four orders of magnitude, from 0.02 Gb in the nematode *Pratylenchus coffeae* to 120 Gb in the marbled

lungfish *Protopterus aethiopicus* (Gregory 2023). Understanding why such a huge variation occurs is a long-standing question in evolutionary biology. It is now established that genome size is not related to organismal complexity (C-value enigma) or the number of coding genes in eukaryotes. Rather, variation in DNA content depends on the amount of non-coding DNA such as Transposable Elements (TEs), introns and pseudogenes (Elliott and Gregory 2015; Kidwell 2002; Lynch et al. 2011). However, the evolutionary mechanisms leading certain lineages to inflate their genome size or to maintain streamlined genomes are still debated (Galtier 2024).

The various hypotheses that have been proposed can be divided into adaptive and non-adaptive. Adaptive theories such as the nucleoskeletal (Cavalier-Smith 1978) and the nucleotypic one (Gregory and Hebert 1999) consider genome size to be mainly composed of “indifferent” DNA (Graur et al. 2015), whose bulk is indirectly selected as a consequence of its effect on nuclear and cellular volume. Cellular phenotypes, in turn, are thought to influence the fitness of organisms by affecting traits such as cell division rate (Bennett and Riley 1997), metabolic rate (Vinogradov 1995; Vinogradov 1997) and developmental time and complexity (Jockusch 1997; Olmo et al. 1989; Gregory 2002). On the other hand, non-adaptive theories emphasize the importance of the neutral processes of mutation and genetic drift in determining genome size (Lynch and Conery 2003; Petrov 2001). In particular, the concepts originally proposed by Lynch and Conery were later on formalized within the framework of the Mutational Hazard Hypothesis (MHH) (Lynch 2007).

The MHH posits that tolerance to the accumulation of non-coding DNA depends on its mutational liability, which is minimized when the mutation rate is low and the effective population size is high (N_e, which is inversely proportional to the intensity of genetic drift). The fundamental assumption is that most of such extra DNA is mildly deleterious and its fate in the population depends on the interplay between selection and genetic drift: a newly emerged nearly-neutral allele with a given negative selective effect (s) should be effectively removed from the population when N_e is high ($|N_e s| >> 1$), while it should have approximately

the same chance of fixation of a neutral allele when N_e is low ($|N_e s| \sim 1$ or lower than 1) as drift will be the predominant force (Ohta 1992).

Because of the pervasiveness of TEs and their generally neutral or slightly deleterious effect (Arkhipova 2018), their dynamics in response to changing N_e are of particular interest in the context of the MHH. TE insertions are expected to drift to fixation as neutral alleles and enlarge genome size in organisms with low N_e, while the genomes of organisms with large N_e are expected to remain streamlined as emerging TEs should be efficiently removed by purifying selection (Lynch 2007). The MHH is highly popular: indeed, it is based on general principles of population genetics and proposes a unifying explanation for the evolution of complex traits of genome architecture without recourse to specific molecular processes. The studies supporting the MHH are based on phylogenetically very diverse datasets as the goal of the theory is to explain broad patterns of complexity emergence and variation (Lynch and Conery 2003; Yi and Streelman 2005; Yi 2006). Nevertheless, other authors pointed out that the application of the MHH to such distantly related taxa could suffer from confounding factors intervening across organisms with very different biologies (Charlesworth and Barton 2004; Daubin and Moran 2004), thus raising the question whether N_e could explain genome size variation patterns at finer scales. Additionally, potential issues of robustness of the original dataset to phylogenetic control have been raised (Lynch 2011; Whitney et al. 2011; Whitney and Garland 2010). On top of that, an alternative TE-host-oriented perspective is that the accumulation of TEs in particular depends on their type of activity and dynamics, as well as on the lineage-specific silencing mechanisms evolved by host genomes (Ågren and Wright 2011).

Recent attempts have been made to assess the impact of increased genetic drift on the genomic TE content and genome size increase across closely related species with similar biological characteristics, employing both genetic diversity data and life-history traits (LHTs) as predictors of N_e. While some studies do not find any evidence supporting the role of N_e in genome size and TE content variation (Bast et al. 2016; Kapheim et al. 2015; Mackintosh et

al. 2019; Roddy et al. 2021; Yang et al. 2024), others do (Chak et al. 2021; Cui et al. 2019; Lefébure et al. 2017; Mérel et al. 2021; Merel et al. 2024). Thus, a univocal conclusion can hardly be drawn from such studies. The MHH has thus been investigated at either very wide - from prokaryotes to multicellular eukaryotes - or narrow phylogenetic scales (i.e., inter-genera or inter-population). However, no study spanning across an exhaustive set of distantly related taxa and relying on a phylogenetic framework has yet been performed to our knowledge. Such an approach would allow a systematic test of the association between N_e variation and long-term patterns of genome and TE expansion at a wider evolutionary scale, while controlling for the effect of phylogenetic inertia. Synonymous genetic diversity is commonly used to inform patterns of N_e (Lynch and Conery 2003; Romiguier et al. 2014; Buffalo 2021; Lynch et al. 2023). However, while its insights are limited to the age of current alleles (mostly less than $10N_e$ generations in diploid organisms), complex genomic features likely have much deeper origins. Comparative measures of divergence, like the genome-wide ratio of non-synonymous substitution rate to synonymous substitution rate (dN/dS), can quantify the level of genetic drift acting on protein-coding sequences since the last common ancestor. This approach accounts for processes occurring over a longer time scale than those responsible for genetic diversity. In fact, polymorphism might reflect relatively recent population size fluctuations (Daubin and Moran 2004) and might even diverge from indices of long-term N_e if the selection-drift equilibrium is not reestablished (Lefébure et al. 2017; Müller et al. 2022). We therefore adopt dN/dS as an index of long-term N_e, as it is more likely to reflect the evolutionary lapse during which the deep changes in genome size and TE content that we are investigating occurred (Whitney et al. 2011; Whitney and Garland 2010).

In this study, we took advantage of 3,214 public metazoan reference genomes and C-value records (i.e., the haploid DNA content of a nucleus) to estimate genome sizes. A subset of 807 species including birds, mammals, ray-finned fishes, insects and molluscs was selected to test the predictions of the MHH, especially through the relationship among the level of drift, genome size and genomic TE content. A phylogeny was computed with metazoan-conserved

genes. N_e was accounted for by the dN/dS and by life-history traits (LHTs) when available; TE contents were estimated *de novo* from read data. Controlling for phylogenetic non-independence, we (1) assessed the contribution of TE quantity to genome size differences and (2) evaluated the efficacy of N_e proxies in explaining genome size and TE content variation, across the whole dataset and within specific clades.

2. Results

2.1. Selection of high-quality assemblies

The reference genomes of the 3,214 metazoan species were downloaded via the NCBI genome database (Supplemental Table S1). From these, the genomes with contig N50 ≥ 50 kb and available C-value record were employed for genome size estimation (see 2.2, 4.2). Based on the assembly quality (contig N50 ≥ 50 kb), the completeness of metazoan core genes (complete BUSCO orthologs ≥ 70%) (Manni et al. 2021) and the raw sequencing data availability, a dataset of 930 genomes was then retained for downstream analyses. For reliable estimation of substitution rates, this dataset was further downsized to 807 representative genomes as species-poor, deep-branching taxa were excluded (Figure 1; Supplemental Table S2).

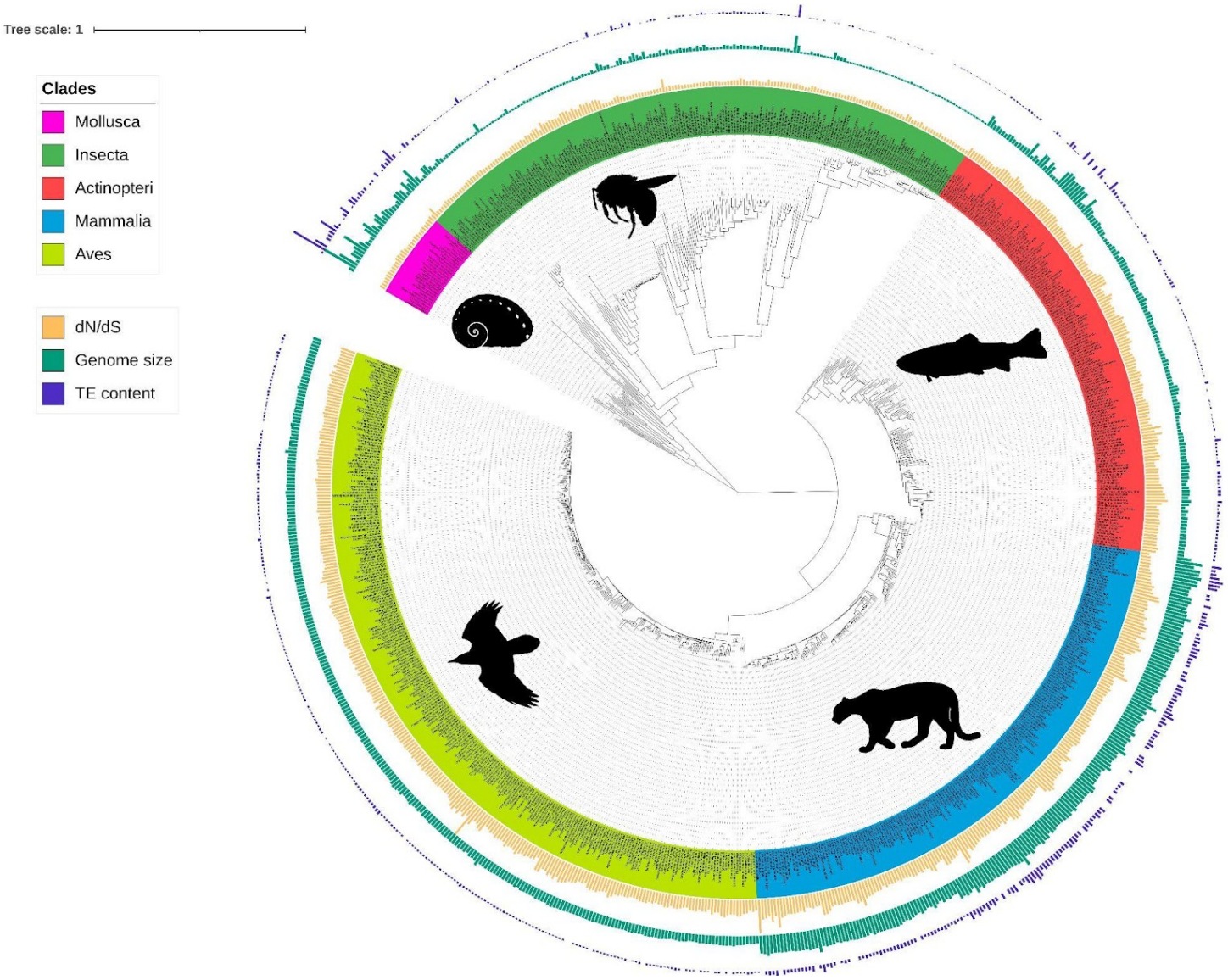


Figure 1. Phylogeny of the 807 species including ray-finned fishes (Actinopteri), birds (Aves), insects (Insecta), mammals (Mammalia), and molluscs (Mollusca) with bars corresponding to TE content (bp, blue), genome size (bp, green), and dN/dS estimations (values between 0 and 1, yellow). Branch lengths are amino acid substitutions calculated on BUSCO genes. The tree was plotted with iTOL (Letunic and Bork 2021).

