Imperial College London

Gene Loss is an Important Signature of Insect Size Evolution

Georgios Kalogiannis
CID: 02431394

Supervised by:
Prof. Samraat Pawar
Dr. Dimitrios-Georgios Kontopoulos

Word count: 5975

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Research at Imperial College London

Submitted for the MRes in
Computational Methods in Ecology and Evolution

Imperial College London
Silwood Park Campus

August, 2024

Declaration of Authorship

I, Georgios Kalogiannis, declare that I am the sole author of this thesis dissertation. I collated data for the project from previously published literature, as well as the literature collated by Dr. Paul Huxley for the VecTraits Database, which I have cited appropriately. Maria Filippou (MRes Ecology, Evolution, and Conservation, 2024) integrated the dry-to-live mass conversions from Studier and Sevick (1992) on the mass data I collated and shared with her for use in her thesis. I received guidance from my supervisors Prof. Samraat Pawar and Dr. Dimitrios-Georgios Kontopoulos throughout the duration my project.

Acknowledgements

I want to express my sincere gratitude to my supervisors Prof. Samraat Pawar and Dr. Dimitrios-Georgios Kontopoulos for their continued guidance, advice and support throughout the duration of my project. This research would not have been possible for me to complete without their invaluable theoretical and practical expertise.

I am also grateful to Dr. Paul Huxley for his assistance and guidance with collecting data during my project's infancy, and my fellow academic weapons, Loveline, Maria, and Danica, for their day-to-day help and support in the office.

I would also like to thank my CMEE Clique comrades (Temea, Zane & Prannoy), Isobel, and all my other friends at Silwood Park, for their friendship and company throughout my Master's.

Finally, I must thank my parents and sister for their constant encouragement and emotional support. Without you, I would not have been able to achieve this.

Abstract

Body size determines all major life-history traits and ecological interactions, shaping how organisms evolve within their environment. Yet, genomic signatures underlying these dynamics remain poorly understood and neglected in research. In bacteria and complex eukaryotic organ systems, genomic adaptations are known to facilitate increases in size. In this study, I examine whether similar patterns of genomic change linked to size evolution are identifiable across insects. By collating extensive body size and genomic data, gene loss is shown to be a consistent and significant feature related to size across insects. Contrary to my expectations, affected regions are not found within Hox gene clusters that control for embryonic body plan development and regulation. Additionally, there is no consistent trend of non-coding DNA increasing with body mass, despite predictions in literature. Based on these findings, I conclude that gene loss is a primary signature of body mass evolution, though the precise mechanisms driving this process remain a topic of future study. This study provides first insights into this emerging field of research.

Contents

Declaration of Authorship i

Acknowledgements ii

Abstract iii

1 Introduction 1

2 Methods 3

2.1 Data Compilation 3

2.2 Genomic Analysis 4

2.3 Statistical Analysis 5

3 Results 7

3.1 Body mass and genome size weakly scale across insects and some orders 7

3.2 Gene loss is linked to body mass 8

3.3 Body mass affects some Hox genes 9

3.4 Transposable elements scale with body mass in some orders 10

4 Discussion 11

Code and Data Availability 15

References 16

Supplementary Information 23

SI 1: Body mass data references 23

SI 2: Conversion equations from size to mass 25

SI 3: Data distributions 26

SI 4: Expanded Hox gene analysis results 28

1 Introduction

Body size is a crucial morphological trait that plays a fundamental role in the biology of organisms (Roff and Baker, 1994). It determines the rates of metabolic processes, such as oxygen consumption and temperature response, through gas diffusion and surface area to volume ratios (Gillooly et al., 2016), and influences the scale and rate of interactions of an organism within its natural environment (Peters, 1983; Enquist et al., 1999; Jetz et al., 2004). As a consequence, body size influences almost all major life-history traits (Gergely and Tökölyi, 2023), and the environment drives organisms to smaller or larger body sizes depending on the fitness advantages (Smith and Belk, 2018).

The evolution of an organism's body size is closely tied with its ecological function and role within its environment, with distinct effects observable at both intra- and interspecific levels. Within species, a larger body size is often correlated with greater reproductive capacity, improved access to mates, and resource intake advantages (Kemp and Alcock, 2003). In contrast, smaller organisms benefit from reduced detectability, lower maintenance energy costs, and reduced reproduction costs, leading to decreased energy expenditure and reproductive investment (Remmel and Tammaru, 2009; Gergely and Tökölyi, 2023). Between species, the relationship between size and function highlights different adaptive strategies. Smaller species generally have more complex life histories and experience greater ecological costs through increased energy expenditures relative to their body mass and their vulnerability to environmental changes (Bakewell et al., 2020). Larger species are also exposed to these pressures (in stressful conditions), but generally experience lower mortality, increased longevity, and higher success per offspring (Brown et al., 2004; Bielby et al., 2007). In general, the interaction between body size and ecological function highlights the adaptive importance of size evolution, as organisms optimise their physical traits to navigate and thrive within their environmental spaces.

In metazoans, increased gene complexity is essential for encoding the wide array of proteins required for sensory perception, movement, and physiological processes. It has been suggested that with increases in body size, the complexity of these systems will increase due to biomechanical limitations (Savage et al., 2008) and changes in body plan (Vinogradov and Anatskaya, 2021), such as nutrient and gas transfer problems caused by decreased surface area to volume ratios in the circulatory system (Ruppert and Carle, 1983). Alternatively, DeLong et al. (2010) suggest that increases in size will lead to more interactions of organisms with their surrounding environment, requiring greater genetic diversity and an increased number of genes to be used by new metabolic pathways to sustain these interactions. Generally, one might expect that co-occurring wider nutrient use and biomechanical requirements would demand simultaneous genomic adaptation, as they are two distinct selective pressures.

In contrast, genetic complexity can come with a substantial metabolic cost, as the maintenance and duplication of large and complex genomes require significant amounts of energy in the form of ATP (Lane and Martin, 2010). During cell division, the entire genome must be replicated accurately, a process that is metabolically intensive because of the need for numerous enzymes and regulatory proteins. Kozłowski et al. (2003) developed a model to study the relationships between metabolic rate, cell size, and body mass, all of which in turn correlate with genome size. Their model assumes non-coding DNA to be responsible for increases in genome sizes, suggesting that there is a strong negative relationship between metabolic rate and genome size, alluding to the importance of non-coding DNA length in this relationship. Despite the expectation that a larger genome would be more costly to

maintain, they suggest that it induces a lower metabolic rate due to the enlarged cell size required to contain it and the subsequent reduced metabolic rate of the cell. In their review, Gardner et al. (2020) found little evidence for the relationship between genome size and metabolic rate in vertebrates, suggesting that body mass and possibly other life-history strategies overshadow the relationship with genome size. However, they did not test the assumptions of the size of the coding versus the non-coding region from Kozłowski et al. (2003), which could provide a better understanding of the relationship between specific genomic features, metabolic rate, and, indirectly, body size.

These mechanisms of genomic and body size evolution lead to three reasonable expectations regarding the evolution of genomes across other parts of the tree of life, than just vertebrates. Firstly, in invertebrates, the size of the genome is expected to increase proportionally to body mass (Glazier, 2021). This is due to the increased genomic complexity required from the increased metabolic demands and structural capability of larger organisms (Savage et al., 2008; Vinogradov and Anatskaya, 2021). However, findings in existing literature on insects are generally variable between orders (Glazier, 2021), possibly due to the use of outdated study methods, such as genomic weights as a proxy of genome size in studies of butterfly and moth genomes (Gregory and Hebert, 2003; Miller, 2014), which can result in less accurate estimates (Totikov et al., 2021). Second, this increase in genomic complexity should be directly related to increases in protein coding genes, since these encode the functional traits of organisms (Morris et al., 2022). Third, the non-coding portion of DNA should be partially responsible for the metabolic rate of the organism, due to the scaling relationship of cellular and genomic size (Kozłowski et al., 2003).

In this study, I will examine these three expectations, reassessing the trends between genome size and body mass to establish a better understanding of this relationship using modern data. I will provide a first understanding of the unexplored in insects patterns of protein-coding genes and non-coding DNA relative to body mass. I have the following hypotheses:

(i) Larger insects have larger genomes.

(ii) Larger insects have an increased number of coding genes.

(iii) Larger insects have more non-coding DNA in their genome.

2 Methods

I chose to test my hypotheses using insects as a sample group. This is partly because they are underrepresented in existing literature exploring this topic, which predominantly focuses on vertebrates and prokaryotes (Gardner et al., 2020). Additionally, they play an important role in ecosystems (Verma et al., 2023), and extensive genome and mass data is available for them (Mei et al., 2022). Newly available whole-genome sequences provide more precise estimates of genome length, enabling greater accuracy analyses of genome characteristics, essential for addressing evolutionary theories (Totikov et al., 2021). Specifically, the quality of genome assemblies determines the effectiveness of identifying non-coding regions and gene gain or loss (Zavala-Paez et al., 2024). Furthermore, insect mass and genome size vary significantly, across several orders of magnitude (Chown and Gaston, 2010), from the smallest genome of 68 megabases in Chironomid midges to approximately 16.6 gigabases in *Podisma pedestris* (Cornette et al., 2015; Hawlitschek et al., 2023). This variation allows for better estimates and a comprehensive understanding of evolutionary trends, making insects a suitable taxonomic group to assess whether findings are reliable.

To address hypothesis (i) that larger insects have larger genomes, I will study the relationship between body mass and genome size, having collated the largest dataset of published insect masses. To evaluate hypothesis (ii) that larger insects have an increased number of coding genes, I will use two approaches. I will use phylogenetically-controlled estimates of gene loss as a proxy for the number of coding genes. These are a good approximation of coding region reduction and reduction in gene complexity, since gene gain is inherently harder to achieve than gene loss (Ku et al., 2015; Albalat and Cañestro, 2016; Deutekom et al., 2019), so it is highly unlikely that gene loss trends will be outweighed by gene gain. I will also study Hox genes as these may have an effect on body mass, by influencing embryonic body plan development and nutrient allocation that ultimately defines the adult body plan (Lewis et al., 1999; Duverger and Morasso, 2008). To address hypothesis (iii) of whether larger genomes have longer non-coding DNA regions, I will assess whether the proportion of non-coding transposable elements (TE) in the genome correlates with body mass. Recent studies show that some non-coding regions of insect genomes, specifically TEs, are longer and more abundant in larger genomes (Cong et al., 2022). Transposable elements are closely associated with non-coding regions (Hadjiargyrou and Delihas, 2013), and often non-coding themselves (Park et al., 2022). As they are highly correlated with genome size (Sproul et al., 2023), they can be useful for estimating the relationship of non-coding regions relative to body mass.

2.1 Data Compilation

I collected species size and mass data from published sources (Supplementary Information 1), standardised these to uniform units and converted them to body mass (mg) using taxon-specific allometric equations (Supplementary Information 2). Equations vary by measurement proxy: forewing length (Lepidoptera), hindwing length (Odonata), and body length (all other orders). I calculated custom conversion factors from dry to live mass from Studier and Sevick (1992), using the most specific taxonomic level available (species, genus, family, or order), and averaged multiple data points per species to a mean species mass.

I accessed and downloaded whole-genome sequences from the NCBI Genomes database (NCBI

Resource Coordinators, 2015). 658 species from 11 orders had both genome and mass information; however, I included only 4 orders (Coleoptera, Diptera, Hymenoptera, and Lepidoptera) due to insufficient numbers of species in the other orders for a comparative analysis, lack of phylogenetic information, and time constraints.

I downloaded a time-uncalibrated phylogeny from the Open Tree of Life project using the 'rotl' package (Michonneau et al., 2016) in R version 4.1.2 (R Core Team, 2021). I used the 'congruify.phylo' function from the 'geiger package (Eastman et al., 2013; Pennell et al., 2014) to calibrate the uncalibrated phylogeny using a reference time-calibrated phylogeny accessed from 'TimeTree' (Kumar et al., 2022).

