Genomes from 117 vertebrate species reveal rapidly evolving segmental duplication landscapes

Authors: Alber Aqil[1,2,†], Saiful Islam[3,†], Faraz Hach[4,5], Ibrahim Numanagić[6], Naoki Masuda[1,7,8,9,*], Omer Gokcumen[2,*]

Affiliations:
1. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA.
2. Department of Biological Sciences, State University of New York at Buffalo, Buffalo, NY, USA.
3. Institute for Artificial Intelligence and Data Science, State University of New York at Buffalo, Buffalo, NY, USA.
4. Vancouver Prostate Centre, Vancouver, British Columbia, Canada.
5. Department of Urologic Sciences, University of British Columbia, Vancouver, British Columbia, Canada.
6. Department of Computer Science, University of Victoria, Victoria, BC, Canada.
7. Department of Mathematics, University of Michigan, Ann Arbor, MI, USA
8. Department of Mathematics, State University of New York at Buffalo, Buffalo, NY, USA.
9. Center for Computational Social Science, Kobe University, Japan.

Correspondence:
Omer Gokcumen, gokcumen@gmail.com
Naoki Masuda, naokimas@gmail.com

† *Contributed equally*
* *Correspondence*

Abstract

Segmental duplications are major drivers of evolutionary innovation, yet their dynamics across vertebrates remain poorly understood. Here, we identify segmental duplications from long-read sequenced genomes of 117 vertebrates and the starfish, generating the largest multi-species dataset of its kind. We find that vertebrate genomes show a higher propensity for tandem duplications than for interspersed duplications. However, when focusing only on subtelomeric regions, avian and mammalian genomes show the opposite propensity toward interspersed duplications. We also observe that, across vertebrates, tandem duplications tend to be larger than interspersed duplications. Next, we construct a segmental duplication network for each species, and use network-derived properties to quantify the duplication landscape for that species. Functional enrichment analysis of hyper-duplicated genes reveals a strong enrichment in platypus for pheromone response, driven by the expansion of the vomeronasal pheromone receptor *V1R* gene family. Overall, our results uncover the general properties of vertebrate segmental duplication, demonstrate the rapid evolution of segmental duplication landscapes, and highlight the utility of network-based approaches for studying genome evolution.

Keywords: segmental duplication, vertebrate evolution, segmental duplication networks, platypus, biological networks, tandem duplication, interspersed duplication.

Significance

Gene and regulatory region duplications are a fundamental source of evolutionary raw material. Here we generate segmental duplication calls from 117 vertebrate species. We find that vertebrate genomes have a bias towards tandem duplications relative to interspersed duplications. However, in the subtelomeric regions, birds and mammals exhibit an opposite bias toward interspersed duplications. Our analysis of segmental duplication networks demonstrate that duplication landscapes evolve rapidly, following species-specific rather than phylogenetic patterns. These findings indicate that the genomic architecture underlying segmental duplications is highly dynamic, uniquely shaping each lineage's potential to adapt. Our study provides the most comprehensive view of vertebrate segmental duplications to date and establishes a network-based framework for studying genomic structural evolution.

Introduction

Duplication of genes and regulatory regions has been proposed to play a major role in vertebrate evolution (Ohno 1970). In particular, segmental duplications can create redundant paralogs of functional regions (Carroll 2005), allowing mutations that would have been deleterious in the ancestral copy to be tolerated. Such duplications can therefore circumvent valleys in the fitness landscape (Banse et al. 2024) and increase the probability of reaching higher adaptive peaks (Li and Zhang 2025) by enabling the co-option of duplicates for new functions (Innan and Kondrashov 2010). The mutation rate is elevated in segmental duplicates relative to single-copy regions (Vollger et al. 2023), further raising the likelihood that duplication will acquire new functions. Alternatively, duplication of genes or regulatory elements can drive dramatic changes in expression levels (Aqil et al. 2025), which may itself confer a fitness advantage.