2.2. Genome size estimation

For the selected genomes, the genome size records available for 465 species (Supplemental Table S3) show that the assembly size is strongly positively correlated with the C-value (Pearson's r = 0.97, p-value < 0.001). This indicates the reliability of the use of assemblies to estimate genome size. Although a non-linear model is not statistically better than the Weighted Least Squares (WLS), assembly sizes tend to underestimate genome size in comparison to C-values, an effect becoming more and more evident as genome size increases (Supplemental Figure S1). According to the metadata that we could retrieve, whether a genome was assembled using long read or uniquely short read data does not affect the slope of the WLS with assembly size as independent variable (T-test: long reads - p-value = 0.88; short reads - p-value = 0.87). Because it is not affected by sequencing biases, C-value was used when available; otherwise, the predicted C-value according to WLS was employed. The value chosen as genome size estimation is reported for each species in Supplemental Tables S1, S2.

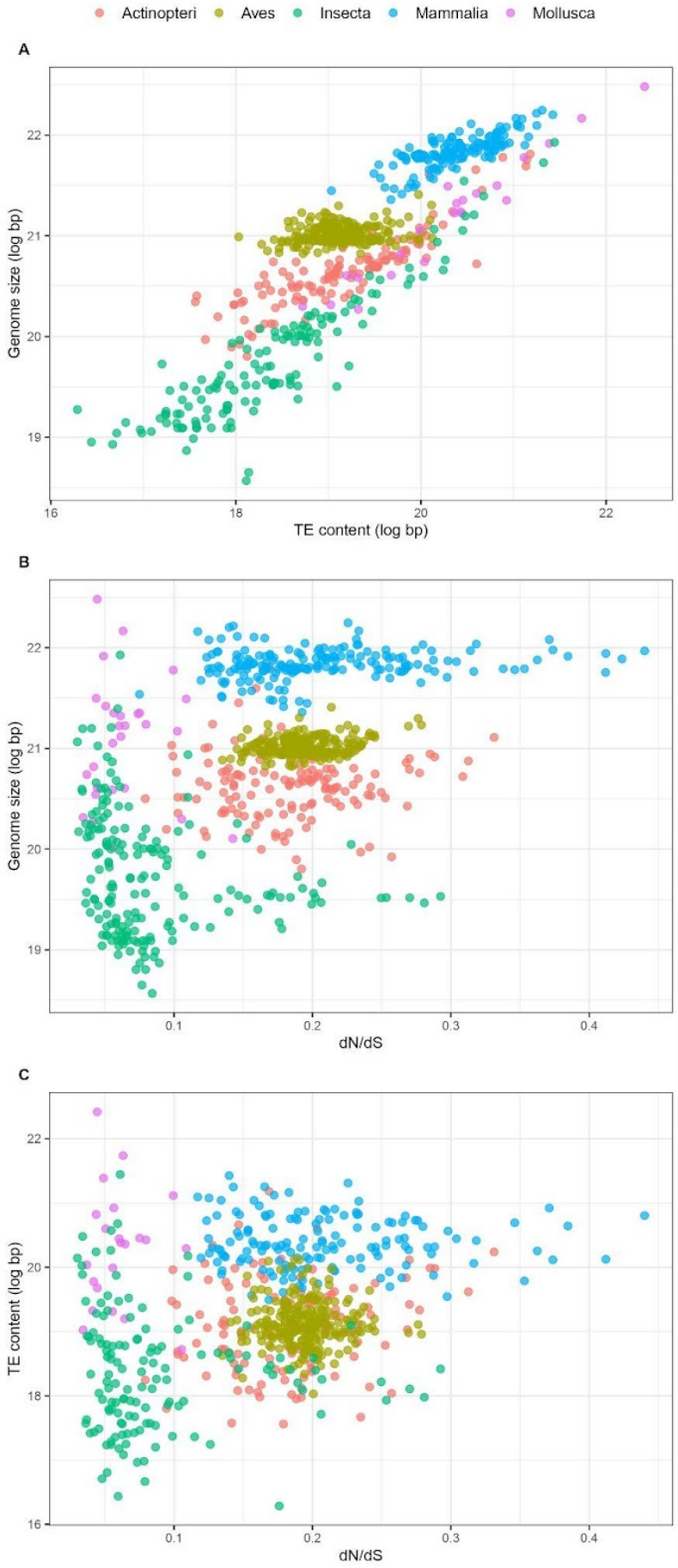
2.3. Transposable elements and genome size variation

Repeat content of a subset of 29 dipteran genomes was previously estimated using EarlGrey v1.3 (Baril et al. 2024) and a wrapper around dnaPipeTE (Goubert et al. 2015), an assembly-based and a read-based pipeline, respectively. dnaPipeTE leverages the sampling of reads at low-coverage to perform *de novo* assembly of TE consensus sequences: this approach has the advantage of being unbiased by repeat sequences potentially missing from genome assemblies. The results of the two methods overall agree with each other across the scanned genomes (Genomic percentage of TEs: Pearson's r = 0.88, p-value < 0.001; TE base pairs: Pearson's r = 0.90, p-value < 0.001), with the most notable difference being the proportion of unknown elements, generally higher in dnaPipeTE estimations (Wilcoxon signed-rank test, p-value < 0.001; Supplemental Table S4; Supplemental Figure S2, S3). We

therefore mined the remaining genomes with dnaPipeTE which is much less computationally intensive. Repeat content could only be estimated for 672 species over the 807 representative genomes (Supplemental Table S2): for the remaining 135, the pipeline could not be run because of unsuitable reads (e.g. only long reads available or too low coverage). A very strong positive correlation between TE content and genome size is found both across the whole dataset and within taxa (Figure 2A; Supplemental Figure S4; Table 1; Supplemental Table S5). A notable exception concerns the avian clade that deviates from this pattern: the range of TE content is wider than the one of genome size compared to the other clades (Figure 2A), resulting in a weaker power of TEs in explaining genome size variation in this group (Table 1; Supplemental Table S5).

PIC Correlations		Regression coefficient	Adjusted-R2	p-value
GS (log bp) ~ TE content (log bp)	Overall Dataset	**0.219** ***	**0.417** ***	< 0.001
	Actinopteri	**0.300** ***	**0.610** ***	< 0.001
	Aves	**0.042** ***	**0.039** **	0.001
	Insecta	**0.356** ***	**0.626** ***	< 0.001
	Mammalia	**0.200** ***	**0.526** ***	< 0.001
	Mollusca	**0.605** ***	**0.895** ***	< 0.001

Table 1. Correlation between genome size and overall TE content based on Phylogenetic Independent Contrasts. Statistics are shown relative to the overall dataset and to each clade. Variables were log-transformed previous to regression. * $0.05 < p \leq 0.01$; ** $0.01 < p \leq 0.001$; *** $p < 0.001$.


Actinopteri
Aves
Insecta
Mammalia
Mollusca
A
Genome size (log bp)
TE content (log bp)
B
Genome size (log bp)
dN/dS
C
TE content (log bp)
dN/dS

182

Figure 2. (A) Relationship between overall TE content and genome size (log-transformed). Slope = 0.718, adjusted-R^2 = 0.751, p-value < 0.001. (B) Relationship between genome size and dN/dS. Slope = 6.100, adjusted-R^2 = 0.275, p-value < 0.001. (C) Relationship between TE content and dN/dS. Slope = 4.253, adjusted-R^2 = 0.092, p-value < 0.001. Statistics refer to linear regression, see Tables 1 and 2 for Phylogenetic Independent Contrasts results.

Alternatively to the impact of TEs on genome size, we investigated whether whole or partial genome duplications could be major factors in genome size variation among animals. BUSCO Duplicated score has indeed a slightly positive correlation with genome size, which is however much weaker than that of TEs (Slope = $6.639 \cdot 10^{-9}$, adjusted-R^2 = 0.022, p-value < 0.001). Of the 24 species with more than 30% of duplicated BUSCO genes, 13 include sturgeon, salmonids and cyprinids, known to have undergone whole genome duplication (Du et al. 2020; Li and Guo 2020; Lien et al. 2016), and five are dipteran species, where gene duplications are common (Ruzzante et al. 2019). In general, TEs appear as the main factor influencing genome size variation across species.

2.4. dN/dS and life history traits as proxies of effective population size

Intensity of effective selection acting on species can be informed by the dN/dS ratio: a dN/dS closer to 1 accounts for more frequent accumulation of mildly deleterious mutations over time due to increased genetic drift, while a dN/dS close to zero is associated with a stronger effect of purifying selection. We therefore employed this parameter as a genomic indicator of N_e, as the two are expected to scale negatively between each other. We compiled several LHTs from different sources (see 4.6) to cross-check our estimations of dN/dS. In general, dN/dS is expected to scale positively with body length, age at first birth, maximum longevity, age at sexual maturity and mass, and to scale negatively with metabolic rate, population density and depth range.

We estimated dN/dS with a mapping method (hereafter referred to as Bio++ dN/dS; Dutheil et al. 2006; Guéguen et al. 2013), and with a bayesian approach using Coevol (hereafter referred to as Coevol dN/dS; Lartillot and Poujol 2011). The two metrics are reported in Supplemental Table S2. Note that for Coevol, we report both the results relative to dN/dS at terminal branches (Table 2) and the correlations inferred by the model (Table 3; Supplemental Table S5).

As expected, Bio++ dN/dS scales positively with body mass and longevity under Phylogenetic Independent Contrasts (PIC) (Table 2; Supplemental Figure S5). dN/dS estimation on the trimmed phylogeny deprived of short and long branches results in a stronger correlation with LHTs, suggesting that short branches might contribute to dN/dS fluctuations (Table 2; Supplemental Figure S6). Coevol dN/dS values are however highly concordant with Bio++ dN/dS (Supplemental Figure S7) and scale positively with body mass and longevity, as well (Table 2).

PIC correlations		Bio++ (full phylogeny)			Bio++ (trimmed phylogeny)			Coevol		
		Regression coefficient	Adjusted-R2	p-value	Regression coefficient	Adjusted-R2	p-value	Regression coefficient	Adjusted-R2	p-value
Body mass (log gr) ~ dN/dS	Overall Dataset	**6.865 *****	**0.036 *****	< 0.001	**23.905 *****	**0.044 *****	< 0.001	**11.422 *****	**0.087 *****	< 0.001
Longevity (log years) ~ dN/dS	Overall Dataset	**2.147 ****	**0.025 ****	0.004	**11.349 *****	**0.097 *****	< 0.001	**2.970 *****	**0.050 *****	< 0.001
GS (log bp) ~ dN/dS	Overall Dataset	0.199	0.001	0.175	0.114	-0.002	0.858	**-0.287 ***	**0.004 ***	0.039
	Actinopteri	0.915	0.016	0.066	0.687	-0.009	0.701	0.282	-0.005	0.583
	Aves	0.109	0.001	0.270	0.709	0.013	0.073	0.238	-0.001	0.407
	Insecta	1.085	-0.002	0.411	-3.703	0.005	0.204	**-1.241 ***	**0.026 ***	0.018
	Mammalia	0.071	-0.005	0.701	0.220	-0.011	0.777	-0.165	0.003	0.227
	Mollusca	3.504	-0.032	0.698	3.504	-0.032	0.698	-1.488	-0.032	0.699
TE content (log bp) ~ dN/dS	Overall Dataset	0.798	0.004	0.062	-0.216	-0.002	0.899	**-0.903 ***	**0.004 ***	0.050
	Actinopteri	2.139	0.015	0.098	1.393	-0.012	0.751	**3.407 ***	**0.046 ***	0.013
	Aves	0.340	-0.002	0.513	-1.492	-0.004	0.551	2.285	0.006	0.129
	Insecta	1.744	-0.004	0.528	-3.096	-0.007	0.602	0.063	-0.007	0.960
	Mammalia	1.001	0.0027	0.238	1.023	-0.012	0.691	**-2.113 ****	**0.063 ****	0.001
	Mollusca	22.930	0.026	0.225	22.930	0.026	0.225	0.881	-0.050	0.936
Recent TE content (log bp) ~ dN/dS	Overall Dataset	**1.963 ****	**0.012 ****	0.003	0.691	-0.002	0.727	-1.225	0.0029	0.089
	Actinopteri	2.271	0.013	0.113	1.241	-0.012	0.793	**4.365 ****	**0.061 ****	0.005
	Aves	0.545	-0.001	0.384	-0.385	-0.006	0.890	**4.982 ****	**0.028 ****	0.006
	Insecta	1.725	-0.004	0.530	-4.792	-0.004	0.428	0.536	-0.006	0.668
	Mammalia	**4.115 ***	**0.024 ***	0.031	3.192	-0.006	0.460	**-3.151 ***	**0.024 ***	0.032
	Mollusca	14.730	-0.024	0.481	14.730	-0.024	0.481	2.986	-0.047	0.8026