Sproul et al. (2023) provided transposable element lengths for the following non-coding families: long and short interspersed nuclear elements (LINE & SINE), long terminal repeats (LTRs), DNA transposons, and tandem repeats. Merging mass & genome data with the species RE information resulted in a sample of 94 species from 9 orders. I summed the proportions of the TE families for a total proportion of TE per species and log10-transformed this.

2.2 Genomic Analysis

For the analysis of coding gene loss, I first aligned genomes using the 'make_lastz_chains' pipeline and then passed these to the 'TOGA' (Tool to infer Orthologs from Genome Alignments) pipeline, both from Kirilenko et al. (2023). TOGA uses a genome annotation of a reference species, along with a whole-genome alignment between the reference genome and a query genome, and infers the orthologous genes present in both, annotating them and classifying them depending on the changes that have occurred (e.g. duplicated, lost). By aligning multiple species within a taxonomic group to the same reference genome, a phylogenetically informed reconstruction of gene loss relative to the reference is constructed and then related to body mass.

I subsetted the larger data set of genomes to those with chromosome-level assemblies, because TOGA does not handle smaller-level assemblies (scaffold, contig) in a reasonable time frame. This resulted in a sample size, including references, of 46 Coleoptera, 36 Diptera, 38 Hymenoptera, and 86 Lepidoptera for TOGA analyses. I used the following reference species to align and annotate genes on query genomes: *Diorhabda carinulata* (Coleoptera), *Drosophila melanogaster* (Diptera), *Apis cerana* (Hymenoptera), and *Helicoverpa armigera* (Lepidoptera). I chose these species because of their high-quality genome annotation and proximity in phylogenetic distance to the other available species.

I removed species from the TOGA analysis if their distance from the reference exceeded the third quartile, as the results were still skewed despite accounting for phylogenetic bias (described in Methods 3.3 below). This resulted in a sample size of query species for TOGA analyses (excluding references) of 36 species of Coleoptera, 29 Diptera, 30 Hymenoptera, and 81 Lepidoptera.

As TOGA functions by aligning existing genes from a reference genome to a query genome, it cannot provide information on gene gain, but it can provide information on the orthology of loci between the two genomes. From the TOGA output, three gene loss metrics were summarised to be used as a proxy for coding genome complexity. Firstly, I / PI, the sum of the number of intact and partially intact genes, which represent probable protein-encoding transcripts in which $\geq 50\%$ of the coding sequence is present and the middle 80% is present without any gene-deactivating mutation present. Second, One-to-None, the number of One-to-None orthology relationships, where TOGA

cannot find a locus in the query genome either because it is lost, it has lost its function, or because there is a gap in the assembly. Third, One-to-None Ratio, the number of One-to-None relationships divided by the sum of all other orthology relationships of One- or Many- loci in the reference to None, One, or Many, depending on whether the loci encode none, one or many functional genes in the query.

To further study the effect of gene loss and gain on body size, I used the HbxFinder pipeline from Zhong and Holland (2011) and Mulhair et al. (2023), which identifies and characterises homeobox, and specifically Hox, genes. The pipeline functions by annotating the Hox gene sequences in a genome from a reference set of Hox gene annotations. It runs best on high-quality assemblies, so I subsetted the dataset into those with N50 scores over 10 megabases, which is in line with the genomes from the 'Darwin Tree of Life Project' that the pipeline was developed with. The N50 score represents the length of the shortest DNA fragment in the assembly (contig or scaffold) at 50% of the total genome length, and is used as a measure of the quality of the genome assembly. Finally, according to the suggestions in Mulhair et al. (2023), I removed Hox genes with <70% identity score and length >150 bp or <210 bp, as in these cases it is likely that the identification of Hox genes has been incorrect (Duverger and Morasso, 2008). This results in a sample size of 50 Coleoptera, 40 Diptera, 41 Hymenoptera and 329 Lepidoptera.

2.3 Statistical Analysis

Following my hypotheses, I expect the log proportion of genomic variables to be dependent on log body mass, because of the multiplicative nature and allometic relationship of body mass scaling with other life-history traits (Kerkhoff and Enquist, 2009). See Supplementary Information 3 for pre- and post-transformation data distributions. As such, the null, linear, and quadratic model which includes possible non-linear responses for this relationship, can be visualised as follows:

- Null: $\log_{10}(C) = \beta_0 + \epsilon$
- Linear: $\log_{10}(C) = \beta_0 + \beta_1 \log_{10}(M) + \epsilon$
- Quadratic: $\log_{10}(C) = \beta_0 + \beta_1 \log_{10}(M) + \beta_2 \left(\log_{10}(M)\right)^2 + \epsilon$

Where C is the genomic variable (either proportion of non-coding DNA (C_{nc}) or the number of Hox genes (C_h)), M is body mass, β_0 is the intercept of the relationship, β_1 the slope of the mass variable, β_2 the slope of the quadratic variable, and ϵ the residual errors.

The models are similar for the TOGA metrics for coding DNA loss (C_c), with the included variable T in all three models and its slope β_T: $\beta_T \log_{10}(T)$. T is a vector of each query species' phylogenetic distance from the reference species within each comparison, to account for TOGA's decreasing ability to locate genes in increasingly distant from the reference species.

I calculated Pagel's λ to estimate the measure of phylogenetic signal in the residuals of the equations, analysed in the form of a linear regression, according to the suggestions of using the 'phylosig' function from the 'phytools' package (Revell, 2024). There was a significant phylogenetic error in the residuals, which violates the assumptions of uncorrelated data in linear models, requiring phylogenetic correction (Revell, 2010; Li and Ives, 2017). As I log10-transformed the results from TOGA and the TE data, and they are both continuous, I used the phylogenetic generalised least squares (PGLS) analysis from the 'nlme' package to analyse the above models (Pinheiro et al., 2024). For each model, I include a phylogenetic correlation structure to account for various modes of trait evolution (see

Table 1 for descriptions of each structure). This results in a model comparison of 15 models for each response variable: null, linear & quadratic, each with 5 error correlation structures. Akaike information criterion (AIC) scores are used for model comparisons, choosing the model with the lowest AIC score as the best fit.

In the Hox gene analyses, I assessed the effect of body mass on both the number of genes per Hox gene family and the total number of Hox genes in a species. I used the same models from the analysis of non-coding DNA to estimate the effects of body mass on the HbxFinder outputs, as they do not suffer from the increased error of TOGA's inability to locate genes in distant genomes. For the comparison across all insect hox genes, I removed Hox genes that were only present in one order, as this would bias the findings within the wider insect comparison. Here, I used a phylogenetic generalised linear mixed model (PGLMM), due to the outputs being positive integer data including gene absence, meaning that the variables could not be log-transformed to be used in a PGLS. Depending on the range of the data, I either used a binomial or Poisson distribution in the model. The response variable was not log-transformed, while body mass remained log-transformed.

Traditional R^2 metrics cannot be estimated for PGLS and PGLMM due to their maximum likelihood estimators and the inclusion of phylogenetically correlated data making variance measurements unclear (Ives, 2018). I used the 'rsquared.gls' function of the 'piecewiseSEM' package to calculate R^2 for the PGLS analyses (Lefcheck, 2016), and the 'R2' function of the 'rr2' package for PGLMM analyses (Ives, 2018).

Table 1: Correlation structures from the 'phytools' package used to account for phylogenetic signal of body mass evolution through various modes of trait evolution (Revell, 2024).

Function	Description	Reference
corBlomberg	Uses Blomberg's K measure of phylogenetic signal to adjust the correlation structure of the phylogeny in a manner relative to the expected evolution of the trait under Brownian motion.	(Blomberg et al., 2003)
corBrownian	Uses a Brownian Motion mode of evolution, which assumes that traits evolve according to a random walk: neutrally at a constant rate along the branches.	(Martins and Hansen, 1997)
corGrafen	Uses Grafen's model of phylogenetic correlation, rescaling the branch lengths of the phylogenetic tree, and allowing variation in the strength of the phylogenetic signal along the phylogeny.	(Grafen and Hamilton, 1989)
corMartins	Uses stabilising selection to model traits under selective pressures, where evolution is constrained toward an optimal value.	(Martins and Hansen, 1997)
corPagel	Uses Pagel's λ to rescale the internal branches of the phylogeny based on the strength of the phylogenetic signal of the trait.	(Pagel, 1999)

3 Results

3.1 Body mass and genome size weakly scale across insects and some orders

Multiple comparisons of log species mass with log genome size showed differing trends between orders of insects. The models fitted to the total insect sample contained 658 species and 11 orders, while the order comparisons contained 55 species in Coleoptera, 41 in Diptera, 7 in Hemiptera, 38 in Hymenoptera, and 338 in Lepidoptera. Other than in Hymenoptera, where the null linear model fit best, PGLS model fitting of the linear model showed the lowest AIC scores and consequently the best fit. On average, mass significantly affected insect size across all insects and within most insect orders. Generally, Blomberg and Brownian correlation structures produced better fits than other correlation structures and were used in the PGLS models of Coleoptera and Diptera, as well as across insects.

Figure 1 shows the results of the model fitting comparisons between PGLS and Linear Regressions. Excluding the case of the Hymenoptera, it is visible from the figure that the PGLS fits produced models with slopes lower than the linear regressions, indicating that including closely related species introduced bias in the samples, due to the non-independent evolution of body and genome size. This is particularly evident in the case of Lepidoptera, where the phylogenetic correction reduced the patterns from significant to non-significant.

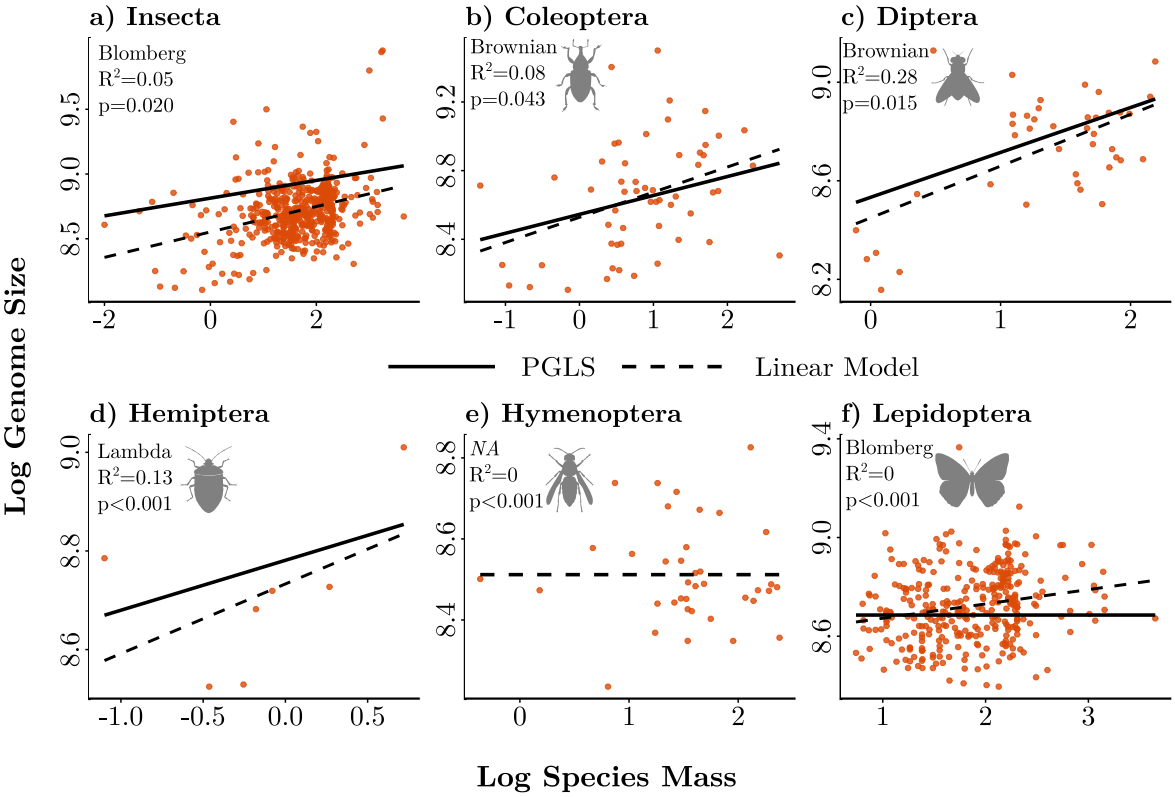

Figure 1: Comparison of the log10 mass of species (mg) to the log10 genome size of their assemblies. PGLS and linear regression analyses were performed in all insects, across 11 orders, (**a**) and individual orders (**b-f**). Linear regression models were performed without phylogenetic correction. Specific PGLS error structures, significance values and pseudo–R^2 values are detailed in the individual panels.