Additionally, genomic regions rich in segmental duplication are prone to accumulating further duplications by increased rates of recombination errors (Kim et al. 2008; Lin and Gokcumen 2019), thereby affecting the potential for adaptation. Finally, as functional analyses extend to non-human species, identifying segmental duplications becomes important, since they can also generate spurious functional associations between genomic loci (Aqil and Gokcumen 2025). For these reasons, identifying segmental duplications and studying the evolution of segmental duplication landscapes across diverse species is imperative.

One way to study the evolution of segmental duplication landscapes is through segmental duplication networks. Previous work has shown that models involving gene duplication followed by divergence between duplicates yield realistic properties in biological networks (Solé et al. 2002; Pastor-Satorras et al. 2003; Vázquez et al. 2003; Lin and Zhang 2010; Lü and Wang 2020; Abdullaev et al. 2021). Although duplication-and-divergence network models were first proposed more than two decades ago (Solé et al. 2002; Pastor-Satorras et al. 2003; Vázquez et al. 2003), they have rarely been tested on large empirical datasets of segmental duplications (Bergman et al. 2006; Vollger et al. 2019; Abdullaev et al. 2021; Abdullaev et al. 2023). Such empirical networks can reveal aspects of the segmental duplication landscape, such as the presence of duplication clusters, that non-network approaches cannot capture.

Here, we call genome-wide (including both genic and non-genic regions) segmental duplications from 117 vertebrate species and the starfish using long-read-sequenced genomes (**Figure 1A–B**), generating the largest multi-species catalog of segmental duplications to date. Using this dataset, we explore the properties of vertebrate segmental duplications, and construct segmental duplication networks to study their evolutionary dynamics (**Figure 1C**).

Overall, our analyses of segmental duplication networks offer new insights into the forces shaping genomic structural variation across vertebrates. By integrating large-scale comparative genomics with network approaches, we address gaps in understanding how the segmental duplication landscape evolves and drives species-specific adaptive potential.