Table 2. PIC results for the correlations of LHTs, genome size, and TE content against dN/dS. Results for Bio++ dN/dS are shown for the full dataset and for the phylogeny deprived of the longest (> 1) and shortest (< 0.01) terminal branches. Results for Coevol dN/dS are relative to the GC3-poor geneset. Only body mass and longevity are reported as LHTs (for an overview of all traits, see Table 3). For genomic traits, statistics are reported relative to the overall dataset and to each clade. Expected significant correlations of dN/dS with LHTs and genomic traits are marked in bold black; significant correlations opposite to the expected trend are marked in bold red. * $0.05 < p \leq 0.01$; ** $0.01 < p \leq 0.001$; *** $p < 0.001$

As for Coevol reconstruction, dN/dS covaries as expected with most of the tested LHTs: dN/dS scales positively with body length, longevity, mass, sexual maturity and depth range in fishes (Table 3); the same is found in mammals, in addition to negative correlations with population density and metabolic rate; in birds, mass, metabolic rate and sexual maturity correlate in the same way with dN/dS, although this is consistently observed only for the GC3-rich gene set (Table 3). Based on the available traits, the estimations of dN/dS ratios obtained using two different methods correspond in general to each other, supporting dN/dS as a meaningful indicator of long-term effective population size, at least for vertebrate clades. Results are not reported for molluscs and insects, as none and very few records of LHTs (seven species with at least one trait) were available, respectively.

Coevol correlations	Actinopteri dN/dS				Aves dN/dS				Insecta dN/dS				Mammalia dN/dS				Mollusca dN/dS			
	GC3-poor		GC3-rich		GC3-poor		GC3-rich		GC3-poor		GC3-rich		GC3-poor		GC3-rich		GC3-poor		GC3-rich	
	CC	PP	CC	PP	CC	PP	CC	PP	CC	PP	CC	PP	CC	PP	CC	PP	CC	PP	CC	PP
Body length (cm)	**0.246**	0.92	**0.289**	0.97									**0.371**	1	**0.271**	0.99				
Basal metabolic rate (ml/O_2/hour)[a]													**-0.258**	0.02	**-0.337**	0.01				
Age at first birth (days)													**0.428**	1	**0.384**	1				
Population density (individuals/km²)													**-0.361**	0.001	**-0.188**	0.09				
Maximum longevity (years)	**0.435**	0.97	**0.330**	0.95	**0.303**	0.92	0.109	0.75	0.0984	0.62	0.157	0.67	**0.295**	0.99	**0.376**	1				
Mass (g)	**0.223**	0.89	**0.308**	0.97	0.096	0.84	**0.155**	0.95	0.0243	0.52	0.221	0.76	**0.412**	1	**0.316**	1				
Metabolic rate (Watt)[b]					0.030	0.57	**-0.244**	0.07	0.201	0.78	-0.038	0.45	**-0.283**	0.02	**-0.377**	0.003				
Sexual maturity (days)	**0.320**	0.91	0.286	0.88	0.032	0.58	**0.284**	0.99	-0.244	0.19	0.072	0.58	**0.468**	1	**0.329**	1				
Depth range	**0.388**	0.93	**0.584**	1																
Genome size (bp)	0.034	0.59	-0.0273	0.42	0.054	0.69	0.072	0.77	**-0.330**	0.03	-0.110	0.24	-0.041	0.35	-0.140	0.12	0.142	0.65	0.306	0.82
Overall TE content (bp)	0.192	0.88	0.167	0.84	0.065	0.73	**-0.195**	0.03	-0.203	0.16	-0.017	0.45	**-0.220**	0.08	**-0.249**	0.06	-0.119	0.39	-0.021	0.47
Recent TE content (bp)	0.205	0.89	**0.258**	0.95	0.109	0.83	**-0.240**	0.01	-0.136	0.25	0.049	0.63	-0.119	0.23	**-0.296**	0.06	-0.172	0.34	-0.004	0.50

Table 3. Correlation coefficients (CC) and posterior probabilities (PP) estimated by Coevol with the GC3-poor and GC3-rich genesets for the coevolution of dN/dS with life history and genomic traits. Different LHTs are shown according to availability for a clade. Posterior probabilities lower than 0.1 indicate significant negative correlations; posterior probabilities higher than 0.9 indicate significant positive correlations. Expected significant correlations of dN/dS with LHTs and genomic traits are marked in bold black; significant correlations opposite to the expected trend are marked in bold red. [a]Pantheria, [b]AnAge.

2.5. dN/dS does not predict genome size and overall TE content across metazoans

If increased genetic drift leads to TE expansions, a positive relationship between dN/dS and TE content, and more broadly with genome size, should be observed. However, we find no statistical support for this relationship across all species in the PIC analysis. Similarly, no association is found when short and long branches are removed (Figure 2B-C; Supplemental Figure S8; Table 2). Contrary to our expectations, Coevol dN/dS scales negatively with genome size across the whole dataset (Slope = -0.287, adjusted-R^2 = 0.004, p-value = 0.039) and within insects (Slope = -1.241, adjusted-R^2 = 0.026, p-value = 0.018).

Surprisingly, different patterns are observed relative to TE content, as a negative correlation with Coevol dN/dS is detected across all species (Slope = -0.903, adjusted-R^2 = 0.004, p-value = 0.050) and within Mammalia (Slope = -2.113, adjusted-R^2 = 0.063, p-value = 0.001), while a positive correlation is found within Actinopteri (Slope = 3.407, adjusted-R^2 = 0.046, p-value = 0.013) (Table 2). Therefore, the two only significant positive PIC correlations found for birds and fishes are contrasted by results with an opposite trend in other groups. However, such correlations are slightly significant and their explained variance is extremely low.

Overall, we find no evidence for a recursive association of long-term N_e variation, as approximated by dN/dS, with genome size and TE content across the analysed animal taxa. PIC analysis without the 24 species with more than 30% of duplicated BUSCO genes produced similar results with dN/dS as independent variable (genome size: slope = 0.234, adjusted-R^2 = 0.002, p-value = 0.102; TE content: slope = 0.819, adjusted-R^2 = 0.004, p-value = 0.054; Recent TE content: slope = 2.002, adjusted-R^2 = 0.012, p-value = 0.003), indicating that genomic duplications have a negligible effect on the missing link between dN/dS and genome size.

2.6. Population size and genome size: a complex relationship across clades

Although no strong signal is found across the full dataset using PIC, different trends within different clades are suggested by both PIC and Coevol approaches.

Coevol infers a negative correlation of dN/dS with genome size in insects (GC3-poor: CC = -0.330, PP = 0.03; GC3-rich: CC = -0.110, PP = 0.24) and TE content in mammals (GC3-poor: CC = -0.220, PP = 0.08; GC3-rich: CC = -0.249, PP = 0.06), and a positive correlation (even though below significance threshold) with TEs in fishes (GC3-poor: CC = 0.192, PP = 0.88; GC3-rich: CC = 0.167, PP = 0.84). Additionally, a negative correlation with TE content is found in birds for the GC3-rich geneset, while a positive – yet not significant – trend is found using the GC3-poor geneset (GC3-poor: CC = 0.065, PP = 0.73; GC3-rich: CC = -0.195, PP = 0.03) (Table 3). Even though available for fewer species, LHTs partially support these trends for vertebrates: Actinopteri display a positive correlation between longevity and recent TE content (see 2.7). Instead, mammalian TE content correlates positively with metabolic rate and population density, and negatively with body length, mass, sexual maturity, age at first birth and longevity (Supplemental Table S5). Within Aves, Coevol predicts opposite results for genome size and TE content: genome size associates positively with longevity and mass, as well as with dN/dS, and negatively with metabolic rate, while TEs correlate positively with metabolic rate (and negatively with dN/dS in one case) (Supplemental Table S5). In summary, N_e seems to negatively affect TE content in fishes, and positively in mammals. Importantly, genome size correlations seem to follow the same trends of TE content in these groups, although correlations are weaker and mostly non-significant. In the case of birds, genome size seems rather to be explained by N_e as expected by MHH but not TE content, which instead might have the opposite trend.

2.7. dN/dS weakly correlates with the recent TE content

The global TE content integrally reflects a long history of TE insertions and deletions. To have a glance at the dynamics of TEs on an evolutionary time comparable to that of the level of drift estimated using dN/dS, we additionally examined the quantity of the youngest elements. From the overall TE insertions, we estimated a recent TE content, defined by reads with less than 5% of divergence from consensus, and included it among the traits to model with PIC and Coevol. In PIC analysis, the variation of recent TE content weakly associates with Bio++ dN/dS across the full dataset (Slope = 1.963, adjusted-R^2 = 0.012, p-value = 0.003) and in mammals (Slope = 4.115, adjusted-R^2 = 0.024, p-value = 0.031). On the contrary, using Coevol dN/dS this correlation is found in Aves (Slope = 4.982, adjusted-R^2 = 0.028, p-value = 0.006) and Actinopteri (Slope = 4.365, adjusted-R^2 = 0.061, p-value = 0.005), but the opposite is detected in mammals (Slope = -3.151, adjusted-R^2 = 0.024, p-value = 0.032) (Table 2). In agreement with PIC, Coevol reconstruction retrieves a positive association of recent TE content with dN/dS and longevity only in fishes, and a relationship opposite to expectations with dN/dS (CC = -0.296, PP = 0.06) and LHTs in mammals. In contrast with PIC results, a negative correlation between recent TE content and dN/dS is found for birds using the GC3-rich genesets (CC = -0.240, PP = 0.01) (Table 3; Supplemental Table S5).

On the whole, only a very weak positive correlation of dN/dS with recent TE insertions is observed across all species. However, considering again the taxa separately, clade-specific patterns emerge: a negative association between population size proxies and recent TE content is jointly found by the two methods only in fishes. Conversely, mammals show a positive correlation between recent TE content and population size proxies. Therefore, the coevolution patterns between population size and recent TE content are consistent with the pictures emerging from the comparison of population size proxies with genome size and overall TE content in the corresponding clades.

3. Discussion

Our results demonstrate the absence of a negative relationship between genome size and effective population size across a large dataset of animals, in contrast to the prediction of the MHH (Lynch and Conery 2003; Lynch 2007). Rather, our results highlight heterogeneous patterns within clades, with no consistent response of genome size and TE dynamics to N_e variations.

3.1. Assembly size underestimate genome size as genomes grow bigger

Assembly size is commonly used as a measure of genome size. However, the difficulty in assembling repetitive regions generally have it underestimate the actual genome size, in particular when only short reads are employed (Benham et al. 2024; Blommaert 2020; Peona et al. 2018). Consequently, methods that directly measure C-value such as flow cytometry (Doležel and Greilhuber 2010) and Feulgen densitometry (Hardie et al. 2002) are normally preferred as they do not rely on sequence data. Despite applying quality criteria to the assembly, the relationship between assembly size and genome size might still be questioned. However, we show that assembly size can overall approximate genome size quite well and, probably because we removed lower quality assemblies, no effect related to read type (short Illumina *vs* long ONT/PacBio) was detected. This suggests that the assemblies selected for our dataset can mostly provide a reliable measurement of genome size, and thus a quasi-exhaustive view of the genome architecture. On the other hand, because a gap with C-value is still present, we integrated this metric to correct assembly size estimations to their “expected C-values”. Similarly, the use of dnaPipeTE allowed us to quantify the repeat content without relying on assembly completeness. In summary, we extensively controlled for the effect of data quality on results and employed methods to minimize it.