3.2 Gene loss is linked to body mass

For the TOGA analyses, there are generally significant trends of gene loss relative to mass across order and metric (see Figure 2). All orders showed a significant negative trend in at least one gene loss metric: number of One-to-None genes in Coleoptera, Hymenoptera, and Lepidoptera, and One-to-None Ratio in Coleoptera and Diptera. For I/PI, there is a slight positive relationship in the number of intact & partially intact genes in Coleoptera, but the trend is not visible or significant in other orders.

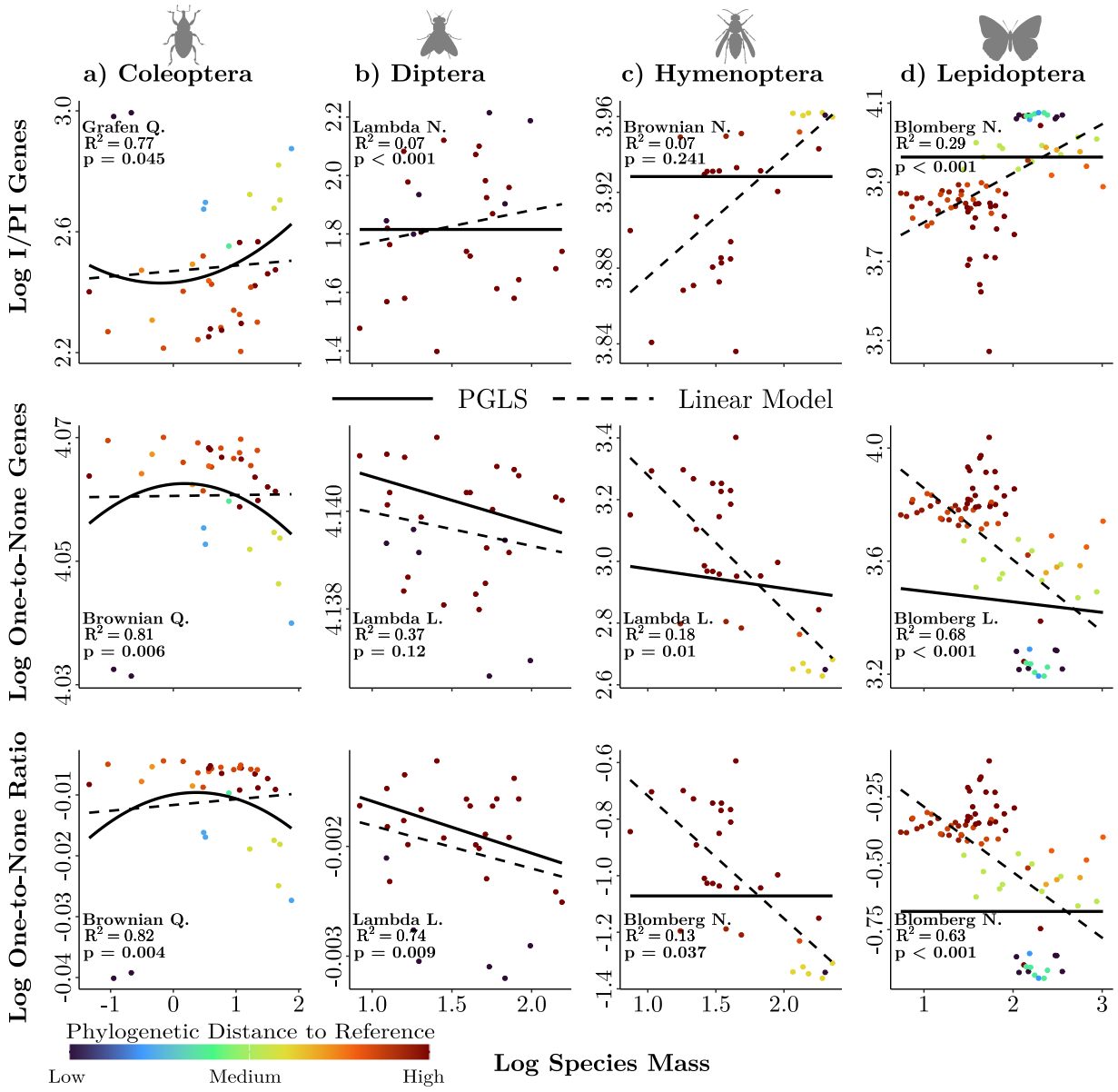

Figure 2: Relationship between the log10 mass of species (mg) and log10 gene loss metrics obtained from TOGA analyses. Columns represent each insect order (**a**–**d**), while rows display each loss metric examined across orders. Phylogenetic distance was included as an independent variable in all models to account for bias in gene identification (see Methods 2.3), and is visualised here in colour. For details on the variables used, refer to Methods 2.2. Figure contains the PGLS error structures, significance values, pseudo–R^2 values, and best model type: null (**N**), linear (**L**) or quadratic. (**Q**)

3.3 Body mass affects some Hox genes

The analyses performed on the number of Hox genes per family and on the total number of Hox genes per species revealed a consistently non-significant relationship with log body mass across and within all insect orders (see Figure 3). However there were two exceptions to this, where in the comparison of all insects the number of *rough* (*Ro*) and *caudal* (*cad*) genes per species was significantly influenced by log body mass. Although both models showed a significant difference (> 2) in their AIC scores compared to the null model ($\Delta\text{AIC}_{Ro} = 17.2$; $\Delta\text{AIC}_{cad} = 44.4$), their R^2 values indicated a poor explanation of the variance in the data. Non-significant findings from all other insect orders can be seen in Supplementary Information 4.

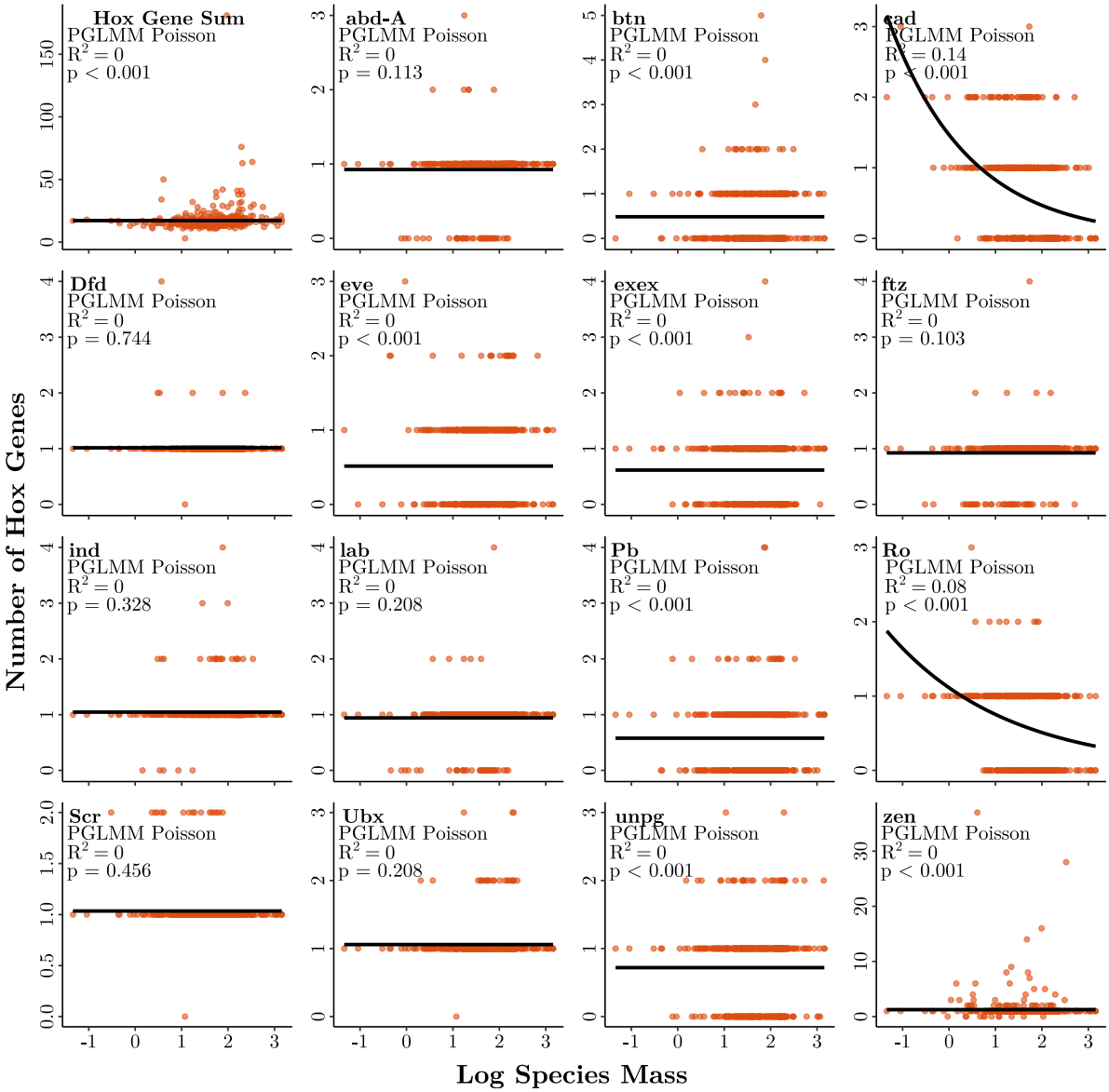

Figure 3: PGLMM predictions for number of Hox genes per species against log10 species mass (mg) estimated from Poisson models across the entire insect sample set. Hox gene name, pseudo–R^2 and p–values included in each panel. See Supplementary Information 4 for the non-significant results in genes within individual orders.

3.4 Transposable elements scale with body mass in some orders

The analyses assessing the relationship of body mass with the proportion of transposable elements in the genome generally showed a lack of influence of body mass on TE length, visible in Figure 4. In all cases, the PGLS analyses with a Lambda error correlation structure fit better than non-phylogenetically controlled linear models and other PGLS' with different error structures; however, many were non-significant. There was only a significant relationship in the Coleoptera and Hemiptera, with the former showing a moderate explanation of the overall variance and the latter minimal (PGLS with Lambda structure; $p < 0.001$, $R^2 = 0.53$ and $p < 0.001$, $R^2 = 0.17$, respectively).