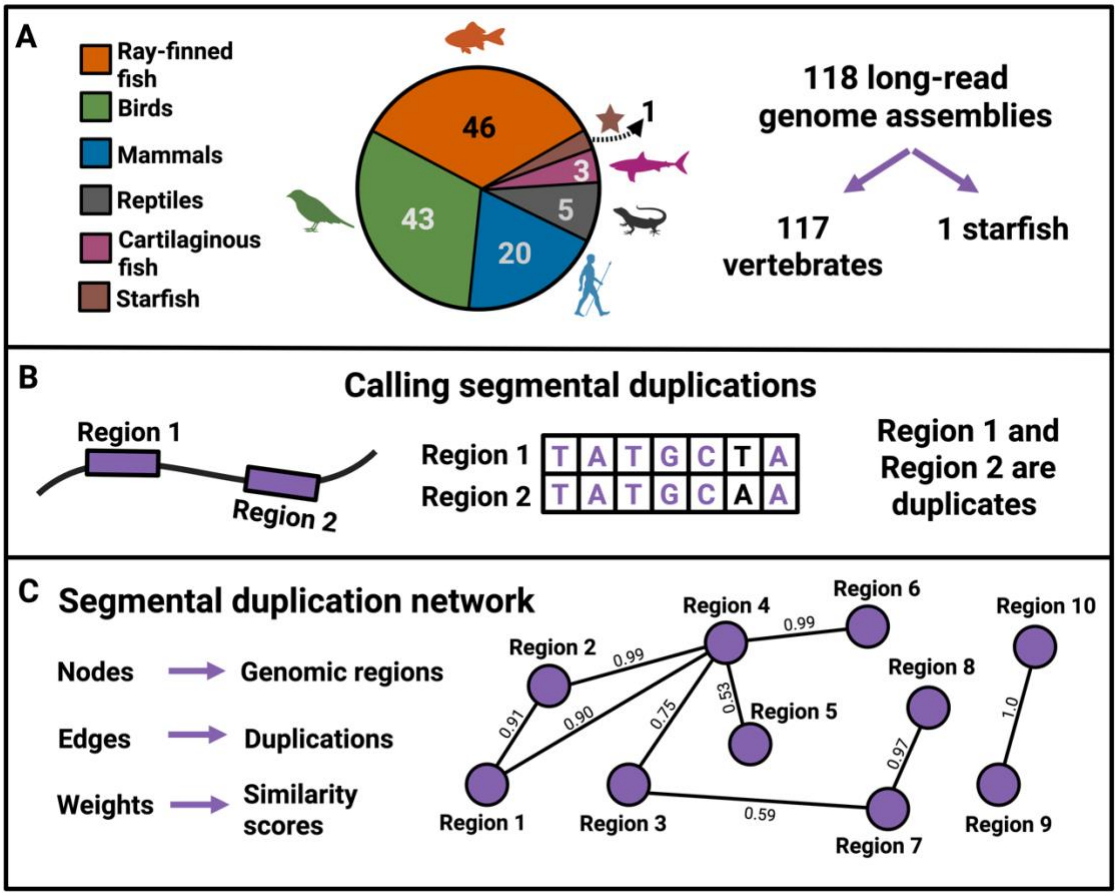

Figure 1. Methodological framework. A. The number of species in each taxonomic class in our data. **B.** A schematic showing segmental duplication calling from genome sequences. The image on the left represents a chromosome with two genomic regions marked as Region 1 and Region 2. The middle image shows that the sequences from the two regions align with a high similarity score. Regions 1 and 2 are, therefore, duplicates. **C.** A schematic segmental duplication network. For each species, we used 13 network properties and the average similarity score across duplications (a non-network property) to quantify the segmental duplication landscape.

Results

Segmental duplication, transposable elements, or ohnologs?

Our approach to detecting segmental duplications relies on identifying regions of sequence similarity with an alignment span of at least 900 bases. Without prior filtering, this approach would capture three categories of duplicated sequence: transposable elements, ohnologs that persist from ancient whole-genome duplications, and segmental duplications.

Transposable elements and short repeats are pervasive in vertebrate genomes (Elliott and Gregory 2015). To remove transposable elements and short repeats, we masked each genome with RepeatMasker using species-specific libraries wherever available. In particular, we used species-specific libraries for all mammals, all birds, all reptiles, 23 out of 46 ray-finned fish species, and 2 out of 3 cartilaginous fish species. For the remaining species, we used class-level repeat libraries. The percentages of each genome masked across taxonomic classes is shown in **Figure 2A.**

In contrast, we do not expect ohnologs to affect large fractions of genomes. In particular, only two species in our dataset (brown trout and sterlet sturgeon; discussed below) have experienced relatively recent whole-genome duplications. Further, duplicates from ancient whole genome duplications (such as two rounds of genome duplications in ancestral vertebrates) are not detectable due to gene loss and sequence divergence after the whole genome duplication (Langham et al. 2004; Brunet et al. 2006; Cheng et al. 2018), i.e. genomic fractionation, as presciently predicted by Haldane (Haldane 1933). Therefore, with the exception of the brown trout and sterlet sturgeon, ohnologs are unlikely to skew our genome-wide analysis of segmental duplications, and no filter was applied to remove them.

After masking genomes for repeats and transposable elements and reasoning that ohnologs are unlikely to bias our results, we called segmental duplications using BISER (Išerić et al. 2022). The percentage of each genome covered by segmental duplications is shown in **Figure 2B.** These percentages are generally higher than those reported elsewhere because BISER relies on a segmental duplication error model that assumes up to 25% sequence dissimilarity rather than the standard 10%. This allows us to discover older segmental duplications. We note that the repeat content and segmental duplication content across genomes are uncorrelated, suggesting that differential masking quality does not introduce a systematic error in segmental duplication calls (**Figure 2C**).