3.2. Transposable elements are major contributors to genome size variation

Genome size and TE content have already been reported as tightly linked in eukaryotes (Elliott and Gregory 2015; Kidwell 2002), arthropods (Sproul et al. 2023; Wu and Lu 2019) and vertebrates (Chalopin et al. 2015). Our results are consistent with this perspective across all the animal species analysed, as well as at the level of ray-finned fishes (Reinar et al. 2023), insects (Heckenhauer et al. 2022; Mérel et al. 2024; Petersen et al. 2019; Sessegolo et al. 2016), mammals (Osmanski et al. 2023), and molluscs (Martelossi et al. 2023), strongly indicating TEs as major drivers of genome size variation in metazoans. It should be noted however that we mainly focused on some vertebrate groups and insects, while leaving out many animal taxa with fewer genomic resources currently available including much of the animal tree of life, such as most molluscs, annelids, sponges, cnidarians and nematodes. Even for better studied vertebrates, our datasets are far from comprehensive. For instance, the genomes of squamate reptiles are relatively stable in size but show a high variability in repeat content (Castoe et al. 2011; Pasquesi et al. 2018). A similar case is represented by bird genomes where, according to our analysis and consistently with other studies (Ji et al. 2022; Kapusta and Suh 2017), repeat content has a lower capacity to explain size compared to other clades. This could be due to satellites, whose contribution to genome size can be highly variable (Flynn et al. 2020; Pasquesi et al. 2018; Peona et al. 2023). While the remarkable conservation of avian genome sizes has prompted interpretations involving further mechanisms (see discussion below), dnaPipeTE is known to generally underestimate satellite content (Goubert et al. 2015). This bias is more relevant for those species that exhibit large fractions of satellites compared to TEs in their repeatome. For instance, the portions of simple and low complexity repeats estimated with dnaPipeTE are consistently smaller than those reported in previous analyses based on assembly annotation for some species, such as *Triatoma infestans* (0.46% *vs* 25%; 7 Mbp *vs* 400 Mbp), *Drosophila*

eugracilis (1.28% *vs* 10.89%; 2 Mbp *vs* 25 Mbp), *Drosophila albomicans* (0.06% *vs* 18 to 38%; 0.12 Mbp *vs* 39 to 85 Mbp) and some other *Drosophila* species (Pita et al. 2017; de Lima and Ruiz-Luano 2022; Supplemental Table S2). Although the accuracy of Coevol analyses might occasionally be affected by such underestimations, the effect is likely minimal on the general trends. Inability to detect ancient TE copies is another relevant bias of dnaPipeTE. However, the strong correlation between repeat content and genome size and the consistency of dnaPipeTE and earlGrey results, even in large genomes such as that of *Aedes albopictus*, indicate that dnaPipeTE method is pertinent for our large-scale analysis. Furthermore, such an approach is especially fitting for the examination of recent TEs, as this specific analysis is not biased by very repetitive new TE families that are problematic to assemble.

Another way for genomes to grow involves genomic duplications. Although a high proportion of duplicated BUSCO genes may indicate a low haplotype resolution of the assembly, many species with a high duplication score in our dataset correspond to documented duplication cases, suggesting that such BUSCO statistics may provide an insight into this biological process. However, the contribution of duplicated genes to genome size is minimal compared to the one of TEs, and removing species with high duplication scores does not affect our results: this implies that duplication is unlikely to be the factor causing the relationship between genome size and dN/dS to deviate from the pattern expected from the MHH. Across the animal species considered here, the activity of TEs is therefore a preponderant mechanism of DNA gain, and their evolutionary dynamics appear of prime importance in driving genome size variation.

3.3. Reduced selection efficacy is not associated with increased genome size and TE content

Our dN/dS calculation included several filtering steps by branch length and topology: indeed, selecting markers by such criteria appears to be an essential step to reconcile estimations

with different methodologies (M Bastian, F Bénitière, A Marino, pers. comm.; Bastian 2024). In addition, our analyses resulted to be robust to species pruning by deviant branch lengths. Müller et al. (2022) showed that recent N_e fluctuations might perturb the expected correlation between long- and short-term estimates of N_e. According to the nearly neutral theory, alleles that start at low relative frequencies have a mean fixation of ~4Ne generations, under the implicit assumption of constant N_e (Kimura and Ohta 1969). This implies that dN/dS, that accounts for the accumulation of substitutions over time, has a weaker sensitivity to short-term changes in N_e compared to estimates based on polymorphism (Müller et al. 2022). Additionally, inferences on simulated and empirical data showed that N_e changes along branches could be captured and generally recapitulated by dN/dS and LHTs in a framework similar to that of Coevol (Latrille et al. 2021). Accordingly, dN/dS assessments by Bio++ and Coevol are highly concordant between each other and with LHTs. Taken together, our results point at the dN/dS found with the two methods as reliable proxies of long-term N_e.

If TEs are ascribable to nearly-neutral mutations, a negative correlation of N_e with TE expansion, and consequently with genome size – equivalent to a positive association with dN/dS – is expected. However, no such correlations are observed across the sampled species. It is important to note how not treating species traits as non-independent leads to artifactual results (Figure 2B-C). For instance, mammals have on average small population sizes and the largest genomes. Conversely, insects tend to have large N_e and overall small genomes. With a high sampling power and phylogenetic inertia being taken into account, our meta-analysis clearly points at a phylogenetic structure in the data: the main clades are each confined to separate genome size ranges regardless of their dN/dS variation. The other way around, variability in genome size can be observed in insects, irrespective of their dN/dS. Relying on non phylogenetically corrected models based on a limited number of species (such as that available at the time of the MHH proposal) can thus result in a spurious positive scaling between genome size and N_e proxies. To account for a shallower phylogenetic scale, we isolated recently active elements and at the same time explored the same relationships within each clade. Indeed, while the selective effect of elements might be slightly negative as

long as they are active, TEs accumulated over long periods of time might be subject to changing dynamics: in the latter case, the pace of sequence erosion could be in the long run independent of drift and lead to different trends of TE retention and degradation in different lineages. Extracting recent elements should thus allow us to have a glimpse of the latest TE colonizations. A positive scaling between the quantity of young TEs and dN/dS found in some cases indicates that relatively recent expansions of TEs could be subject to a more effective negative selection. However, this trend is always very weak and often summarizes that of full TE content within clades. A potential limit of this analysis lies in the application of the same similarity threshold to all species to delimit recent elements. While this is not problematic when comparing species that recently split apart (e.g., Yang et al. 2024), some noise might be introduced at large scale, as the quantity of young repeats that evolved on the same time scale can vary according to the mutation rate and generation time of a species.

Interestingly, the correlation patterns between population size proxies and genomic traits emerging within single clades are distinct and sometimes opposite to the expectations of MHH. Mammals display a negative correlation of dN/dS with TE content, a pattern that is uniformly confirmed by LHTs. Not only does this result corroborates previous findings of no relationship between N_e and genome size in mammals (Roddy et al. 2021), but it supports a correlation opposite to the predictions of the MHH. On the other hand, the observed positive scaling between dN/dS and TE content in ray-finned fishes might lend support to a role of drift on genome size in this clade: this result is also consistent with a previous study which found a negative scaling between genome size and heterozygosity in this group (Yi and Streelman 2005). In birds, population size seems to negatively affect genome size and positively the TE content, a decoupling that is however not surprising given the higher variation of TE load compared to the restricted genome size range. Contrasting signals from the two genomic traits have already been observed by Ji et al. (2022) who also reported a positive correlation between assembly size and mass, but a negative correlation between TE content and generation time. As previous studies find relatively weak correlations between TE content and genome size in birds (Ji et al. 2022; Kapusta and Suh 2017), it is possible for

the very narrow variation of the avian genome sizes to impair the detection of consistent signals. On the other hand, it is conceivable the avian TE diversity to be underappreciated due to the limits of sequencing technologies used so far in resolving complex repeat-rich regions. For instance, employment of long-reads technologies allowed to reveal more extended repeated regions that were previously ignored with short read assemblies (Kapusta and Suh 2017; Benham et al. 2024). Besides, quite large fractions might indeed be satellite sequences constituting relevant fractions of the genome that are challenging to identify with reference- or read-based methods (Edwards et al. 2025). An “accordion” dynamic has been proposed whereby higher TE loads are paralleled by equally strong deletional pressures, which could contribute to the maintenance of remarkably small and constant genome sizes in birds, in spite of ongoing TE activity (Kapusta et al. 2017; Kapusta and Suh 2017). Finally, the diffused evidence for a positive and a negative correlation of genome size with body mass and metabolic rate, respectively, is also compatible with the adaptationist perspective of powered flight indirectly maintaining small genome sizes in birds as a consequence of the metabolic needs (Wright et al. 2014; Zhang and Edwards 2012). In insects, dN/dS scales negatively with genome size, but never with TE content. As eusociality appears to bring about selection relaxation (Imrit et al. 2020; Kapheim et al. 2015; Weyna and Romiguier 2021), several studies explored the link between N_e and genome size in this taxon by focussing on social complexity as a proxy, but with contrasting outcomes: Mikhailova et al. (2023) find bigger genome size associated with eusociality in Hymenoptera, but the opposite trend in Blattodea; in contrast and partially in accordance with our findings, Kapheim et al. (2015) and Koshikawa et al. (2008) report less abundant TEs in eusocial hymenopters and smaller genomes in eusocial termites, respectively. While the approximation of N_e based on dN/dS should allow for a quantification of selection efficacy in wider terms than sociality traits, the investigated evolutionary scale might hold an important role in the outcome of such analyses. First, and specifically relative to insects, genome size seems to be subject to different evolutionary pressures – either selective or neutral – within different insect orders (Cong et al. 2022), implying that increased drift might not necessarily produce the same

effect on genome size across all insect groups. More generally, the five defined clades cover quite different time scales: insects and molluscs have much more ancient origins than mammals and birds, and such distant groups also evolve at very different evolutionary rates, making it difficult to characterize the evolution of their traits on the same evolutionary scale. Nevertheless, the results are still valuable in highlighting the absence of relationship between genetic drift and genome size variation in the long-term evolution of such broad groups, in contrast to previous work focusing at the population level or on recently diverged species (Cui et al. 2019; Mérel et al. 2021; Yang et al. 2024). At the same time, as noted by Mérel et al. (2024), comparing very distantly related species - as the insect and molluscan species of our dataset - might overshadow any relationship between genome size and N_e, either due to dN/dS predicting power being weakened by branch saturation, deep N_e fluctuations not being detected by our methods, or to additional factors affecting long-term genome size evolution.