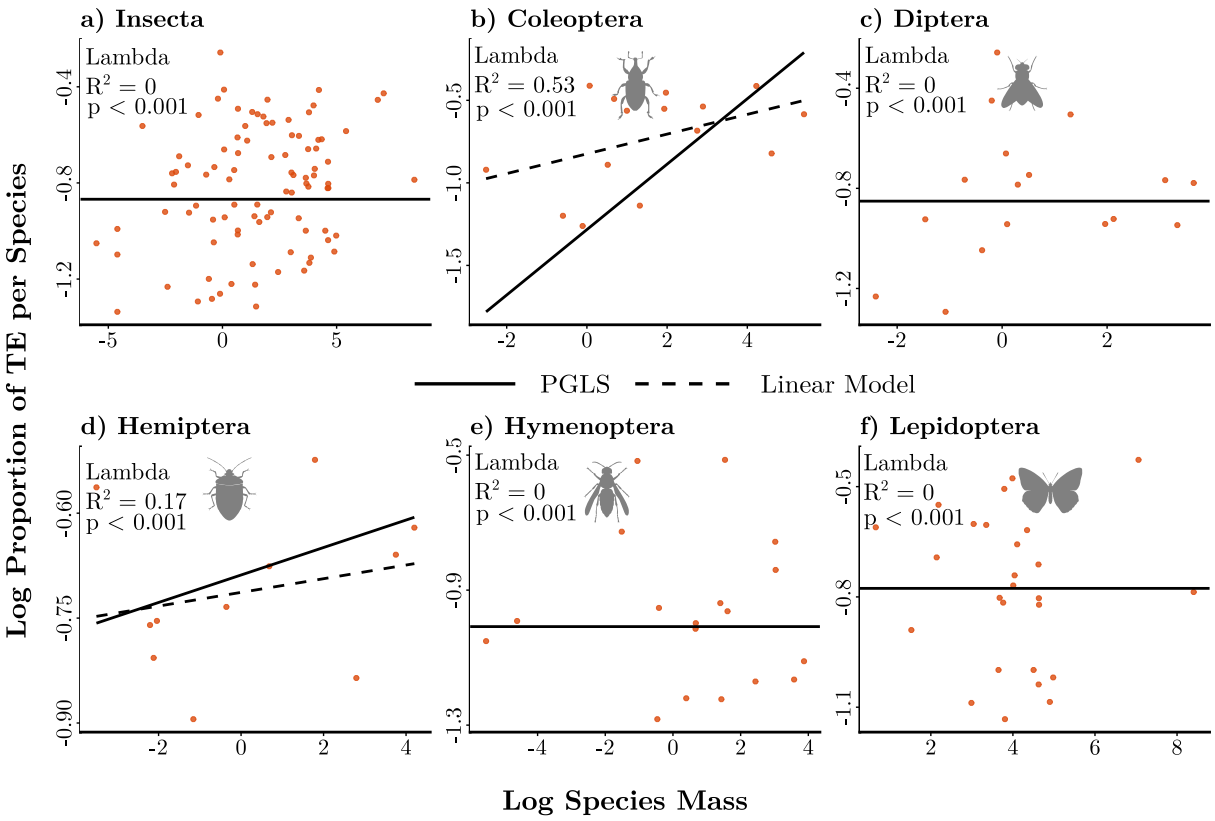

Figure 4: Comparison of log10 mass of species (mg) to the log10 proportion of transposable elements (TE) in their genome. Models were fit across all insects (**a**) and individual orders (**b-f**). Simple linear regression models were fit without phylogenetic correction, and null PGLS visualised as the intercept of the null PGLS model when it was the best fitting model. Data on transposable elements was obtained from Sproul et al. (2023).

4 Discussion

In this study, I sought to address three specific hypotheses regarding the coevolution of insect genomes and their mass. In the first part, I assessed whether larger species exhibit larger genome sizes. Here, I show a significant relationship between insect body mass and genome size when including 11 orders of insect, that has not been previously identified (Glazier, 2021). Within orders, the findings in Coleoptera and Hemiptera contradict the existing literature that suggests that there is no relationship within these orders. The weak positive correlation suggests a link between the DNA content of a cell influencing the size of the organism, possibly indicating that a larger genome is, minimally, necessary to support the increased metabolic and structural demands with increased size. Thus, larger genomes may provide increased genetic capacity for the development and maintenance of more complex body structures and metabolic pathways. However, Glazier (2021) questioned this with their recent research in crustaceans, a group within invertebrates alongside insects, suggesting that propagule size (egg and/or sperm) correlates better than body size with genome size. This is because larger genomes often lead to larger cell sizes and because invertebrate life begins at the single cell, it is more likely to be influenced by genome size. Church et al. (2019) found contradictory results in their study of insects, in which egg size did not correlate with genome size; however, their study includes relatively small and uneven sample sizes across orders, likely skewing their comparisons.

There are two exceptions where larger insects do not have larger genomes: Hymenoptera and Lepidoptera. This indicates a more complex relationship between genome size and mass, with different eco-evolutionary forces at play. In Hymenoptera, social organisation may influence genome evolution. For example, eusocial bees and wasps have smaller genomes than their non-eusocial counterparts (Harrison et al., 2018; Chak et al., 2021). As eusocial insects share reproductive labour, they have a reduced effective population size, making them increasingly vulnerable to invasions of deleterious mutations from mobile genetic elements (Chapman and Bourke, 2001). Their smaller genome size is maintained through increased genetic recombination (Kent and Zayed, 2013), where genetic material is exchanged along or between chromosomes, breaking the coding sequences of mobile genetic elements, and prohibiting the introduction of deleterious mutations (Wilfert et al., 2007; Sirviö et al., 2011). In Lepidoptera, species life history may obscure genome size-mass trends. Specifically, development rate is negatively correlated with the genome size of some moth families, but not in others (Miller, 2014). Additional effects come from feeding preferences, where polyphagous stem borers have much longer genomes than monophagous species (Calatayud et al., 2016). Generally, traits such as development rate, host-plant specificity, and feeding preferences are phylogenetically-correlated traits (Kawahara et al., 2023). This indicates the need to study the evolution of the lepidopteran genome at closer taxonomic levels, to avoid the confounding effects of ecological traits. Such practices should be applied to the study of all insect orders, as these downfalls of comparative genomic analyses exist throughout the insect tree of life (Alfsnes et al., 2017).

I also investigated whether there were patterns of gene loss associated with size. I find a negative relationship between gene loss and body mass in a metric of gene loss, with fewer functional genes found as body mass increases. This result challenges my initial hypothesis that suggests there should be a greater number of genes as body mass increases. I constructed my hypothesis on the basis of bacterial systems and eukaryotic complexity. The number of genes in the bacterial genome is expected to increase proportionally to size, because bacterial genomes are highly specialised in function and the

volume of bacterial cells limits the capacity a genome can occupy (DeLong et al., 2010). Furthermore, the increase in the number of unique adaptations for multicellular organisms is believed to require increased genomic complexity to encode structure and function (Lynch and Conery, 2003; Bingham and Ratcliff, 2024), and the need for more structure will increase with size due to nonlinear scaling of surface area-to-volume ratios (Ruppert and Carle, 1983; Savage et al., 2008). My findings contradict these expectations and suggest a different evolutionary trajectory in which gene loss is more crucial in the evolution of body size.

The observation that gene loss could play an important role in body size evolution challenges traditional views of genome evolution, whereby gene loss has long been assumed to be a neutral process in which redundancy decreases with little evolutionary change (Albalat and Cañestro, 2016). Instead, it aligns with the broader framework of genome streamlining theory, suggesting that gene loss can act as a method to increase genome architecture efficiency by removing redundant or unnecessary functional gene networks (Olson, 1999). In the specific case of gene loss and body size, rather than larger organisms requiring an increased number of genes to support more robust body structures and complex metabolic pathways, genome streamlining suggests that a more efficient genome will facilitate species adaptation to ecological niches that favour certain life histories (Cheatle Jarvela and Wexler, 2023), and in this case a larger body size. For example, in some *Drosophila* fly species, the loss of odorant receptor genes has facilitated their expansion into new ecological niches, with associated losses of behaviour and newly obtained diets (Goldman-Huertas et al., 2015). Although examples like this illustrate genome streamlining along a shorter evolutionary time frame, it is unclear whether this mechanism explains the macroevolutionary findings of this study.

A second possible mechanism of gene loss is that of regressive evolution. Loss of use of a trait puts the genes that encode it under weaker selection and allows mutations that ultimately stop gene function to accumulate (Albalat and Cañestro, 2016). Regressive evolution often occurs in species that have either transitioned into new ecological niches or experienced environmental changes that render previously useful traits obsolete. Examples of regressive evolution include flightless birds and cave-dwelling species (Zess et al., 2022), such as dytiscid water beetles in aquifers losing eye pigment genes due to the lack of light in their environment (Leys et al., 2005). In the context of the correlation of body size and gene loss, regressive evolution suggests that genes encoding ecologically essential functions in smaller species become redundant as they adapt to niches that favour larger size and will be lost over time. Regressive evolution, genome streamlining, or a combination of the two could be the mechanism responsible for the resulting trends of gene loss in this analysis.

Furthermore, my analysis of the number of Hox genes indicated a decrease in the number of caudal (*cad*) and rough (*Ro*) genes alongside increase in mass. These genes are critical in the development of the anal plates and sensory mechanisms of the posterior region of the insect body, and photoreceptors in the eye, respectively (Tomlinson et al., 1988; Moreno and Morata, 1999). Despite the statistical significance of this finding, the lack of biological reasoning for these genes being affected by mass and the low R^2 values indicate a weak correlation that may be induced by technical or statistical errors. This is reinforced by my finding that these genes have been lost in many species, which is known to cause detrimental effects to their development, though studied in individual organisms and not across species (Tomlinson et al., 1988; Moreno and Morata, 1999). In addition, these findings are only significant when the pattern is examined across all insects and not within orders. Thus, these results

likely do not represent true evolutionary trends, but are the result of inaccurate gene loss detection. The lack of consistent findings in the broader Hox gene analysis, considering both within and between insect orders, suggests that body size evolution may be influenced by genomic factors further than traditional developmental pathways. Although Hox genes play a crucial role in the development of the insect body plan (Duverger and Morasso, 2008), my analyses suggest that other genetic elements, possibly related to metabolism, growth regulation, or non-coding regions, could be responsible in driving the relationship of gene loss in the evolution of body size.

My final hypothesis proposed that larger organisms should have more non-coding DNA. I studied this in the form of transposable elements because they are well documented and there is considerable available data on their accumulated proportion in insect genomes (Sproul et al., 2023). My hypothesis was based on the theory by Kozłowski et al. (2003) that non-coding DNA is linked to body mass through the effect that genome size has on cell size. Higher accumulation of non-coding DNA leads to larger genomes and larger cells to support this (Cavalier-Smith, 1978), with lower metabolic rates as an inherent advantage of this and, in turn, increased body mass (Glazier, 2022; DeLong et al., 2010). However, the results from my analysis challenge their conclusions, indicating that across insects there is no clear link between body size and non-coding DNA. The discrepancy between the predictions by Kozłowski et al. (2003) and my findings lies in their link between cell size and body mass. Cellular metabolism of a single cell organism is closely related to its size, while organismal metabolism is more closely related to body size, which is determined by a variety of factors, including the number of cells in an organism, their methods of organisation in tissues, and the specific metabolic rates of those structures (He and Huang, 2006). Taking into account my findings that identified a positive correlation between genome size and body mass, as well as consistent gene loss in larger individuals, the lack of correlation of transposable elements with body size suggests that there are alternative signatures within the genome that have evolved alongside body size. If non-coding DNA is in fact correlated with body mass, it is unlikely to contain regions of non-coding transposable elements, as these seem to evolve independently of body mass when studied across all insects.