Next, we examined the potential determinants of genome sizes across species. Consistent with previous work (Hughes and Hughes 1995; Gregory et al. 2009; Kapusta et al. 2017), we find that bird genomes are significantly smaller than those of other amniotes (Cohen's $d = 4.5; p < 2.2 \times 10^{-16}$; **Figure 2D**). We find that while repeat content, a rough proxy for transposable elements content (Wang et al. 2025), strongly predicts the size of genomes across species (**Figure 2E**), consistent with previous findings (Elliott and Gregory 2015; Osmanski et al. 2023), segmental duplications do not (**Figure 2F**).

However, the sterlet sturgeon and the brown trout, stand out among ray-finned with unusually large genomes and high apparent coverage by segmental duplications (**Figure 2G**). These signals likely reflect the persistence of ohnologs from whole-genome duplications rather than an excess of true segmental duplications, as discussed below.

The sterlet sturgeon (*Acipenser ruthenus*) is a member of the non-teleost "living fossil" sturgeon family. Its lineage has undergone a distinct third round of whole genome duplication

(Acipenseriformes-specific third-round or As3R WGD) after divergence from the teleost fish and sometime before 200 million years ago (Redmond et al. 2023). Mirroring its morphological stasis, the sterlet sturgeon's molecular evolution (Krieger and Fuerst 2002; Du et al. 2020; Redmond et al. 2023), in terms of both substitutions and rearrangements, is so remarkably slow that the duplication event was initially misdated to 21.3 million years ago (Cheng et al. 2019). Indeed, the sterlet sturgeon retains an exceptionally high proportion (70%) of duplicated genes from the As3R event (Bi et al. 2021). This combination of ancient whole genome duplication, high gene retention, and slow evolutionary change likely drives the apparent association between segmental duplication and the sterlet sturgeon's large genome size.

The brown trout (*Salmo trutta*), a teleost in the salmonid family, is another notable exception. This species, introduced worldwide as a game fish, is recognized as one of the most invasive fish species globally (Lowe et al. 2000). In addition to the teleost-specific whole-genome duplication (Ts3R WGD), the salmonid family underwent a fourth, salmonid-specific genome duplication (Ss4R) between 50 and 80 million years ago (Alexandrou et al. 2013; Macqueen and Johnston 2014; Lien et al. 2016). This extra round of duplication is associated with larger genome sizes in salmonids (Zhu et al. 2022), and as the only salmonid in our dataset, the brown trout has the largest genome among the ray-finned fish we analyzed. Moreover, ~10-15% of salmonid genomes still exhibit tetrasomic inheritance (Allendorf et al. 2015), promoting gene conversion in these regions (Campbell et al. 2019). These gene conversion events help maintain sequence similarity among duplicated regions, preserving the remnants of the Ss4R event, detectable as spurious segmental duplicates. Thus, the combination of a relatively recent whole-genome duplication and ongoing gene conversion in tetrasomically inherited regions likely explains the brown trout's large genome and its high fraction of apparent segmental duplications.