3.4. Do lineage-specific TE dynamics affect genome size evolution?

Our findings do not support the quantity of non-coding DNA being driven in a nearly-neutral fashion by genetic drift. Notably, these results not only reject the theory of extra non-coding DNA being costly for its point mutational risk, but also challenges the more general idea of its accumulation depending on other kinds of detrimental effects, such as increased replication, pervasive transcription, or ectopic recombination. Therefore, our results can be considered more general than a mere rejection of the MHH hypothesis, as they do not support any theory predicting that species with low N_e would accumulate more non-coding DNA. In agreement with previous analyses (Pasquesi et al. 2018), we find that the proliferation of TEs in particular can, under comparable drift levels, give place to lineage-specific outcomes that mostly do not seem to depend on effective population size. These results contrast with those of other large-scale analyses which instead support the predictions of the drift-barrier hypothesis for a general impact of N_e on other genomic features, notably mutation rate

(Bergeron et al. 2023; Lynch et al. 2023; Wang and Obbard 2023) and splicing accuracy (Bénitière et al. 2024). To put this in perspective, it should be emphasized that, in the framework of the MHH, the success of nearly-neutral alleles depends on the combination of both N_e and liability to mutation of non-coding DNA (Lynch et al. 2011). Overall, we studied N_e variation without accounting for the different mutagenic burden posed by non-coding DNA across different lineages. In the case of TEs, inherently assuming the same distribution of selective effect and a constant activity in all species and among TE insertions was assumed. However, it is known that TEs are subject to waves of activity rather than a uniform pace of transposition (Arkhipova 2018). Moreover, given the broad phylogenetic scale of our dataset, it is likely for different levels of hazard to be acting across genomes due to different "host-parasite" dynamics in different animal groups (Ågren and Wright 2011). Such coevolutionary dynamics are, for example, determined by TE silencing mechanisms, which evolve differently across lineages and might influence the degree of genome expansion (Lechner et al. 2013; Zhou et al. 2020; Wang et al. 2023). In general, because of their complex interactions with genomes, TEs are especially likely to deviate from the assumption of gradually mutating sequences. Therefore, treating them as universally slightly deleterious alleles might be an oversimplified model. For instance, while the big genomes of salamanders are not related to small N_e, the low synonymous substitution rates and low degree of deletions due to ectopic recombination suggest weak mutational hazard of TEs that possibly contributes to the maintenance of genomic gigantism in this group (Frahry et al. 2015; Mohlhenrich and Mueller 2016; Rios-Carlos et al. 2024). Additionally, lineage-specific TE dynamics themselves might underlie different genomic architectures: for example, mammalian genomes are generally characterized by one preponderant type of active element and by a long-term retention of old TEs (Osmanski et al. 2023; Sotero-Caio et al. 2017), as in human where a very small proportion of active elements (<0.05%) is unlikely to impose a mutation rate causing genome size variation (Mills et al. 2007). Conversely, squamate and teleost fish genomes are smaller and characterized by several, simultaneously active and less abundant TE types (Duvernell et al. 2004; Furano et al. 2004; Novick et al.

2009; Pasquesi et al. 2018; Volff et al. 2003). These different patterns of genomic organization seem overall associated with different rates of elements' turnover (Blass et al. 2012; Lavoie et al. 2013; Novick et al. 2009; Volff et al. 2003). All such variables might alter the selective effect and differentiate TEs from gradually and constantly evolving alleles, eventually contributing to the lack of association between long-term N_e and genome size. Finally, Kapusta et al. (2017) showed that large-scale deletions can be as important as DNA gain in determining genome size, thus questioning the assumption of the rate of elements insertion being greater than their removal rate (Lynch 2007). This implies that the contribution of TEs constitutes just one side of the coin and that deletion bias could also drive the divergence of genome size across lineages, as suggested by several studies linking negatively deletion rates with genome size (Frahry et al. 2015; Ji et al. 2022; Kapusta et al. 2017; Wang et al. 2014).

3.5. Perspectives

Evidence for signatures of negative selection against TE proliferation exist at various degrees. In *Anolis* lizards, the ability of TEs to reach fixation can vary between populations of the same species according to population size (Ruggiero et al. 2017; Tollis and Boissinot 2013). Furthermore, N_e was found to negatively correlate with genome size and TE expansion at the intraspecific level in *Drosophila suzukii* (Mérel et al. 2021) and at the interspecific level in fruit flies (Mérel et al. 2024), asellid isopods (Lefébure et al. 2017) and killifishes (Cui et al. 2019), supporting the role of genetic drift in determining recent differences in genome size among closely related animal species. Given the very different taxonomic scale of such works and ours, and with the perspective of lineage-specific interaction between genome and genomic parasites in mind, our negative results for the MHH at metazoan scale are not incompatible with an effect of N_e on genome size within specific clades. In a nutshell, although an increase in genetic drift seems to lead to the short-term accumulation of transposable elements, this process is not visible in the long-term,

suggesting that it fades over time. A general mechanism of selection preventing the proliferation of non-coding DNA and TEs in animals might exist but its results be detectable only at a sufficiently short evolutionary time. In this sense, the lack of evidence for MHH in other clade-specific studies might be due to the methodological challenges of either estimating a suitable marker of N_e or investigating too distantly related lineages. Moreover, the contrasting outcomes of such studies might reflect an actual variability in the selective effect of TEs not compatible with a general selection mechanism. Further reducing the phylogenetic scale under study and systematically exploring the consequences of N_e variation within independent biological systems could therefore provide an alternative way to test the impact of drift, while removing the confounding effects due to different genomic backgrounds.

4. Methods

4.1. Dataset

All the metazoan reference assemblies available as of November 14th 2021 were used, except for insect genomes which were drawn from Sproul et al. (2023), for a total of 3,214 assemblies. For each assembly, quality metrics were computed with Quast 5.0.2 (Gurevich et al. 2013) and genome completeness was assessed with BUSCO 5.2.2 using the 954 markers of the metazoa_odb10 geneset (Supplemental Table S1). Availability of raw reads was verified with SRA Explorer (https://github.com/ewels/sra-explorer). All the assemblies with either a contig N50 smaller than 50 kb, less than 70% of complete BUSCO orthologs, or without available reads were excluded from TE and dN/dS analyses. The subdivision into Actinopteri (N = 148), Aves (N = 260), Insecta (N = 189), Mammalia (N = 182), Mollusca (N = 28) was adopted to perform alignments, phylogenies, dN/dS estimation with Bio++ and Coevol runs (see below).

4.2. Genome size estimation

Assembly sizes and C-values were jointly used to estimate genome size. C-values measured by either flow cytometry (FCM), Feulgen densitometry (FD) or Feulgen image analysis densitometry (FIA) were collected from https://www.genomesize.com/ (last accessed 6 october 2022) for all available species of our initial dataset with contig N50 ≥ 50 kb, totalling 465 measurements for 365 species (Supplemental Table S3). To assign a unique C-value, when multiple values were present for one species, the most recent one was retained and, if dates were the same, the average was used. For all the other species having contig N50 ≥ 50 kb but with no available C-value record, genome size was calculated as an expected C-value predicted from a WLS where the 465 FCM, FD and FIA estimations were the dependent variables and assembly sizes were the independent variables (for details see https://github.com/albmarino/Meta-analysis_scripts). Out of all the records in this training dataset for genome size, 93 correspond to ray-finned fishes, 93 to mammals, 92 to birds, 106 to insects, and 9 to molluscs, overall mirroring the taxa represented in the final dataset. For the purpose of our analysis, C-values were used for the species for which such data were available, while the expected C-value was used as genome size estimation in all the other cases, regardless of the type of sequencing data used for the assembly (Supplemental Tables S1, S3).

4.3. Gene alignment

The 954 annotated single-copy BUSCO genes were aligned with the pipeline OMM_MACSE 11.05 using MACSE 2.06 (Ranwez et al. 2018; Scornavacca et al. 2019). Alignments were performed separately within each clade - Actinopteri, Aves, Insecta, Mammalia and Mollusca.

4.4. Phylogeny

Phylogenies were computed separately for each clade with IQ-TREE 1.6.12 (Nguyen et al. 2015). JTT+F+R10 substitution model was selected with ModelFinder (-m MFP option) (Kalyaanamoorthy et al. 2017). For reasons of computing power and time, we have reconstructed the phylogenies of each clade independently and then grouped them together to create a single complete phylogenetic tree (see below). The same set of 107 concatenated BUSCO amino-acid sequences was used to calculate all the phylogenies. However, since this produced a spurious relationship in the mammalian tree with paraphyly of primates, an alternative set of randomly selected 100 genes was used instead for the phylogeny of Mammalia (Supplemental Table S6). Each phylogeny was rooted using an outgroup species belonging to its respective sister clade: the outgroup sequences were added to gene alignments with the enrichAlignment function from MACSE; the outgroup + clade gene alignments were concatenated and used to recompute the outgroup + clade phylogeny taking into account the previously computed tree topology of the clade. The outgroups were then removed, and rooted clade phylogenies were merged together manually using the tree editor program Baobab (Dutheil and Galtier 2002). Finally, 50 top-shared genes across all species were chosen among the set of 107 genes (Supplemental Table S6) to recalculate the branch lengths of the whole dataset phylogeny: with the MACSE program alignTwoProfiles the nucleotide gene alignment of one clade was joined to the one of its respective sister clade until achievement of the whole dataset alignment. Branch lengths were then estimated based on the 50-genes concatenate and the tree topology (for the detailed workflow, see https://github.com/albmarino/Meta-analysis_scripts).

4.5. dN/dS estimation

When genetic drift is strong, slightly deleterious mutations are more likely to reach fixation than under conditions of high N_e and more effective selection (Ohta 1992). The genome-wide

fixation rate of non-synonymous mutations is expected to drive the dN/dS ratio due to nearly-neutral mutations responding to different N_e: a higher dN/dS accounts for more frequent accumulation of mildly deleterious mutations over time due to small N_e, while lower dN/dS is associated with a stronger effect of selection against slightly deleterious non-synonymous mutations due to high N_e (James et al. 2016; Romiguier et al. 2014; Weyna and Romiguier 2021; Woolfit and Bromham 2005). This is also supported at the polymorphism level, with higher pN/pS and accumulation of slightly deleterious mutations in smaller populations (Leroy et al. 2021; Dussex et al. 2023).

Before dN/dS calculation, sequences with more than 10% of their length occupied by insertions were preemptively removed from BUSCO alignments. Estimation of dN/dS on either very long or short terminal branches might lead to loss of accuracy due to branch saturation (Weber et al. 2014) or to a higher variance of substitution rates, respectively. Furthermore, shared polymorphism can be captured in the substitution rates when closely related species are compared, and further contribute to bias dN/dS (Mugal et al. 2020). To correct for such issues, genes with deviant topology were identified and removed from every clade with PhylteR using default parameters (Comte et al. 2023). Moreover, genes exhibiting branch lengths shorter than 0.001, for which dN/dS could have a large variance, were also not integrated in the dN/dS calculation of a species.

We then used bppml and mapnh from the Bio++ libraries to estimate dN/dS on terminal branches (Dutheil et al. 2006; Guéguen et al. 2013; Guéguen and Duret 2018; Romiguier et al. 2012). bppml calculates the parameters under a homogenous codon model YN98 (F3X4). Next, mapnh maps substitutions along the tree branches and estimates dN and dS. More precisely, the substitution rate is given by the number of substitutions mapped according to the model parameters normalized by the number of substitutions of the same category (i.e. synonymous, non-synonymous) that would occur under the same neutral model (Bolívar et al. 2019). Therefore, dN and dS are calculated for each species as follows:

$$\frac{\sum_{i=1}^{n} K(i)}{\sum_{i=1}^{n} \frac{O(i)}{l(i)}}$$

Where *n* is the number of genes, *K* is the substitutions count as mapped by the substitution model calculated for the gene, *O* is the substitutions count as mapped under the same neutral model, and *l* is the branch length of the given species for that gene. In addition to the gene filtering described above, Bio++ dN/dS was recalculated on a reduced dataset where the longest (> 1) and shortest (< 0.01) branches were removed, in order to ensure that substitution saturation and segregating polymorphism did not influence the results. Terminal branches with more than 1 and less than 0.01 amino-acid substitutions per site were removed, and dN/dS recalculated on the trimmed phylogenies with the same method described above.