In this study, I improve on previous literature with my use of physiological, phylogenetic, and genomic information on insects; however, it suffers from a limited number of constraints that are visible primarily from my chosen method of estimating the size and complexity of the coding genome. A major limitation of TOGA is its ability to detect only gene loss and not gene gain, due to its use of existing genome annotations (Kirilenko et al., 2023). Although gene gain is less common than loss, it is possible especially on a macroevolutionary scale (Ku et al., 2015; Albalat and Cañestro, 2016; Deutekom et al., 2019). TOGA works best in pairwise comparisons, using each species as a reference, which requires annotations for all species to determine gene presence or absence. Whole transcriptomes would also allow for a similar comparative analysis (Orr and Goodisman, 2023). In the absence of such data (Mei et al., 2022), computational methods could have provided approximate estimates of gene gain using stochastic simulations under evolutionary expectations, such as 'CAFE'(De Bie et al., 2006), 'BadiRate' (Librado et al., 2011) software packages, but limited time did not allow for this.

Furthermore, the overlapping mass and genomic data set includes species with different life history traits that could have impacted their rates of gene loss (Kawahara et al., 2023; Alfsnes et al., 2017). My analysis of genome size and body mass could have benefited from the inclusion of life-history traits; however, such data is only vastly available for limited orders of insects (Shirey et al., 2022).

Regardless, the trends identified using phylogenetically-controlled methods are strong and consistent enough that this did not hinder its ability to provide insightful results.

Finally, despite the use of transposable elements enabling the study of non-coding DNA, this methodology may potentially be a limitation of this analysis. Transposable elements are significant components of the non-coding region of genomes and are associated with genomic expansions because of their ability to insert themselves and multiply within the genome (Hadjiargyrou and Delihas, 2013). Although this makes them an appropriate group of non-coding DNA to study in relation to body mass, transposable elements can be actively involved in host adaptation. González et al. (2008) demonstrated that transposable elements contributed to the adaptation of *Drosophila melanogaster* to temperate climates by influencing gene expression and introducing genetic diversity. The correlation between transposable elements and body mass in Coleoptera and Hemiptera may therefore indicate a more dynamic interaction of transposable elements in the evolution of mass in these orders. Although this may indicate misuse of transposable elements in this analysis, it also implies that non-coding DNA is likely more than just unnecessary excess in the genome and is an important genomic factor contributing to species evolution along wider time scales.

To conclude, I identified consistent trends of larger insects having larger genome sizes, supporting the hypothesis that a larger size requires increased genomic content to encode metabolic pathways and structure. Exceptions to this rule suggest that ecological and behavioural traits, such as eusociality and feeding preferences, as well as evolutionary history, may influence genome evolution. The finding of gene loss being correlated with larger size challenges traditional views of genomic evolution and instead suggests more complex evolutionary pathways, involving genome streamlining and regressive evolution. Furthermore, the lack of a clear correlation between transposable elements and body mass indicates that genomic features further to non-coding DNA could be relevant to body size evolution in insects. Despite this study advancing the understanding of comparative genomics of insects, the limitations in detecting gene gain, use of transposable elements as a proxy of non-coding DNA and inclusion of species with diverse life histories, regardless of their evolutionary implications, suggest areas for future research in this field. Overall, the findings presented in this study offer insight into the complex dynamics between genomes and morphology.

Code and Data Availability

The code and data required for the collation of mass data are available on GitHub – Repository 1. The remaining code and data required for repeating the data analysis and visualisation of this study are available on GitHub – Repository 2.

References

Albalat, R. and Cañestro, C. (2016), 'Evolution by gene loss', *Nat. Rev. Genet.* **17**(7), 379–391.

Alfsnes, K., Leinaas, H. P. and Hessen, D. O. (2017), 'Genome size in arthropods; different roles of phylogeny, habitat and life history in insects and crustaceans', *Ecology and Evolution* **7**(15), 5939–5947.

Bakewell, A. T., Davis, K. E., Freckleton, R. P., Isaac, N. J. B. and Mayhew, P. J. (2020), 'Comparing life histories across taxonomic groups in multiple dimensions: How mammal-like are insects?', *The American Naturalist* **195**(1), 70–81.

Bielby, J., Mace, G., Bininda-Emonds, O., Cardillo, M., Gittleman, J., Jones, K., Orme, C. and Purvis, A. (2007), 'The Fast-Slow Continuum in Mammalian Life History: An Empirical Reevaluation.', *The American Naturalist* **169**(6), 748–757.

Bingham, E. P. and Ratcliff, W. C. (2024), 'A nonadaptive explanation for macroevolutionary patterns in the evolution of complex multicellularity', *Proceedings of the National Academy of Sciences* **121**(7), e2319840121.

Blomberg, S. P., Garland JR., T. and Ives, A. R. (2003), 'Testing for phylogenetic signal in comparative data: behavioral traits are more labile', *Evolution* **57**(4), 717–745.

Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. M. and West, G. B. (2004), 'Toward a Metabolic Theory of Ecology', *Ecology* **85**(7), 1771–1789.

Calatayud, P.-A., Petit, C., Burlet, N., Dupas, S., Glaser, N., Capdevielle-Dulac, C., Le Ru, B., Jacquin-Joly, E., Kaiser-Arnauld, L., Harry, M. and Vieira, C. (2016), 'Is genome size of Lepidoptera linked to host plant range?', *Entomologia Experimentalis et Applicata* **159**(3), 354–361.

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', *Journal of Cell Science* **34**(1), 247–278.

Chak, S. T. C., Harris, S. E., Hultgren, K. M., Jeffery, N. W. and Rubenstein, D. R. (2021), 'Eusociality in snapping shrimps is associated with larger genomes and an accumulation of transposable elements', *Proceedings of the National Academy of Sciences* **118**(24), e2025051118.

Chapman, R. E. and Bourke, A. F. G. (2001), 'The influence of sociality on the conservation biology of social insects', *Ecology Letters* **4**(6), 650–662.

Cheatle Jarvela, A. M. and Wexler, J. R. (2023), 'Advances in genome sequencing reveal changes in gene content that contribute to arthropod macroevolution', *Dev. Genes Evol.* **233**(2), 59–76.

Chown, S. L. and Gaston, K. J. (2010), 'Body size variation in insects: a macroecological perspective', *Biological Reviews* **85**(1), 139–169.

Church, S. H., Donoughe, S., de Medeiros, B. A. S. and Extavour, C. G. (2019), 'Insect egg size and shape evolve with ecology but not developmental rate', *Nature* **571**(7763), 58–62.

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

Cornette, R., Gusev, O., Nakahara, Y., Shimura, S., Kikawada, T. and Okuda, T. (2015), 'Chironomid Midges (Diptera, Chironomidae) Show Extremely Small Genome Sizes', *Zoological Science* **32**(3), 248 – 254.

De Bie, T., Cristianini, N., Demuth, J. P. and Hahn, M. W. (2006), 'CAFE: a computational tool for the study of gene family evolution', *Bioinformatics* **22**(10), 1269–1271.

DeLong, J. P., Okie, J. G., Moses, M. E., Sibly, R. M. and Brown, J. H. (2010), 'Shifts in metabolic scaling, production, and efficiency across major evolutionary transitions of life', *Proceedings of the National Academy of Sciences* **107**(29), 12941–12945.

Deutekom, E. S., Vosseberg, J., van Dam, T. J. P. and Snel, B. (2019), 'Measuring the impact of gene prediction on gene loss estimates in Eukaryotes by quantifying falsely inferred absences', *PLOS Computational Biology* **15**(8), 1–15.

Duverger, O. and Morasso, M. I. (2008), 'Role of homeobox genes in the patterning, specification, and differentiation of ectodermal appendages in mammals', *J. Cell. Physiol.* **216**(2), 337–346.

Eastman, J. M., Harmon, L. J. and Tank, D. C. (2013), 'Congruification: support for time scaling large phylogenetic trees', *Methods in Ecology and Evolution* **4**(7), 688–691.

Enquist, B., West, G., Charnov, E. and Brown, J. (1999), 'Allometric Scaling of Production and Life History Variation in Vascular Plants', *Santa Fe Institute, Working Papers* **401**, 907–911.

Gardner, J. D., Laurin, M. and Organ, C. L. (2020), 'The relationship between genome size and metabolic rate in extant vertebrates', *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**(1793), 20190146.

Gergely, R. and Tökölyi, J. (2023), 'Resource availability modulates the effect of body size on reproductive development', *Ecol. Evol.* **13**(1), e9722.

Gillooly, J. F., Gomez, J. P., Mavrodiev, E. V., Rong, Y. and McLamore, E. S. (2016), 'Body mass scaling of passive oxygen diffusion in endotherms and ectotherms', *Proc. Natl. Acad. Sci. U. S. A.* **113**(19), 5340–5345.

Glazier, D. S. (2021), 'Genome Size Covaries More Positively with Propagule Size than Adult Size: New Insights into an Old Problem', *Biology* **10**(4).

Glazier, D. S. (2022), 'How metabolic rate relates to cell size', *Biology (Basel)* **11**(8), 1106.

Goldman-Huertas, B., Mitchell, R. F., Lapoint, R. T., Faucher, C. P., Hildebrand, J. G. and Whiteman, N. K. (2015), 'Evolution of herbivory in Drosophilidae linked to loss of behaviors, antennal responses, odorant receptors, and ancestral diet', *Proc. Natl. Acad. Sci. U. S. A.* **112**(10), 3026–3031.

González, J., Lenkov, K., Lipatov, M., Macpherson, J. M. and Petrov, D. A. (2008), 'High Rate of Recent Transposable Element–Induced Adaptation in Drosophila melanogaster', *PLOS Biology* **6**(10), 1–21.

Grafen, A. and Hamilton, W. D. (1989), 'The phylogenetic regression', *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* **326**(1233), 119–157.

Gregory, T. R. and Hebert, P. D. (2003), 'Genome size variation in lepidopteran insects', *Canadian Journal of Zoology* **81**(8), 1399–1405.

Hadjiargyrou, M. and Delihas, N. (2013), 'The Intertwining of Transposable Elements and Non-Coding RNAs', *International Journal of Molecular Sciences* **14**(7), 13307–13328.

Harrison, M. C., Jongepier, E., Robertson, H. M., Arning, N., Bitard-Feildel, T., Chao, H., Childers, C. P., Dinh, H., Doddapaneni, H., Dugan, S., Gowin, J., Greiner, C., Han, Y., Hu, H., Hughes, D. S. T., Huylmans, A.-K., Kemena, C., Kremer, L. P. M., Lee, S. L., Lopez-Ezquerra, A., Mallet, L., Monroy-Kuhn, J. M., Moser, A., Murali, S. C., Muzny, D. M., Otani, S., Piulachs, M.-D., Poelchau, M., Qu, J., Schaub, F., Wada-Katsumata, A., Worley, K. C., Xie, Q., Ylla, G., Poulsen, M., Gibbs, R. A., Schal, C., Richards, S., Belles, X., Korb, J. and Bornberg-Bauer, E. (2018), 'Hemimetabolous genomes reveal molecular basis of termite eusociality', *Nat. Ecol. Evol.* **2**(3), 557–566.

Hawlitschek, O., Sadílek, D., Dey, L.-S., Buchholz, K., Noori, S., Baez, I. L., Wehrt, T., Brozio, J., Trávníček, P., Seidel, M. and Husemann, M. (2023), 'New estimates of genome size in Orthoptera and their evolutionary implications', *PLOS ONE* **18**(3), 1–20.

He, J.-H. and Huang, Z. (2006), 'A novel model for allometric scaling laws for different organs', *Chaos, Solitons & Fractals* **27**(4), 1108–1114.

Ives, A. R. (2018), 'R^2s for Correlated Data: Phylogenetic Models, LMMs, and GLMMs', *Systematic Biology* p. syy060.

Jetz, W., Carbone, C., Fulford, J. and Brown, J. H. (2004), 'The Scaling of Animal Space Use', *Science* **306**(5694), 266–268.