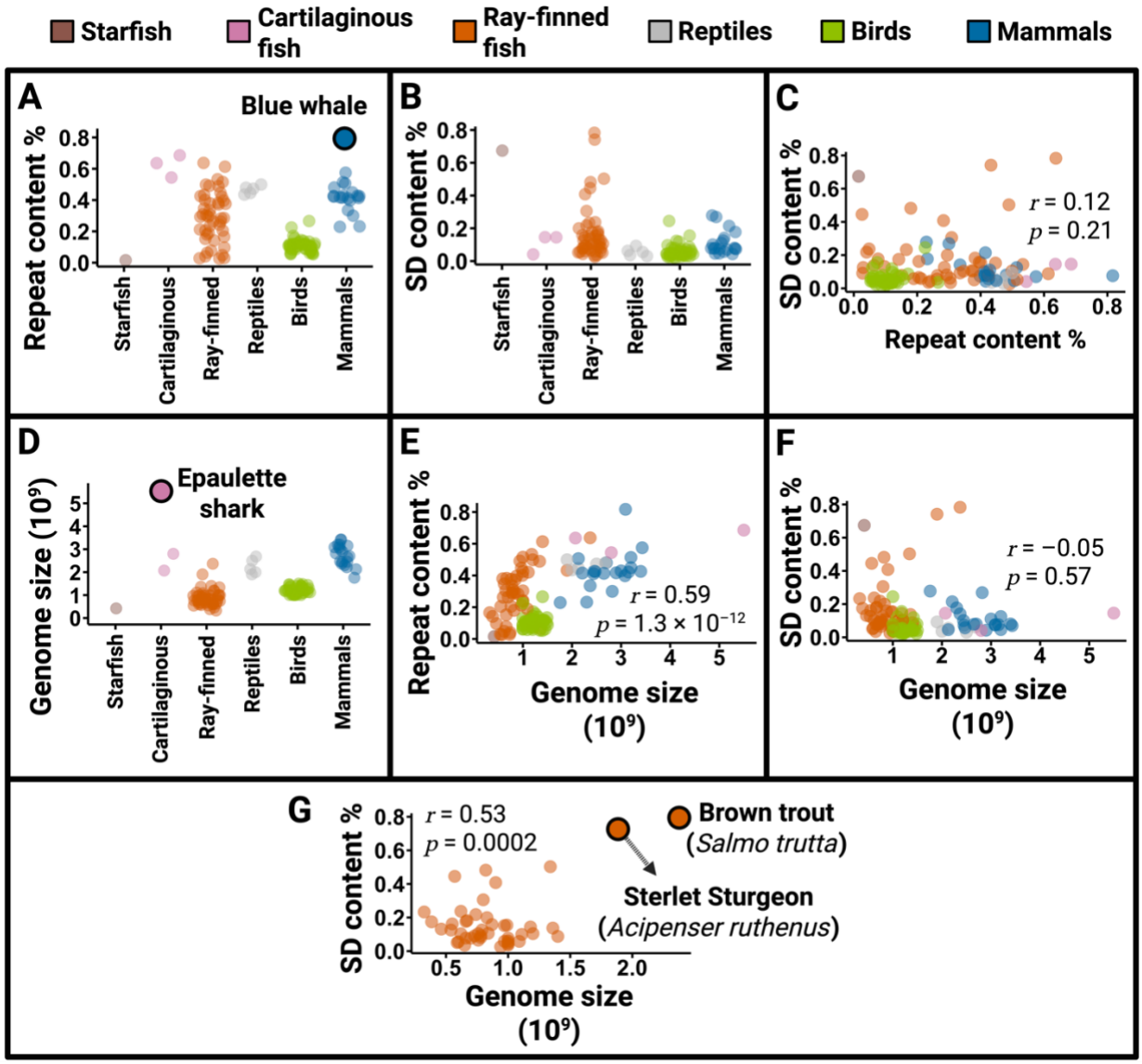

Figure 2. Segmental duplications, repeat content, and ohnologs. A. Distribution of the percentage of the genome covered by repeats (transposable elements and simple repeats) across species. The blue whale (*Balaenoptera musculus*) is an outlier among mammals for its high repeat content. **B.** Percentage of the genome covered by segmental duplications across species. **C.** Repeat content and segmental duplication content are uncorrelated across species. **D.** Distribution of genome size across species. The epaulette shark (*Hemiscyllium ocellatum*) is an outlier among our sample of vertebrates for its large genome size (>5 billion base pairs). **E.** Genome size is correlated with repeat content (presumably driven by transposable elements) across species. **F.** Genome size and segmental duplication content are uncorrelated. **G.** The correlation between genome size and the fraction of the genome covered by duplications among ray-finned fish is driven by two outliers: the sterlet sturgeon and the brown trout (a salmonid). After removing these two species, the correlation coefficient drops to –0.05. The apparently high segmental duplication content of brown trout and sterlet sturgeon actually represents a high ohnolog content.