The same metric was estimated with Coevol 1.6 (Lartillot and Poujol 2011). Coevol models the co-evolution of dN/dS and continuous traits along branches following a multivariate Brownian diffusion process, thus reducing the variance in the dN/dS of the smallest branches (Lartillot and Delsuc 2012). Bio++ dN/dS was therefore compared with dN/dS estimated by Coevol on terminal branches to verify the consistency between the two methods.

4.6. Compilation of life history traits

Life-history traits – body mass, longevity, generation time, among others – are found to be related to N_e in mammals, birds, and amniotes in general (Bolívar et al. 2019; Figuet et al. 2016; Nikolaev et al. 2007; Popadin et al. 2007). Available LHTs were assigned to the species of our dataset using information from several resources. Adult body mass, body length, maximum longevity, basal metabolic rate, age at first birth, population size and population density were assigned to mammalian species using PanTHERIA (Jones et al. 2009). For birds, body mass information was extracted from Dunning (2007). Shallow to deep depth range, longevity in the wild, body length and body mass were compiled for ray-finned fishes from www.fishbase.org using the rfishbase R package (Boettiger et al. 2012).

Additionally, age at sexual maturity, adult body mass, maximum longevity and metabolic rate were extracted from AnAge (Tacutu et al. 2018), as well as body mass and metabolic rate from AnimalTraits (Herberstein et al. 2022): such data were used to complement information when missing from the databases cited above. All the retrieved LHTs and their relative source are reported in Supplemental Table S2.

4.7. TE quantification

TEs were annotated with a pipeline employing dnaPipeTE (Goubert et al. 2015) in two rounds (https://github.com/sigau/pipeline_dnapipe). Raw reads are filtered with UrQt (Modolo and Lerat 2015) or fastp (Chen et al. 2018) and undergo a first “dirty” dnaPipeTE round. The obtained dnaPipeTE contigs are mapped against a database of organellar, fungal, bacterial and archaean reference sequences with Minimap2 (Li 2018) and the original reads matching contaminant sequences removed with SAMtools (Danecek et al. 2021). Finally, the quality- and contaminant-filtered reads are used to perform a second “clean” dnaPipeTE round. dnaPipeTE was configured with the Dfam 3.5 and RepBaseRepeatMaskerEdition-20181026 repeat libraries and was run with a genome coverage of 0.25.

To verify the consistency of dnaPipeTE estimations, the dnaPipeTE-based pipeline was benchmarked on a subset of 29 dipteran species against EarlGrey 1.3, an automated pipeline performing TE annotation on genome assemblies (Baril et al. 2024). EarlGrey was configured with the same libraries as dnaPipeTE and was run with ‘metazoa’ as search term.

The results of the second dnaPipeTE round were used to extrapolate the total and recent TE content, the latter being defined by all the reads below 5% of divergence from the corresponding consensus. The two contents were extracted by adapting dnaPT_landscapes.sh from the dnaPT_utils repository (https://github.com/clemgoub/dnaPT_utils) to a custom R script (https://github.com/albmarino/Meta-analysis_scripts).

4.8. Gene duplication

To account for the effect of whole genome or segmental duplications, we used the BUSCO Duplicated score: if big-scale duplication events recently took place, a higher score should be observed genome-wide even for conserved genes. As many of the genomes with BUSCO Duplicated score above 30% corresponded to reported cases of genomic duplication, we used this threshold to perform PIC analysis with and without species whose genome size is potentially more affected by duplication events.

4.9. Phylogenetic independent contrasts and Coevol reconstruction

The correlations of Bio++ and Coevol dN/dS with LHTs, as well as with genome size and TE content, were tested with PIC to correct for the covariation of traits due to the phylogenetic relatedness of species (Felsenstein 1985). PIC were performed on the whole dataset, the trimmed dataset, and within every clade with the R packages ape 5.7.1 (Paradis et al. 2004), nlme 3.1.162, and caper 1.0.1. Results were plotted with the ggplot2 3.4.2 package. Additionally, Coevol 1.6 was run to test for the coevolution of dN/dS and traits: sequence substitution processes and quantitative traits such as LHTs, genome size, TE content, are here assumed to covary along the phylogeny as a multivariate Brownian motion process. Coevol infers trait values on internal nodes and terminal branches (those used for PIC), as well as correlation coefficients and their relative posterior probabilities. Due to computational limitations, Coevol analysis was carried out separately on every clade and on a limited number of genes. Genes were selected according to their GC content at the third codon position (GC3). Indeed, mixing genes with heterogeneous base composition in the same concatenate might result in an alteration of the calculation of codon frequencies, and consequently impair the accuracy of the model estimating substitution rates (Mérel et al. 2024). Moreover, genes with different GC3 levels can reflect different selective pressures, as

highly expressed genes should be enriched in optimal codons as a consequence of selection on codon usage. In *Drosophila*, where codon usage bias is at play, most optimal codons present G/C bases at the third position (Duret and Mouchiroud, 1999), meaning that genes with high GC3 content should evolve under stronger purifying selection than GC3-poor genes. Accordingly, Mérel et al. (2024) do find a stronger relationship between dN/dS and genome size when using GC3-poor genes, as compared to GC3-rich genes or gene concatenates of random GC3 composition. Finally, dN/dS can be influenced by GC-biased gene conversion (Bolívar et al. 2019; Ratnakumar et al. 2010), and the strength at which such substitution bias acts can be reflected by base composition. For these reasons, two sets of 50 genes with similar GC3 content were defined in order to employ genes undergoing similar evolutionary regimes. Markers and species were subject to a more stringent filtering in order to have as much information as possible for each species. 16 species with less than 50% of single-copy orthologs were further filtered out from the dataset. In addition to the PhylteR step, only genes represented in at least 95% of the species of a clade were retained. From those, the 50 GC3-poorest and the 50 GC3-richest genes were chosen. Coevol was then run with both the gene sets for every clade. Convergence of the MCMC chains were checked visually by plotting the evolution of statistics. Likelihood values and correlations were estimated, running the chains for a minimum of 1,000 steps and discarding the first 400 steps as burn-in.

5. Data access

The accession numbers, SRA experiments and C-values used in this study are reported in Supplemental Table S1. C-values are available at https://www.genomesize.com/. The respective sources of LHTs are reported in Supplemental Table S2. Detailed commands and custom scripts can be found at https://github.com/albmarino/Meta-analysis_scripts.

6. Competing interest statement

The authors declare that they have no conflict of interest.

7. Acknowledgements

We are grateful to Laurent Duret, Nicolas Lartillot, Tristan Lefébure, Mélodie Bastian and Florian Bénitière for the helpful discussions. We thank Mélodie Bastian and Florian Bénitière for methodological exchanges to ensure to have a dN/dS accurately estimated. We thank Laurent Guéguen for his help with dN/dS analyses. Analyses were run on the servers of the Montpellier Bioinformatics Biodiversity platform (MBB). We are grateful to three anonymous reviewers whose comments helped improve the quality of this manuscript.

This research was funded by the French National Research Agency (ANR-20-CE02-0008-01 ”NeGA”). A CC-BY public copyright licence has been applied by the authors to the present document and will be applied to all subsequent versions up to the Author Accepted Manuscript arising from this submission, in accordance with the grant’s open access conditions.

8. References

Ågren JA, Wright SI. 2011. Co-evolution between transposable elements and their hosts: a major factor in genome size evolution? *Chromosome Res* **19(6):** 777–786.

Arkhipova IR. 2018. Neutral Theory, Transposable Elements, and Eukaryotic Genome Evolution. *Mol Biol Evol* **35(6):** 1332–1337.

Baril T, Galbraith J, Hayward A. 2024. Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline. *Mol Biol Evol*, **41(4):** msae068. doi: 10.1093/molbev/msae068

Bast J, Schaefer I, Schwander T, Maraun M, Scheu S, Kraaijeveld K. 2016. No Accumulation of Transposable Elements in Asexual Arthropods. *Mol Biol Evol* **33(3):** 697–706.

Bastian M. 2024. Génomique des populations intégrative: de la phylogénie à la génétique des populations. Doctoral dissertation, Université Lyon 1.

Benham PM, Cicero C, Escalona M, Beraut E, Fairbairn C, Marimuthu MPA, Nguyen O, Sahasrabudhe R, King BL, Thomas WK et al. 2024. Remarkably High Repeat Content in the Genomes of Sparrows: The Importance of Genome Assembly Completeness for Transposable Element Discovery. *Genome Biol Evol*, **16(4):** evae067. doi: 10.1093/gbe/evae067

Bénitière F, Necsulea A, Duret L. 2024. Random genetic drift sets an upper limit on mRNA splicing accuracy in metazoans. *Elife 13*, RP93629. doi: 10.7554/eLife.93629.3

Bennett MD, Riley R. 1997. The duration of meiosis. *Proc Royal Soc B* **178(1052):** 277–299.

Blass E, Bell M, Boissinot S. 2012. Accumulation and Rapid Decay of Non-LTR Retrotransposons in the Genome of the Three-Spine Stickleback. *Genome Biol Evol* **4(5):** 687–702.

Bergeron LA, Besenbacher S, Zheng J, Li P, Bertelsen MF, Quintard B, Hoffman JI, Li Z, St Leger J, Shao C, et al. 2023. Evolution of the germline mutation rate across vertebrates. *Nature* **615(7951)**: 285-291.

Blommaert J. 2020. Genome size evolution: towards new model systems for old questions. *Proc R Soc B Biol Sci* **287(1933):** 20201441. doi: 10.1098/rspb.2020.1441

Boettiger C, Lang DT, Wainwright PC. 2012. rfishbase: exploring, manipulating and visualizing FishBase data from R. *J Fish Biol* **81(6):** 2030–2039.

Bolívar P, Guéguen L, Duret L, Ellegren H, Mugal CF. 2019. GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biol* **20(1):** 1-13.

Brevet M, Lartillot N. 2021. Reconstructing the History of Variation in Effective Population Size along Phylogenies. *Genome Biol Evol* **13(8):** evab150. doi: 10.1093/gbe/evab150

Buffalo V, 2021. Quantifying the relationship between genetic diversity and population size suggests natural selection cannot explain Lewontin's Paradox. *Elife* **10:** e67509. doi: 10.7554/eLife.67509

Castoe TA, Hall KT, Guibotsy Mboulas ML, Gu W, de Koning APJ, Fox SE, Poole AW, Vemulapalli V, Daza JM, Mockler T et al. 2011. Discovery of Highly Divergent Repeat Landscapes in Snake Genomes Using High-Throughput Sequencing. *Genome Biol Evol* **3:** 641–653.

Cavalier-Smith T. 1978. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *J Cell Sci* **34(1):** 247–278.

Chak STC, Harris SE, Hultgren KM, Jeffery NW, Rubenstein DR. 2021. Eusociality in snapping shrimps is associated with larger genomes and an accumulation of transposable elements. *PNAS* **118(24):** e2025051118. doi: 10.1073/pnas.2025051118

Chalopin D, Naville M, Plard F, Galiana D, Volff J-N. 2015. Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates. *Genome Biol Evol* **7(2):** 567–580.

Charlesworth B, Barton N. 2004. Genome Size: Does Bigger Mean Worse? *Curr Biol* **14(6):** R233–R235.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34(17):** i884–i890.