Kawahara, A. Y., Storer, C., Carvalho, A. P. S., Plotkin, D. M., Condamine, F. L., Braga, M. P., Ellis, E. A., St Laurent, R. A., Li, X., Barve, V., Cai, L., Earl, C., Frandsen, P. B., Owens, H. L., Valencia-Montoya, W. A., Aduse-Poku, K., Toussaint, E. F. A., Dexter, K. M., Doleck, T., Markee, A., Messcher, R., Nguyen, Y.-L., Badon, J. A. T., Benítez, H. A., Braby, M. F., Buenavente, P. A. C., Chan, W.-P., Collins, S. C., Rabideau Childers, R. A., Dankowicz, E., Eastwood, R., Fric, Z. F., Gott, R. J., Hall, J. P. W., Hallwachs, W., Hardy, N. B., Sipe, R. L. H., Heath, A., Hinolan, J. D., Homziak, N. T., Hsu, Y.-F., Inayoshi, Y., Itliong, M. G. A., Janzen, D. H., Kitching, I. J., Kunte, K., Lamas, G., Landis, M. J., Larsen, E. A., Larsen, T. B., Leong, J. V., Lukhtanov, V., Maier, C. A., Martinez, J. I., Martins, D. J., Maruyama, K., Maunsell, S. C., Mega, N. O., Monastyrskii, A., Morais, A. B. B., Müller, C. J., Naive, M. A. K., Nielsen, G., Padrón, P. S., Peggie, D., Romanowski, H. P., Sáfián, S., Saito, M., Schröder, S., Shirey, V., Soltis, D., Soltis, P., Sourakov, A., Talavera, G., Vila, R., Vlasanek, P., Wang, H., Warren, A. D., Willmott, K. R., Yago, M., Jetz, W., Jarzyna, M. A., Breinholt, J. W., Espeland, M., Ries, L., Guralnick, R. P., Pierce, N. E. and Lohman, D. J. (2023), 'A global phylogeny of butterflies reveals their evolutionary history, ancestral hosts and biogeographic origins', *Nat. Ecol. Evol.* **7**(6), 903–913.

Kemp, D. J. and Alcock, J. (2003), 'Lifetime resource utilization, flight physiology, and the evolution of contest competition in territorial insects', *The American Naturalist* **162**(3), 290–301.

Kent, C. F. and Zayed, A. (2013), 'Evolution of recombination and genome structure in eusocial insects', *Communicative & Integrative Biology* **6**(2), e22919.

Kerkhoff, A. J. and Enquist, B. J. (2009), 'Multiplicative by nature: Why logarithmic transformation is necessary in allometry', *Journal of Theoretical Biology* **257**(3), 519–521.

Kirilenko, B. M., Munegowda, C., Osipova, E., Jebb, D., Sharma, V., Blumer, M., Morales, A. E., Ahmed, A.-W., Kontopoulos, D.-G., Hilgers, L., Lindblad-Toh, K., Karlsson, E. K., Consortium, Z. and Hiller, M. (2023), 'Integrating gene annotation with orthology inference at scale', *Science* **380**(6643), eabn3107.

Kozłowski, J., Konarzewski, M. and Gawelczyk, A. T. (2003), 'Cell size as a link between noncoding DNA and metabolic rate scaling', *Proceedings of the National Academy of Sciences* **100**(24), 14080–14085.

Ku, C., Nelson-Sathi, S., Roettger, M., Sousa, F. L., Lockhart, P. J., Bryant, D., Hazkani-Covo, E., McInerney, J. O., Landan, G. and Martin, W. F. (2015), 'Endosymbiotic origin and differential loss of eukaryotic genes', *Nature* **524**(7566), 427–432.

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., Stecher, G. and Hedges, S. B. (2022), 'TimeTree 5: An Expanded Resource for Species Divergence Times', *Molecular Biology and Evolution* **39**(8), msac174.

Lane, N. and Martin, W. (2010), 'The energetics of genome complexity', *Nature* **467**(7318), 929–934.

Lefcheck, J. S. (2016), 'piecewiseSEM: Piecewise structural equation modelling in r for ecology, evolution, and systematics', *Methods in Ecology and Evolution* **7**(5), 573–579.

Lewis, D. L., DeCamillis, M. A., Brunetti, C. R., Halder, G., Kassner, V. A., Selegue, J. E., Higgs, S. and Carroll, S. B. (1999), 'Ectopic gene expression and homeotic transformations in arthropods using recombinant Sindbis viruses', *Curr. Biol.* **9**(22), 1279–1287.

Leys, R., Cooper, S. J. B., Strecker, U. and Wilkens, H. (2005), 'Regressive evolution of an eye pigment gene in independently evolved eyeless subterranean diving beetles', *Biol. Lett.* **1**(4), 496–499.

Li, D. and Ives, A. R. (2017), 'The statistical need to include phylogeny in trait-based analyses of community composition', *Methods Ecol. Evol.* **8**(10), 1192–1199.

Librado, P., Vieira, F. G. and Rozas, J. (2011), 'BadiRate: estimating family turnover rates by likelihood-based methods', *Bioinformatics* **28**(2), 279–281.

Lynch, M. and Conery, J. S. (2003), 'The Origins of Genome Complexity', *Science* **302**(5649), 1401–1404.

Martins, E. P. and Hansen, T. F. (1997), 'Phylogenies and the Comparative Method: A General Approach to Incorporating Phylogenetic Information into the Analysis of Interspecific Data', *The American Naturalist* **149**(4), 646–667.

Mei, Y., Jing, D., Tang, S., Chen, X., Chen, H., Duanmu, H., Cong, Y., Chen, M., Ye, X., Zhou, H., He, K. and Li, F. (2022), 'InsectBase 2.0: a comprehensive gene resource for insects', *Nucleic Acids Res.* **50**(D1), D1040–D1045.

Michonneau, F., Brown, J. W. and Winter, D. J. (2016), 'rotl: an R package to interact with the Open Tree of Life data', *Methods Ecol. Evol.* **7**(12), 1476–1481.

Miller, W. E. (2014), 'Phenotypic Correlates of Genome Size in Lepidoptera', *The Journal of the Lepidopterists' Society* **68**(3), 203 – 210.

Moreno, E. and Morata, G. (1999), '*Caudal* is the Hox gene that specifies the most posterior Drosophile segment', *Nature* **400**(6747), 873–877.

Morris, R., Black, K. A. and Stollar, E. J. (2022), 'Uncovering protein function: from classification to complexes', *Essays Biochem.* **66**(3), 255–285.

Mulhair, P. O., Crowley, L., Boyes, D. H., Harper, A., Lewis, O. T., Darwin Tree of Life Consortium and Holland, P. W. H. (2023), 'Diversity, duplication, and genomic organization of homeobox genes in Lepidoptera', *Genome Res.* **33**(1), 32–44.

NCBI Resource Coordinators (2015), 'Database resources of the National Center for Biotechnology Information', *Nucleic Acids Res* **44**(D1), D7–19.

Olson, M. V. (1999), 'When less is more: gene loss as an engine of evolutionary change', *Am. J. Hum. Genet.* **64**(1), 18–23.

Orr, S. E. and Goodisman, M. A. (2023), 'Social insect transcriptomics and the molecular basis of caste diversity', *Current Opinion in Insect Science* **57**, 101040.

Pagel, M. (1999), 'Inferring the historical patterns of biological evolution', *Nature* **401**(6756), 877–884.

Park, E. G., Ha, H., Lee, D. H., Kim, W. R., Lee, Y. J., Bae, W. H. and Kim, H.-S. (2022), 'Genomic analyses of non-coding RNAs overlapping transposable elements and its implication to human diseases', *Int. J. Mol. Sci.* **23**(16), 8950.

Pennell, M. W., Eastman, J. M., Slater, G. J., Brown, J. W., Uyeda, J. C., Fitzjohn, R. G., Alfaro, M. E. and Harmon, L. J. (2014), 'geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees', *Bioinformatics* **30**, 2216–2218.

Peters, R. H. (1983), *The Ecological Implications of Body Size*, Cambridge Studies in Ecology, Cambridge University Press.

Pinheiro, J., Bates, D. and R Core Team (2024), *nlme: Linear and Nonlinear Mixed Effects Models.* R package version 3.1-166.

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Remmel, T. and Tammaru, T. (2009), 'Size-dependent predation risk in tree-feeding insects with different colouration strategies: a field experiment', *Journal of Animal Ecology* **78**(5), 973–980.

Revell, L. J. (2010), 'Phylogenetic signal and linear regression on species data', *Methods in Ecology and Evolution* **1**(4), 319–329.

Revell, L. J. (2024), 'phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things).', *PeerJ* **12**, e16505.

Roff, D. A. and Baker, J. A. (1994), 'The evolution of life histories: theory and analysis', *Ecology* **75**(1), 261.

Ruppert, E. and Carle, K. J. (1983), 'Morphology of metazoan circulatory systems', *Zoomorphology* **103**(3), 193–208.

Savage, V. M., Deeds, E. J. and Fontana, W. (2008), 'Sizing Up Allometric Scaling Theory', *PLOS Computational Biology* **4**(9), 1–17.

Shirey, V., Larsen, E., Doherty, A., Kim, C. A., Al-Sulaiman, F. T., Hinolan, J. D., Itliong, M. G. A., Naive, M. A. K., Ku, M., Belitz, M., Jeschke, G., Barve, V., Lamas, G., Kawahara, A. Y., Guralnick, R., Pierce, N. E., Lohman, D. J. and Ries, L. (2022), 'LepTraits 1.0 A globally comprehensive dataset of butterfly traits', *Sci. Data* **9**(1), 382.

Sirviö, A., Johnston, J. S., Wenseleers, T. and Pamilo, P. (2011), 'A high recombination rate in eusocial Hymenoptera: evidence from the common wasp Vespula vulgaris', *BMC Genet.* **12**, 95.

Smith, A. N. and Belk, M. C. (2018), 'Does body size affect fitness the same way in males and females? A test of multiple fitness components', *Biological Journal of the Linnean Society* **124**(1), 47–55.

Sproul, J. S., Hotaling, S., Heckenhauer, J., Powell, A., Marshall, D., Larracuente, A. M., Kelley, J. L., Pauls, S. U. and Frandsen, P. B. (2023), 'Analyses of 600+ insect genomes reveal repetitive element dynamics and highlight biodiversity-scale repeat annotation challenges', *Genome Res.* **33**(10), 1708–1717.

Studier, E. H. and Sevick, S. H. (1992), 'Live mass, water content, nitrogen and mineral levels in some insects from south-central lower michigan', *Comparative Biochemistry and Physiology Part A: Physiology* **103**(3), 579–595.

Tomlinson, A., Kimmel, B. E. and Rubin, G. M. (1988), '*rough*, a Drosophila homeobox gene required in photoreceptors R2 and R5 for inductive interactions in the developing eye', *Cell* **55**(5), 771–784.

Totikov, A., Tomarovsky, A., Prokopov, D., Yakupova, A., Bulyonkova, T., Derezanin, L., Rasskazov, D., Wolfsberger, W. W., Koepfli, K.-P., Oleksyk, T. K. and Kliver, S. (2021), 'Chromosome-level genome assemblies expand capabilities of genomics for conservation biology', *Genes (Basel)* **12**(9), 1336.

Verma, R. C., Waseem, M. A., Sharma, N., Bharathi, K., Singh, S., A., A. R., Pandey, S. K. and Singh, B. V. (2023), 'The Role of Insects in Ecosystems, an in-depth Review of Entomological Research', *International Journal of Environment and Climate Change* **13**(10), 4340–4348.

Vinogradov, A. E. and Anatskaya, O. V. (2021), 'Growth of biological complexity from prokaryotes to hominids reflected in the human genome', *Int. J. Mol. Sci.* **22**(21), 11640.