Duplications close together and duplications far away

We find that vertebrate segmental duplications are not distributed randomly across genomes. In particular, a genomic region on a given chromosome is more likely to be duplicated on the same chromosome than on a different chromosome (**Figure 3A**). This is not to say that every genome contains more intrachromosomal duplications in absolute number. Interchromosomal events are more numerous simply because there are many possible non-homologous chromosome pairs. However, when corrected for this larger search space, all vertebrates have a higher density for intrachromosomal relative to interchromosomal duplications. Moreover, even within the same chromosome, a region is more likely to be duplicated in tandem (within 1MB) than farther away (**Figure 3B, Figure S1**). One consequence of this higher propensity for tandem duplications is the formation of tandem arrays of related genes (Pan and Zhang 2008).

Interestingly, this density bias toward proximal duplications is reversed near chromosome ends (subtelomeric duplications). Subtelomeric regions in mammals, birds, and about half of ray-finned fish species show a density bias towards interchromosomal duplications (**Figure 3C**). This result mirrors previous findings in humans showing that subtelomeric regions are a pastiche of interchromosomal segmental duplications formed by repeated translocations between chromosome ends, facilitated by non-homologous end joining (Linardopoulou et al. 2005).

These spatial patterns suggest that two broad mechanisms shape duplication landscapes. Local errors in recombination and DNA replication create a propensity toward tandem duplications, whereas non-homologous end joining creates a subtelomeric propensity toward interchromosomal duplications.

Finally, we examined how the size (alignment span) and sequence divergence (mismatch score) of duplications differ between tandem (<1 Mb apart) and interspersed (>1 Mb or interchromosomal) duplicates. We find that the median divergence between duplicates is low (**Figure 3D; Figure S2**), and there is no consistent difference in divergence between tandem and interspersed duplications across species (**Figure 3E**).

More interestingly, we observe that while most duplications are small, on the order of 10^3 bp (**Figure 3F; Figure S3**), tandem duplications tend to be larger than interspersed ones (**Figure 3G**). This size difference is consistent with observations in humans (Zhang et al. 2005), mice (Francoeur et al. 2025), and even drosophila (Fiston-Lavier et al. 2007), suggesting that it may be a general feature of animal genomes. This result may also explain the finding in mammals that interchromosomal daughter gene copies evolve faster relative to their parent paralog than do intrachromosomal daughter copies (Moffett et al. 2025). Because interchromosomal duplicates are smaller, they are less likely to include the parent gene's cis-regulatory elements, leaving the parent constrained and the daughter free to diverge; by contrast, larger intrachromosomal duplicates more often retain regulatory sequences, so both copies begin under similar constraint.