Comte A, Tricou T, Tannier E, Joseph J, Siberchicot A, Penel S, Allio R, Delsuc F, Dray S, De Vienne DM. 2023. PhylteR: efficient identification of outlier sequences in phylogenomic datasets. *Mol Biol Evol* **40(11):** msad234. doi: 10.1093/molbev/msad234

Cong Y, Ye X, Mei Y, He K, Li F. 2022. Transposons and non-coding regions drive the intrafamily differences of genome size in insects. *iScience* **25(9):** 104873. doi: 10.1016/j.isci.2022.104873

Cui R, Medeiros T, Willemsen D, Iasi LNM, Collier GE, Graef M, Reichard M, Valenzano DR. 2019. Relaxed Selection Limits Lifespan by Increasing Mutation Load. *Cell* **178(2):** 385-399.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* **10(2):** giab008. doi: 10.1093/gigascience/giab008

Daubin V, Moran NA. 2004. Comment on “The Origins of Genome Complexity.” *Science* **306(5698):** 978–978.

de Lima LG, Ruiz-Ruano FJ. 2022. In-depth satellitome analyses of 37 Drosophila species illuminate repetitive DNA evolution in the Drosophila genus. *Genome Biol Evol* **14(5)**: evac064. doi: 10.1093/gbe/evac064

Doležel J, Greilhuber J. 2010. Nuclear genome size: Are we getting closer? *Cytometry A*

77(7): 635–642.

Du K, Stöck M, Kneitz S, Klopp C, Woltering JM, Adolfi MC, Feron R, Prokopov D, Makunin A, Kichigin I et al. 2020. The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization. *Nat Ecol Evol* **4(6):** 841–852.

Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *PNAS* **96(8):** 4482-4487.

Dussex N, Morales HE, Grossen C, Dalén L, van Oosterhout C. 2023. Purging and accumulation of genetic load in conservation. *Trends Ecol Evol* **38(10):** 961-969.

Dutheil J, Gaillard S, Bazin E, Glémin S, Ranwez V, Galtier N, Belkhir K. 2006. Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinform* **7(1):** 1-6.

Dutheil J, Galtier N. 2002. BAOBAB: a Java editor for large phylogenetic trees. *Bioinformatics* **18(6):** 892–893.

Duvernell DD, Pryor SR, Adams SM. 2004. Teleost Fish Genomes Contain a Diverse Array of L1 RetrotransposonLineages That Exhibit a Low Copy Number and High Rate of Turnover. *J Mol Evol* **59(3):** 298–308.

Edwards SV, Fang B, Khost D, Kolyfetis GE, Cheek RG, DeRaad D, Chen N, Fitzpatrick JW, McCormack JE, Funk WC et al. 2025. Comparative population pangenomes reveal unexpected complexity and fitness effects of structural variants. bioRxiv doi: 10.1101/2025.02.11.637762

Elliott TA, Gregory TR. 2015. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Proc R Soc B Biol Sci* **370(1678):** 20140331. doi: 10.1098/rstb.2014.0331

Felsenstein J. 1985. Phylogenies and the Comparative Method. *Am Nat* **125(1):** 1–15.

Figuet E, Nabholz B, Bonneau M, Mas Carrio E, Nadachowska-Brzyska K, Ellegren H, Galtier N. 2016. Life History Traits, Protein Evolution, and the Nearly Neutral Theory in Amniotes. *Mol Biol Evol* **33(6):** 1517–1527.

Flynn JM, Long M, Wing RA, Clark AG. 2020. Evolutionary Dynamics of Abundant 7-bp Satellites in the Genome of Drosophila virilis. *Mol Biol Evol* **37(5):** 1362–1375.

Frahry MB, Sun C, Chong RA, Mueller RL. 2015. Low Levels of LTR Retrotransposon Deletion by Ectopic Recombination in the Gigantic Genomes of Salamanders. *J Mol Evol* **80(2):** 120–129.

Furano AV, Duvernell DD, Boissinot S. 2004. L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. *Trends Genet* **20(1):** 9–14.

Galtier N. 2024. Half a Century of Controversy: The Neutralist/Selectionist Debate in Molecular Evolution. *Genome Biol Evol* **16(2):** evae003. doi: 10.1093/gbe/evae003

Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M. 2015. De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes aegypti). *Genome Biol Evol* **7(4):** 1192–1205.

Graur D, Zheng Y, Azevedo RBR. 2015. An Evolutionary Classification of Genomic Function. *Genome Biol Evol* **7(3):** 642–645.

Gregory TR. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol rev biol proc Camb Philos Soc* **76(1):** 65–101.

Gregory TR. 2002. Genome size and developmental complexity. *Genetica* **115(1):** 131–146.

Gregory TR, Hebert PDN. 1999. The Modulation of DNA Content: Proximate Causes and Ultimate Consequences. *Genome Res* **9(4):** 317–324.

Guéguen L, Duret L. 2018. Unbiased Estimate of Synonymous and Nonsynonymous

Substitution Rates with Nonstationary Base Composition. *Mol Biol Evol* **35(3):** 734–742.

Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, Fournier D, Pouyet F, Cahais V et al. 2013. Bio++: Efficient Extensible Libraries and Tools for Computational Molecular Evolution. *Mol Biol Evol* **30(8):** 1745–1750.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29(8):** 1072–1075.

Hardie DC, Gregory TR, Hebert PDN. 2002. From Pixels to Picograms: A Beginners' Guide to Genome Quantification by Feulgen Image Analysis Densitometry. *J Histochem Cytochem* **50(6):** 735–749.

Heckenhauer J, Frandsen PB, Sproul JS, Li Z, Paule J, Larracuente AM, Maughan PJ, Barker MS, Schneider JV, Stewart RJ, Pauls SU. 2022. Genome size evolution in the diverse insect order Trichoptera. *GigaScience* **11:** giac011. doi: 10.1093/gigascience/giac011

Herberstein ME, McLean DJ, Lowe E, Wolff JO, Khan MK, Smith K, Allen AP, Bulbert M, Buzatto BA, Eldridge MDB et al. 2022. AnimalTraits - a curated animal trait database for body mass, metabolic rate and brain size. *Sci Data* **9(1):** 265. doi: 10.1038/s41597-022-01364-9

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35(2):** 518–522.

Imrit MA, Dogantzis KA, Harpur BA, Zayed A. 2020. Eusociality influences the strength of negative selection on insect genomes. *Proc Royal Soc B* **287(1933):** 20201512. doi: 10.1098/rspb.2020.1512

James JE, Lanfear R, Eyre-Walker A. 2016. Molecular Evolutionary Consequences of Island Colonization. *Genome Biol Evol* **8(6):** 1876–1888.

Ji Y, Feng S, Wu L, Fang Q, Brüniche-Olsen A, DeWoody JA, Cheng Y, Zhang D, Hao Y, Song G et al. 2022. Orthologous microsatellites, transposable elements, and DNA deletions

correlate with generation time and body mass in neoavian birds. *Sci Adv* **8(35):** eabo0099. doi: 10.1126/sciadv.abo0099

Jockusch EL. 1997. An evolutionary correlate of genome size change in plethodontid salamanders. *Proc R Soc B Biol Sci* **264(1381):** 597–604.

Jones KE, Bielby J, Cardillo M, Fritz SA, O'Dell J, Orme CDL, Safi K, Sechrest W, Boakes EH, Carbone C et al. 2009. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* **90(9):** 2648–2648.

Dunning Jr JB. 2007. *CRC handbook of avian body masses.* CRC Press, Ann Arbor, Michigan, USA.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14(6):** 587–589.

Kapheim KM, Pan H, Li C, Salzberg SL, Puiu D, Magoc T, Robertson HM, Hudson ME, Venkat A, Fischman BJ et al. 2015. Genomic signatures of evolutionary transitions from solitary to group living. *Science* **348(6239):** 1139–1143.

Kapusta A, Suh A. 2017. Evolution of bird genomes—a transposon's-eye view. *Ann N Y Acad Sci* **1389(1):** 164–185.

Kapusta A, Suh A, Feschotte C. 2017. Dynamics of genome size evolution in birds and mammals. *PNAS* **114(8):** E1460–E1469.

Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115(1):** 49–63.

Kimura M, Ohta T. 1969. The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, **61(3):** 763.

Koshikawa S, Miyazaki S, Cornette R, Matsumoto T, Miura T. 2008. Genome size of termites (Insecta, Dictyoptera, Isoptera) and wood roaches (Insecta, Dictyoptera, Cryptocercidae). *Sci*

Nat **95(9):** 859–867.

Lartillot N, Delsuc F. 2012. Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. *Evolution* **66(6):** 1773–1787.

Lartillot N, Poujol R. 2011. A Phylogenetic Model for Investigating Correlated Evolution of Substitution Rates and Continuous Phenotypic Characters. *Mol Biol Evol* **28(1):** 729–744.

Latrille T, Lanore V, Lartillot N. 2021. Inferring long-term effective population size with mutation–selection models. *Mol Biol Evol* **38(10):** 4573-4587.

Lavoie CA, Platt RN, Novick PA, Counterman BA, Ray DA. 2013. Transposable element evolution in Heliconius suggests genome diversity within Lepidoptera. *Mob DNA* **4(1):** 1-10.

Lechner M, Marz M, Ihling C, Sinz A, Stadler PF, Krauss V. 2013. The correlation of genome size and DNA methylation rate in metazoans. *Theory Biosci* **132:** 47-60.

Lefébure T, Morvan C, Malard F, François C, Konecny-Dupré L, Guéguen L, Weiss-Gayet M, Seguin-Orlando A, Ermini L, Sarkissian CD et al. 2017. Less effective selection leads to larger genomes. *Genome Res* **27(6):** 1016–1028.

Leroy T, Rousselle M, Tilak MK, Caizergues AE, Scornavacca C, Recuerda M, Fuchs J, Illera JC, De Swardt DH, Blanco G et al. 2021. Island songbirds as windows into evolution in small populations. *Curr Biol* **31(6):** 1303-1310.

Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49(W1):** W293–W296.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34(18):** 3094–3100.

Li X, Guo B. 2020. Substantially adaptive potential in polyploid cyprinid fishes: evidence from biogeographic, phylogenetic and genomic studies. *Proc Royal Soc B* **287(1920):** 20193008. doi: 10.1098/rspb.2019.3008

Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, Hvidsten TR, Leong JS, Minkley DR, Zimin A et al. 2016. The Atlantic salmon genome provides insights into rediploidization. *Nature* **533(7602):** 200–205.

Lynch M. 2007. *The origins of genome architecture.* Sinauer Associates, Sunderland, Massachusetts, USA.

Lynch M. 2011. Statistical Inference on the Mechanisms of Genome Evolution. *PLoS Genet* **7(6):** e1001389. doi: 10.1371/journal.pgen.1001389

Lynch M, Ali F, Lin T, Wang Y, Ni J, Long H. 2023. The divergence of mutation rates and spectra across the Tree of Life. *EMBO Rep* **24(10)**: e57561. doi: 10.15252/embr.202357561

Lynch M, Bobay L-M, Catania F, Gout J-F, Rho M. 2011. The Repatterning of Eukaryotic Genomes by Random Genetic Drift. *Annu Rev Genomics Hum Genet* **12(1):** 347–366.

Lynch M, Conery JS. 2003. The Origins of Genome Complexity. *Science* **302(5649):** 1401–1404.

Mackintosh A, Laetsch DR, Hayward A, Charlesworth B, Waterfall M, Vila R, Lohse K. 2019. The determinants of genetic diversity in butterflies. *Nat Commun* **10(1):** 3466. doi: 10.1038/s41467-019-11308-4

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol* **38(10):** 4647–4654.

Martelossi J, Nicolini F, Subacchi S, Pasquale D, Ghiselli F, Luchetti A. 2023. Multiple and diversified transposon lineages contribute to early and recent bivalve genome evolution. *BMC Biol* **21(1):** 1-23.

Mérel V, Gibert P, Buch I, Rodriguez Rada V, Estoup A, Gautier M, Fablet M, Boulesteix M, Vieira C. 2021. The Worldwide Invasion of Drosophila suzukii Is Accompanied by a Large

Increase of Transposable Element Load and a Small Number of Putatively Adaptive Insertions. *Mol Biol Evol* **38(10):** 4252–4267.