Wilfert, L., Gadau, J. and Schmid-Hempel, P. (2007), 'Variation in genomic recombination rates among animal taxa and the case of social insects', *Heredity (Edinb.)* **98**(4), 189–197.

Zavala-Paez, M., Holliday, J. and Hamilton, J. A. (2024), 'Leveraging whole-genome sequencing to estimate telomere length in plants', *Molecular Ecology Resources* **24**(2), e13899.

Zess, E. K., Dagdas, Y. F., Peers, E., Maqbool, A., Banfield, M. J., Bozkurt, T. O. and Kamoun, S. (2022), 'Regressive evolution of an effector following a host jump in the irish potato famine pathogen lineage', *PLoS Pathog.* **18**(10), e1010918.

Zhong, Y.-f. and Holland, P. W. H. (2011), 'HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology', *Evolution & Development* **13**(6), 567–568.

Supplementary Information

SI 1: Body mass data references

Supplementary table with sources of insect size and mass used in the study. Information includes the corresponding numbers of orders and species included in the source data, as well as whether they were estimated through regressions, the trait measured, and the state of the insects at measurement.

Citation	Orders	Species	Regressions	Size/Mass	Live/Dry
Hu et al. (2022)	1	2390	Yes	Mass	Live
Kühsel et al. (2017)	4	113	No	Mass	Live
Fielding and DeFoliart (2008)	1	32	No	Mass	Live
Brückner et al. (2017)	3	113		Mass	Dry
Kinsella et al. (2020)	1	1054	Yes	Mass	Dry
Horne et al. (2018)	8	118	No	Mass	Dry
Kendall et al. (2019)	2	494		Mass	Dry
Leiva et al. (2019)	8	123		Mass	
Brose et al. (2005)	11	108	Yes	Mass	Live
Waller et al. (2019)	1	1011	No	Size	
Middleton-Welling et al. (2020)	1	542	No	Size	
Shirey et al. (2022)	1	12172	No	Size	
Hagge et al. (2021)	1	1157	No	Size	
Gillespie et al. (2017)	1	588	No	Size	
White et al. (2022)			No	Mass	Live
Ehnes et al. (2011)	18	437	No	Mass	Live
Dillon and Frazier (2013)	16	361	No	Mass	Dry
Smit (2017)	1	1	No	Mass	Live

References

Brose, U., Cushing, L., Berlow, E. L., Jonsson, T., Banasek-Richter, C., Bersier, L.-F., Blanchard, J. L., Brey, T., Carpenter, S. R., Blandenier, M.-F. C., Cohen, J. E., Dawah, H. A., Dell, T., Edwards, F., Harper-Smith, S., Jacob, U., Knapp, R. A., Ledger, M. E., Memmott, J., Mintenbeck, K., Pinnegar, J. K., Rall, B. C., Rayner, T., Ruess, L., Ulrich, W., Warren, P., Williams, R. J., Woodward, G., Yodzis, P. and Martinez, N. D. (2005), 'Body sizes of consumers and their resources', *Ecology* **86**(9), 2545–2545.

Brückner, A., Heethoff, M. and Blüthgen, N. (2017), 'The relationship between epicuticular long-chained hydrocarbons and surface area - volume ratios in insects (Diptera, Hymenoptera, Lepidoptera)', *PLoS One* **12**(4), e0175001.

Dillon, M. E. and Frazier, M. R. (2013), 'Thermodynamics constrains allometric scaling of optimal development time in insects', *PLoS One* **8**(12), e84308.

Ehnes, R. B., Rall, B. C. and Brose, U. (2011), 'Phylogenetic grouping, curvature and metabolic scaling in terrestrial invertebrates', *Ecol. Lett.* **14**(10), 993–1000.

Fielding, D. J. and DeFoliart, L. S. (2008), 'Relationship of metabolic rate to body size in Orthoptera', *J. Orthoptera Res.* **17**(2), 301–306.

Gillespie, M. A. K., Birkemoe, T. and Sverdrup-Thygeson, A. (2017), 'Interactions between body size, abundance, seasonality, and phenology in forest beetles', *Ecol. Evol.* **7**(4), 1091–1100.

Hagge, J., Müller, J., Birkemoe, T., Buse, J., Christensen, R. H. B., Gossner, M. M., Gruppe, A., Heibl, C., Jarzabek-Müller, A., Seibold, S., Siitonen, J., Soutinho, J. G., Sverdrup-Thygeson, A., Thorn, S. and Drag, L. (2021), 'Morphological trait database of saproxylic beetles'.

Horne, C. R., Hirst, A. G. and Atkinson, D. (2018), 'Insect temperature–body size trends common to laboratory, latitudinal and seasonal gradients are not found across altitudes', *Funct. Ecol.* **32**(4), 948–957.

Hu, J., Pentinsaari, M. and Hebert, P. D. N. (2022), 'Measuring mass: variation among 3,161 species of Canadian Coleoptera and the prospects of a mass registry for all insects', *PeerJ* **10**, e12799.

Kendall, L. K., Rader, R., Gagic, V., Cariveau, D. P., Albrecht, M., Baldock, K. C. R., Freitas, B. M., Hall, M., Holzschuh, A., Molina, F. P., Morten, J. M., Pereira, J. S., Portman, Z. M., Roberts, S. P. M., Rodriguez, J., Russo, L., Sutter, L., Vereecken, N. J. and Bartomeus, I. (2019), 'Pollinator size and its consequences: Robust estimates of body size in pollinating insects', *Ecol. Evol.* **9**(4), 1702–1714.

Kinsella, R. S., Thomas, C. D., Crawford, T. J., Hill, J. K., Mayhew, P. J. and Macgregor, C. J. (2020), 'Unlocking the potential of historical abundance datasets to study biomass change in flying insects', *Ecol. Evol.* **10**(15), 8394–8404.

Kühsel, S., Brückner, A., Schmelzle, S., Heethoff, M. and Blüthgen, N. (2017), 'Surface area-volume ratios in insects', *Insect Sci.* **24**(5), 829–841.

Leiva, F. P., Calosi, P. and Verberk, W. C. E. P. (2019), 'Scaling of thermal tolerance with body mass and genome size in ectotherms: a comparison between water- and air-breathers', *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**(1778), 20190035.

Middleton-Welling, J., Dapporto, L., García-Barros, E., Wiemers, M., Nowicki, P., Plazio, E., Bonelli, S., Zaccagno, M., Šašić, M., Liparova, J., Schweiger, O., Harpke, A., Musche, M., Settele, J., Schmucki, R. and Shreeve, T. (2020), 'A new comprehensive trait database of European and Maghreb butterflies, Papilionoidea', *Sci. Data* **7**(1), 351.

Shirey, V., Larsen, E., Doherty, A., Kim, C. A., Al-Sulaiman, F. T., Hinolan, J. D., Itliong, M. G. A., Naive, M. A. K., Ku, M., Belitz, M., Jeschke, G., Barve, V., Lamas, G., Kawahara, A. Y., Guralnick, R., Pierce, N. E., Lohman, D. J. and Ries, L. (2022), 'LepTraits 1.0 A globally comprehensive dataset of butterfly traits', *Sci. Data* **9**(1), 382.

Smit, E. R. (2017), 'Host specificity testing of Diorhabda carinulata (Coleoptera: Chrysomelidae) as a biological control agent of Tamarix spp. (Tamaricaceae)', *University of the Witwatersrand, Johannesburg* [**Dissertation**].

Waller, J. T., Willink, B., Tschol, M. and Svensson, E. I. (2019), 'The odonate phenotypic database, a new open data resource for comparative studies of an old insect order', *Sci. Data* **6**(1), 316.

White, C. R., Alton, L. A., Bywater, C. L., Lombardi, E. J. and Marshall, D. J. (2022), 'Metabolic scaling is the product of life-history optimization', *Science* **377**(6608), 834–839.

SI 2: Conversion equations from size to mass

Supplementary table with conversion equations used to estimate insect mass from size. Conversions varied by measurement proxy: forewing length for Lepidoptera, hindwing length for Odonata, and whole body length for other orders and suborders.

Order & Suborder	Mass Equation	Citation
Blattodea	$0.0494 \times \text{Size}^{2.344}$	Hódar (1996)
Coleoptera	$0.0389 \times \text{Size}^{2.492}$	Sample et al. (1993)
Diptera	$0.0414 \times \text{Size}^{2.213}$	Sample et al. (1993)
Ephemeroptera	$0.0070 \times \text{Size}^{2.880}$	Smock (1980)
Hemiptera		
Heteroptera	$0.0084 \times \text{Size}^{3.075}$	Sample et al. (1993)
Auchenorrhyncha	$0.0594 \times \text{Size}^{2.225}$	Sample et al. (1993)
Sternorrhyncha	$0.0594 \times \text{Size}^{2.225}$	Sample et al. (1993)
Hymenoptera	$0.0138 \times \text{Size}^{2.696}$	Sample et al. (1993)
Lepidoptera	$-2.137 + 2.772 \times \text{Size}$	García-Barros (2015)
Odonata		
Zygoptera	$10^{-0.854} \times \text{Size}^{1.855}$	Aromaa et al. (2019)
Epiprocta	$10^{-0.979} \times \text{Size}^{2.218}$	Aromaa et al. (2019)
Orthoptera	$0.0488 \times \text{Size}^{2.515}$	Rogers et al. (1977)
Thysanoptera	$0.0071 \times \text{Size}^{2.537}$	Hódar (1996)

References

Aromaa, S., Ilvonen, J. J. and Suhonen, J. (2019), 'Body mass and territorial defence strategy affect the territory size of odonate species', *Proc. Biol. Sci.* **286**(1917), 20192398.

García-Barros, E. (2015), 'Multivariate indices as estimates of dry body weight for comparative study of body size in Lepidoptera', *Nota Lepidopterol.* **38**(1), 59–74.

Hódar, J. A. (1996), 'The use of regression equations for estimation of arthropod biomass in ecological studies', *Acta Oecologica-international Journal of Ecology* **17**, 421–433.

Rogers, L. E., Buschbom, R. L. and Watson, C. R. (1977), 'Length-weight relationships of shrub-steppe Invertebrates', *Ann. Entomol. Soc. Am.* **70**(1), 51–53.

Sample, B. E., Cooper, R. J., Greer, R. D. and Whitmore, R. C. (1993), 'Estimation of insect biomass by length and width', *Am. Midl. Nat.* **129**(2), 234.

Smock, L. A. (1980), 'Relationships between body size and biomass of aquatic insects', *Freshw. Biol.* **10**(4), 375–383.

SI 3: Data distributions

In this subsection of the supplementary information, the distributions of data collated in this study can be seen in Figures 1–3, as partial justification for their log-transformation.