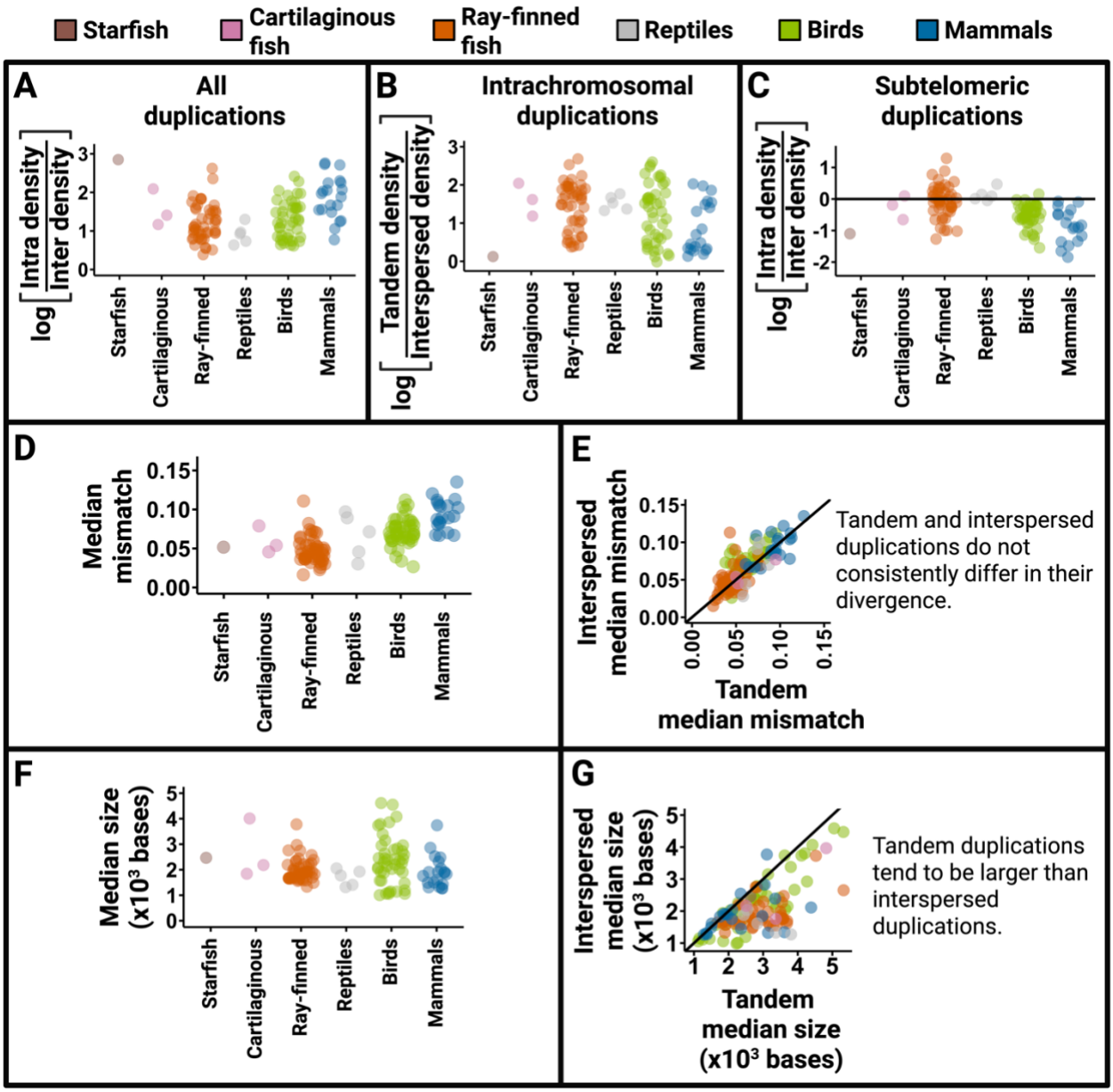

Figure 3. Tandem versus interspersed segmental duplications. A. All vertebrate species show a propensity toward intrachromosomal segmental duplications. The y-axis is the logarithm (base 10) of intrachromosomal to interchromosomal duplication densities. A y-axis value > 0 implies intrachromosomal bias. **B.** Intrachromosomal segmental duplications are biased toward tandem (<1 MB apart) as opposed to interspersed (>1 MB apart) events. The y-axis is the logarithm (base 10) of the ratio of tandem to interspersed duplication density. A y-axis value > 0 implies a bias toward tandem duplications. **C.** Subtelomeric duplications in birds, mammals, and approximately half of ray-finned fish species show a propensity toward interchromosomal events. The y-axis is the logarithm (base 10) of the ratio of intrachromosomal to interchromosomal subtelomeric duplication density. A y-axis value > 0 implies an intrachromosomal bias, and a value < 0 implies an interchromosomal bias. **D.** Distribution of median divergence (mismatch score) between partner duplicates across species in different taxonomic classes. **E.** Comparison of median mismatch scores between tandem and interspersed duplications. The diagonal line represents equal mismatch scores of tandem and interspersed duplications. There is no consistent difference between tandem and interspersed mismatch scores. **F.** Distribution of median segmental duplication size across species in different taxonomic classes. **G.** Comparison of median duplication sizes between tandem and interspersed duplications. The diagonal line represents equal median sizes of tandem and interspersed duplications. Species to the right of this line tend to have larger tandem duplications relative to interspersed duplications. Most species (110/118) have larger tandem duplications relative to interspersed duplications.