Mérel V, Tricou T, Burlet N, Haudry AV. 2024 Relaxed purifying selection is associated with an accumulation of transposable elements in flies. bioRxiv doi: 10.1101/2024.01.23.576885

Mikhailova AA, Rinke S, Harrison MC. 2023. Genomic signatures of eusocial evolution in insects. *Curr Opin Insect Sci* **61:** 101136. doi: 10.1016/j.cois.2023.101136

Mills RE, Bennett EA, Iskow RC, Devine SE. 2007. Which transposable elements are active in the human genome? *Trends Genet* **23(4):** 183-191.

Modolo L, Lerat E. 2015. UrQt: an efficient software for the Unsupervised Quality trimming of NGS data. *BMC Bioinform* **16(1):** 1-8.

Mohlhenrich ER, Mueller RL. 2016. Genetic drift and mutational hazard in the evolution of salamander genomic gigantism. *Evolution* **70(12):** 2865–2878.

Mugal CF, Kutschera VE, Botero-Castro F, Wolf JBW, Kaj I. 2020. Polymorphism Data Assist Estimation of the Nonsynonymous over Synonymous Fixation Rate Ratio ω for Closely Related Species. *Mol Biol Evol* **37(1):** 260–279.

Müller R, Kaj I, Mugal CF. 2022. A nearly neutral model of molecular signatures of natural selection after change in population size. *Genome Biol Evol* **14(5):** evac058. doi: 10.1093/gbe/evac058

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* **32(1):** 268–274.

Nikolaev SI, Montoya-Burgos JI, Popadin K, Parand L, Margulies EH, National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program, Antonarakis SE. 2007. Life-history traits drive the evolutionary rates of mammalian coding and noncoding

genomic elements. *PNAS* **104(51):** 20443–20448.

Novick PA, Basta H, Floumanhaft M, McClure MA, Boissinot S. 2009. The Evolutionary Dynamics of Autonomous Non-LTR Retrotransposons in the Lizard Anolis Carolinensis Shows More Similarity to Fish Than Mammals. *Mol Biol Evol* **26(8):** 1811–1822.

Ohta T. 1992. The Nearly Neutral Theory of Molecular Evolution. *Annu Rev Ecol Evol Syst* **23(1):** 263–286.

Olmo E, Capriglione T, Odierna G. 1989. Genome size evolution in vertebrates: Trends and constraints. *Comp. Biochem. Physiol. B* **92(3):** 447–453.

Osmanski AB, Paulat NS, Korstian J, Grimshaw JR, Halsey M, Sullivan KAM, Moreno-Santillán DD, Crookshanks C, Roberts J, Garcia C et al. 2023. Insights into mammalian TE diversity through the curation of 248 genome assemblies. *Science* **380(6643):** eabn1430. doi: 10.1126/science.abn1430

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20(2):** 289–290.

Pasquesi GIM, Adams RH, Card DC, Schield DR, Corbin AB, Perry BW, Reyes-Velasco J, Ruggiero RP, Vandewege MW, Shortt JA et al. 2018. Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. *Nat Commun* **9(1):** 2774. doi: 10.1038/s41467-018-05279-1

Peona V, Kutschera VE, Blom MPK, Irestedt M, Suh A. 2023. Satellite DNA evolution in Corvoidea inferred from short and long reads. *Mol Ecol* **32(6):** 1288–1305.

Peona V, Weissensteiner MH, Suh A. 2018. How complete are “complete” genome assemblies?—An avian perspective. *Mol Ecol Resour* **18(6):** 1188-1195.

Petersen M, Armisén D, Gibbs RA, Hering L, Khila A, Mayer G, Richards S, Niehuis O, Misof B. 2019. Diversity and evolution of the transposable element repertoire in arthropods with

particular reference to insects. *BMC Ecol Evol* **19(1):** 1-15.

Petrov DA. 2001. Evolution of genome size: new approaches to an old problem. *Trends Genet* **17(1):** 23–28.

Pita S, Panzera F, Mora P, Vela J, Cuadrado A, Sánchez A, Palomeque T, Lorite, P. 2017. Comparative repeatome analysis on Triatoma infestans Andean and Non-Andean lineages, main vector of Chagas disease. *PLoS One* **12(7)**: e0181635. doi: 10.1371/journal.pone.0181635

Popadin K, Polishchuk LV, Mamirova L, Knorre D, Gunbin K. 2007. Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *PNAS* **104(33):** 13390–13395.

Ranwez V, Douzery EJP, Cambon C, Chantret N, Delsuc F. 2018. MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. *Mol Biol Evol* **35(10):** 2582–2584.

Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Philos Trans R Soc B Biol Sci* **365(1552):** 2571-2580.

Reinar WB, Tørresen OK, Nederbragt AJ, Matschiner M, Jentoft S, Jakobsen KS. 2023. Teleost genomic repeat landscapes in light of diversification rates and ecology. *Mob DNA* **14(1):** 14.

Rios-Carlos H, Segovia-Ramírez MG, Fujita MK, Rovito SM. 2024. Genomic Gigantism is not Associated with Reduced Selection Efficiency in Neotropical Salamanders. *J Mol Evol*: 1-10.

Roddy AB, Alvarez-Ponce D, Roy SW. 2021. Mammals with Small Populations Do Not Exhibit Larger Genomes. *Mol Biol Evol* **38(9):** 3737–3741.

Romiguier J, Figuet E, Galtier N, Douzery EJP, Boussau B, Dutheil JY, Ranwez V. 2012. Fast

and Robust Characterization of Time-Heterogeneous Sequence Evolutionary Processes Using Substitution Mapping. *PLoS One* **7(3):** e33852. doi: 10.1371/journal.pone.0033852

Romiguier J, Lourenco J, Gayral P, Faivre N, Weinert LA, Ravel S, Ballenghien M, Cahais V, Bernard A, Loire E, Keller L, Galtier N. 2014. Population genomics of eusocial insects: the costs of a vertebrate-like effective population size. *J Evol Biol* **27(3):** 593–603.

Ruggiero RP, Bourgeois Y, Boissinot S. 2017. LINE Insertion Polymorphisms are Abundant but at Low Frequencies across Populations of Anolis carolinensis. *Front genet* **8:** 44. doi: 10.3389/fgene.2017.00044

Ruzzante L, Reijnders MJMF, Waterhouse RM. 2019. Of Genes and Genomes: Mosquito Evolution and Diversity. *Trends Parasitol* **35(1):** 32–51.

Scornavacca C, Belkhir K, Lopez J, Dernat R, Delsuc F, Douzery EJP, Ranwez V. 2019. OrthoMaM v10: Scaling-Up Orthologous Coding Sequence and Exon Alignments with More than One Hundred Mammalian Genomes. *Mol Biol Evol* **36(4):** 861–862.

Sessegolo C, Burlet N, Haudry A. 2016. Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. *Biol Lett* **12(8):** 20160407. doi: 10.1098/rsbl.2016.0407

Sotero-Caio CG, Platt RN, Suh A, Ray DA. 2017. Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biol Evol* **9(1):** 161–177.

Sproul JS, Hotaling S, Heckenhauer J, Powell A, Marshall D, Larracuente AM, Kelley JL, Pauls SU, Frandsen PB. 2023. Analyses of 600+ insect genomes reveal repetitive element dynamics and highlight biodiversity-scale repeat annotation challenges. *Genome Res* **33(10):** 1708–1717.

Tacutu R, Thornton D, Johnson E, Budovsky A, Barardo D, Craig T, Diana E, Lehmann G, Toren D, Wang J, Fraifeld VE, de Magalhães JP. 2018. Human Ageing Genomic Resources: new and updated databases. *Nucleic Acids Res* **46(D1):** D1083–D1090.

Tollis M, Boissinot S. 2013. Lizards and LINEs: Selection and Demography Affect the Fate of L1 Retrotransposons in the Genome of the Green Anole (Anolis carolinensis). *Genome Biol Evol* **5(9):** 1754–1768.

Vinogradov AE. 1995. Nucleotypic Effect in Homeotherms: Body-Mass-Corrected Basal Metabolic Rate of Mammals Is Related to Genome Size. *Evolution* **49(6):** 1249–1259.

Vinogradov AE. 1997. Nucleotypic Effect in Homeotherms: Body-Mass Independent Resting Metabolic Rate of Passerine Birds Is Related to Genome Size. *Evolution* **51(1):** 220–225.

Volff J-N, Bouneau L, Ozouf-Costaz C, Fischer C. 2003. Diversity of retrotransposable elements in compact pufferfish genomes. *Trends Genet* **19(12):** 674–678.

Wang J, Yuan L, Tang J, Liu J, Sun C, Itgen MW, Chen G, Sessions SK, Zhang G et al. 2023. Transposable element and host silencing activity in gigantic genomes. *Front cell dev biol* **11:** 1124374. doi: 10.3389/fcell.2023.1124374

Wang X, Fang X, Yang P, Jiang X, Jiang F, Zhao D, Li B, Cui F, Wei J, Ma C et al. (2014). The locust genome provides insight into swarm formation and long-distance flight. *Nat Commun* **5(1):** 2957. doi: 10.1038/ncomms3957

Wang Y, Obbard DJ. 2023. Experimental estimates of germline mutation rate in eukaryotes: a phylogenetic meta-analysis. *Evol Lett* **7(4)**: 216-226.

Weber CC, Nabholz B, Romiguier J, Ellegren H. 2014. K_r/K_c but not d_N/d_S correlates positively with body mass in birds, raising implications for inferring lineage-specific selection. *Genome Biol* **15:** 1-13. doi: 10.1186/s13059-014-0542-8

Weyna A, Romiguier J. 2021. Relaxation of purifying selection suggests low effective population size in eusocial Hymenoptera and solitary pollinating bees. *Peer Community J* **1:** e2. doi: 10.24072/pcjournal.3

Whitney KD, Boussau B, Baack EJ, Garland Jr T. 2011. Drift and Genome Complexity

Revisited. *PLoS Genet* **7(6):** e1002092. doi: 10.1371/journal.pgen.1002092

Whitney KD, Garland Jr T. 2010. Did Genetic Drift Drive Increases in Genome Complexity? *PLoS Genet* **6(8):** e1001080. doi: 10.1371/journal.pgen.1001080

Woolfit M, Bromham L. 2005. Population size and molecular evolution on islands. *Proc Royal Soc B* **272(1578):** 2277–2282.

Wright NA, Gregory TR, Witt CC. 2014. Metabolic 'engines' of flight drive genome size reduction in birds. *Proc Royal Soc B* **281(1779):** 20132780. doi: 10.1098/rspb.2013.2780

Wu C, Lu J. 2019. Diversification of Transposable Elements in Arthropods and Its Impact on Genome Evolution. *Genes* **10(5):** 338. doi: 10.3390/genes10050338

Yang H, Goubert C, Cotoras DD, Dimitrov D, Graham NR, Cerca J, Gillespie RG. 2024. Consistent accumulation of transposable elements in species of the Hawaiian Tetragnatha spiny-leg adaptive radiation across the archipelago chronosequence. *Evol J Linn Soc* **3(1):** kzae005. doi: 10.1093/evolinnean/kzae005

Yi S, Streelman JT. 2005. Genome size is negatively correlated with effective population size in ray-finned fish. *Trends Genet* **21(12):** 643–646.

Yi SV. 2006. Non-adaptive evolution of genome complexity. *Bioessays* **28(10):** 979–982.

Zhang Q, Edwards SV. 2012. The Evolution of Intron Size in Amniotes: A Role for Powered Flight? *Genome Biol Evol* **4(10):** 1033–1043.

Zhou W, Liang G, Molloy PL, Jones PA. 2020. DNA methylation enables transposable element-driven genome expansion. *PNAS* **117(32):** 19359-19366.