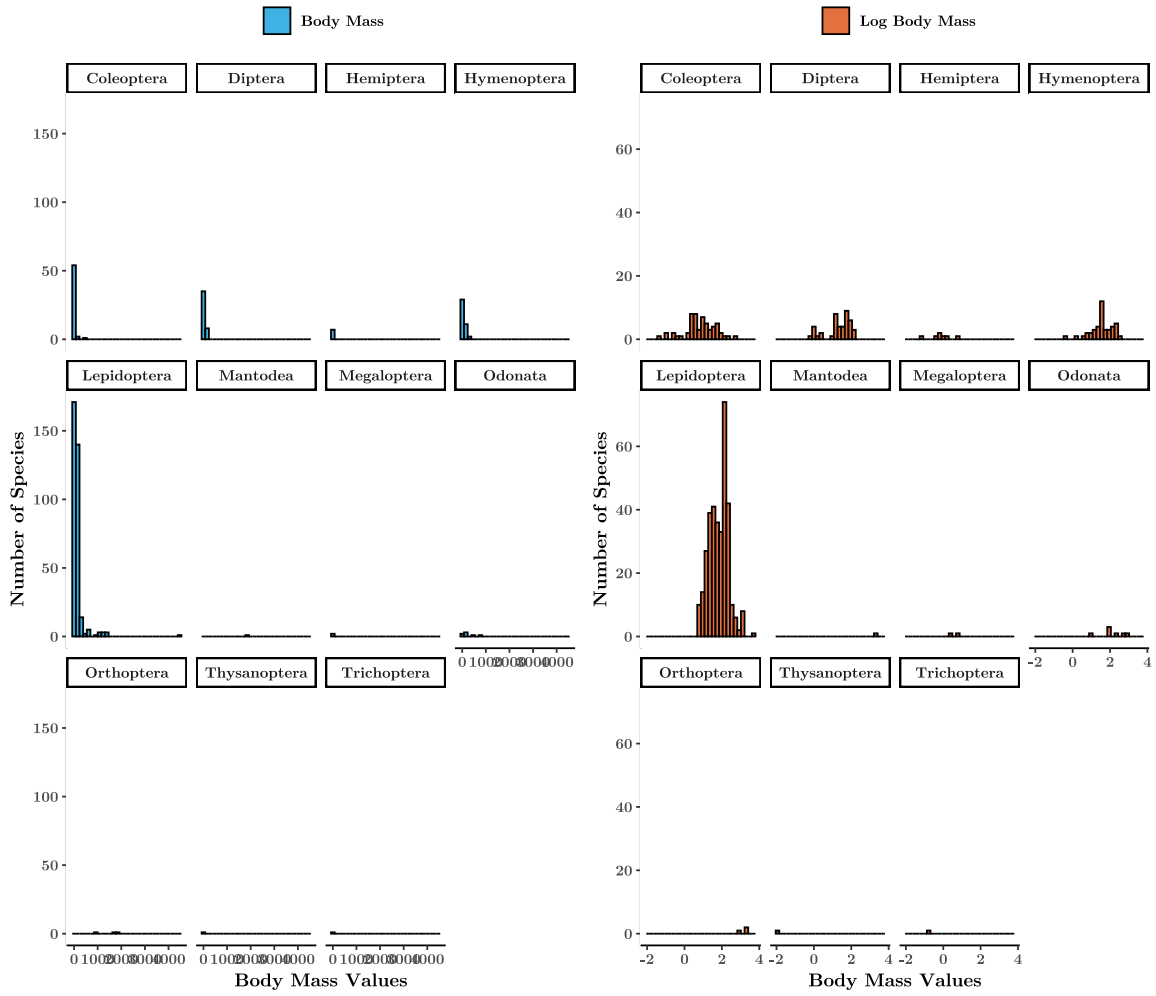

Figure 1: Comparison of body mass distributions in original values (mg) and log10-transformed across orders of insects.

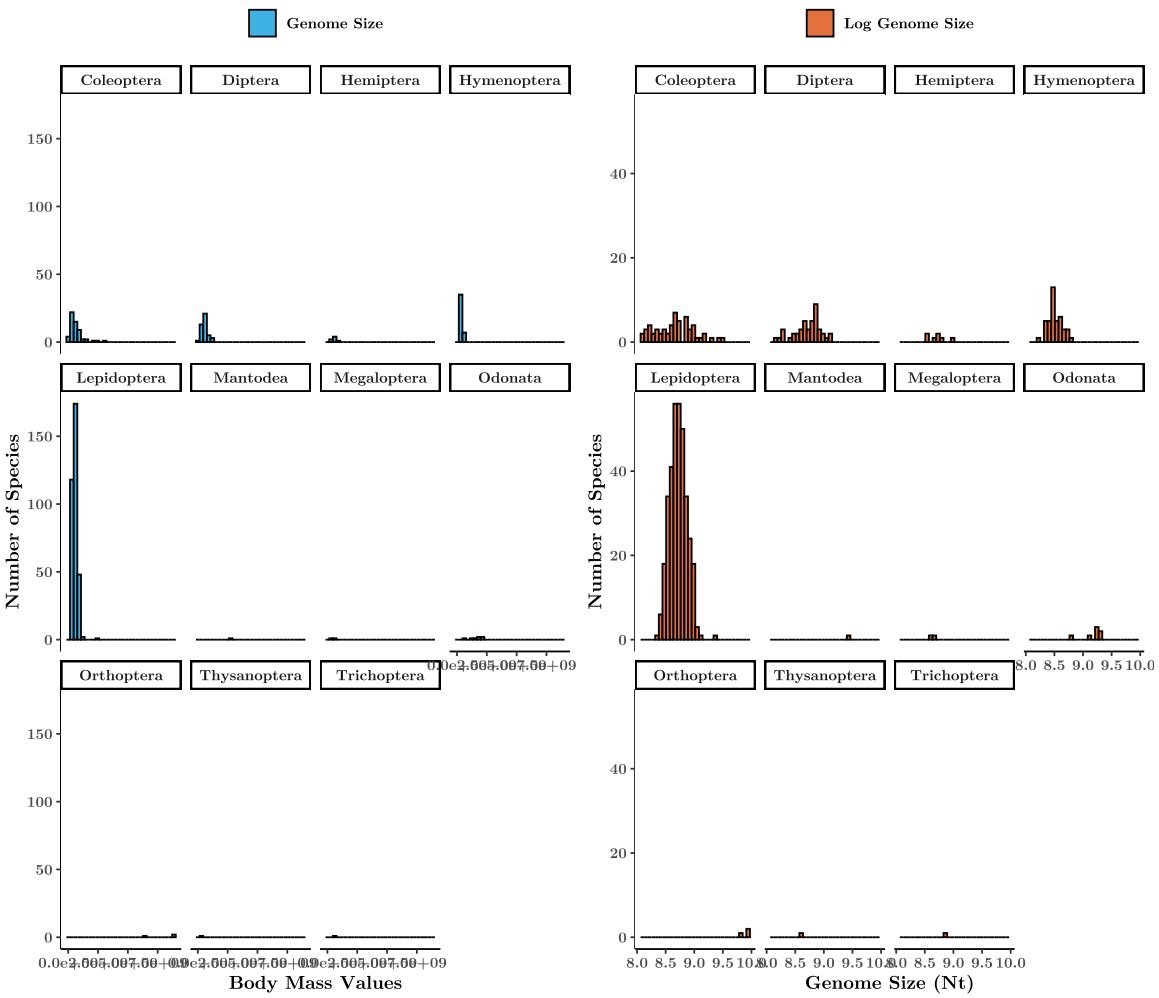

Figure 2: Comparison of genome size distributions in original values (Nt – nucleotides) and log10-transformed across orders of insects.

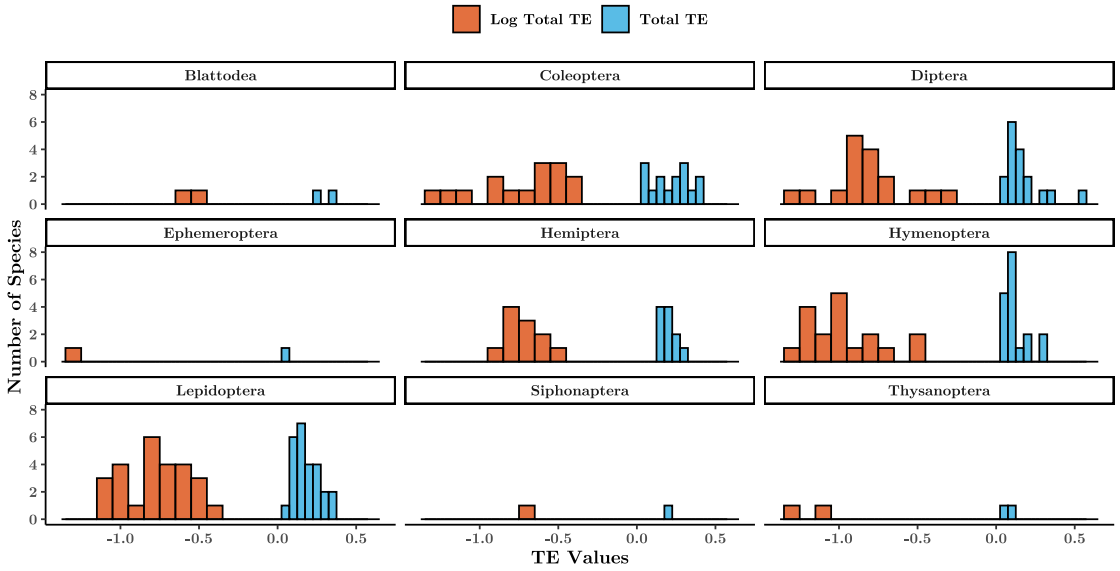

Figure 3: Comparison of transposable element distributions in original values (proportion of the genome – %) and log10-transformed across orders of insects.

SI 4: Expanded Hox gene analysis results

Table of all Hox gene analyses for individual insect orders. All linear and quadratic models were nonsignificant following model comparisons. Information on data distributions (Poisson or Binomial), whether the models were null, linear or quadratic, p-values, and pseudo-R^2 values is included.

	Coleoptera				**Diptera**			
Hox Gene	**Structure**	**Model**	P	R^2	**Structure**	**Model**	P	R^2
Abd–A	Poisson	Null	0.888	0				
Abd–B					Binomial	Null	0.008	0
btn	Poisson	Null	< 0.001	0	Poisson	Null	0.003	0
cad	Poisson	Null	< 0.001	0	Poisson	Null	0.874	0
Dfd	Poisson	Null	0.671	0				
eve	Poisson	Null	0.002	0	Poisson	Null	0.874	0
exex	Poisson	Null	0.003	0	Poisson	Null	0.752	0
ftz	Binomial	Null	0.999	0	Poisson	Null	0.752	0
ind	Poisson	Null	0.888	0	Poisson	Null	0.527	0
lab	Poisson	Null	0.888	0	Binomial	Quadratic	0.155	0.05
Pb	Poisson	Null	0.017	0				
Scr	Poisson	Null	0.258	0				
Ubx	Poisson	Null	0.888	0				
unpg	Poisson	Null	0.888	0	Binomial	Null	0.209	0
zen	Poisson	Null	< 0.001	0	Poisson	Null	< 0.001	
Sum	Poisson	Null	< 0.001	0	Poisson	Null	< 0.001	0

	Hymenoptera				**Lepidoptera**			
Hox Gene	**Structure**	**Model**	P	R^2	**Structure**	**Model**	P	R^2
Abd–A	Poisson	Null	0	0	Poisson	Null	0.956	0
btn	Poisson	Null	< 0.001	0	Poisson	Null	< 0.001	0
cad	Poisson	Null	0.212	0	Poisson	Null	< 0.001	0
eve	Poisson	Null	0.044	0	Poisson	Null	< 0.001	0
exex	Poisson	Null	0.009	0	Poisson	Null	< 0.001	0
ftz	Binomial	Null	0.001	0	Poisson	Null	0.783	0
ind					Poisson	Null	0.408	0
lab					Poisson	Null	0.999	0
Pb	Poisson	Null	0.161	0	Poisson	Null	< 0.001	0
Ro	Poisson	Null	0.876	0	Poisson	Null	< 0.001	0
Scr	Poisson	Null	0.876	0	Poisson	Null	0.869	0
ShxA					Poisson	Null	< 0.001	0
ShxB					Poisson	Quadratic	0.145	0.07
ShxC					Poisson	Linear	0.056	0.01
ShxD					Poisson	Null	< 0.001	0
Ubx					Poisson	Null	0.225	0
unpg	Poisson	Null	0.876	0	Poisson	Null	< 0.001	0
zen	Poisson	Null	0.876	0	Poisson	Null	0.101	0
Sum	Poisson	Null	< 0.001	0	Poisson	Null	< 0.001	0