Rapidly evolving segmental duplication landscapes

To quantify the segmental duplication landscapes, we constructed a separate segmental duplication network for each species, where nodes represent genomic regions and edges represent duplications. Edge weights indicate similarity scores, representing how recently the duplication emerged. For each species, we calculated 13 network properties and the average similarity score across duplications (a non-network property). We used these 14 measures collectively to quantify the segmental duplication landscape for that species (**see Methods**).

To assess whether our measures of segmental duplication landscapes were influenced by assembly qualities, we calculated correlations between each of the 14 landscape metrics and four assembly quality indicators: coverage width, contig NG50, scaffold NG50, and percentage of the assembly composed of unplaced contigs. Only four of the 14 measures (network density, weighted network density, β, and mean similarity score) were moderately correlated with at least one assembly quality metric ($|r| \geq 0.3$ and raw $p < 0.01$). Thus, most features of the SD landscape are largely independent of assembly quality (**Figures S4-7**), indicating that our cross-species comparisons are robust.

Next, conducted comparative analyses of segmental duplication landscapes across the 118 genomes to investigate how genomic structural evolvability varies across vertebrate species. Specifically, we tested three hypotheses: (1) the "selective constraint hypothesis," which states that segmental duplication patterns are highly conserved across species, showing minimal variation; (2) the "phylogenetic drift hypothesis," which states that the segmental duplication landscape varies in accordance with phylogenetic distance between species; and (3) the "species-specific dynamics hypothesis" stating that the segmental duplication landscape evolves so rapidly that that it does not correlate with phylogeny (**Figure 4A**). We note that these three hypotheses are not mutually exclusive, but it is possible to parse which of these three is most dominant in shaping segmental duplication landscapes.

To test the selective constraint hypothesis (that evolutionary pressure conserves segmental duplication patterns across species), we analyzed all 14 measures across species. Following previous studies (Al-Sallami et al. 2014; Kumar et al. 2018), we considered a measure's variation across species to be low if its coefficient of variation across species was less than 10%. We found that none of the segmental duplication landscape measures exhibited low variation (**Table S1**), suggesting that segmental duplication landscapes are not conserved in vertebrates. In fact, six out of the 14 measures, including the weighted density and mean node strength, exhibit coefficients of variation exceeding 100%. Based on these findings, we reject the selective constraint hypothesis and shift focus to the two alternative hypotheses (i.e., phylogenetic drift and species-specific dynamics) to explain the observed variation in segmental duplication landscapes.

To differentiate between the phylogenetic drift hypothesis and the species-specific dynamics hypothesis, we calculated pairwise distances for each of the 14 segmental duplication measures across species pairs. We correlated these with phylogenetic distances obtained from TimeTree (**see Methods**) (Kumar et al. 2022). If a species was not available in TimeTree, we used a closely related substitute (**Table S2**). If most of these 14 measures correlate with phylogeny, we favor the phylogenetic drift hypothesis; otherwise, we favor the species-specific dynamics hypothesis.

We found that none of the 14 measures of segmental duplication landscape correlates with phylogeny across vertebrates ($|r| > 0.3$; See **Methods** for justification of the threshold; **Table S3**). Instead, the evolution of segmental duplication landscapes largely follows species-specific

dynamics. This lack of correlation led us to reject the phylogenetic drift hypothesis in favor of the species-specific dynamics hypothesis. We demonstrate this reasoning using the average clustering coefficient, one of the 14 measures, in **Figure 4B**. We found that even very closely related species pairs (time since divergence < 50 million years) exhibit no greater similarities in segmental duplication trends than more distantly related species pairs (Cohen's $|d|$ < 0.2 across measures; **Table S4**). These findings suggest that the evolution of segmental duplication patterns, driven by species-specific dynamics, occurs at such a rapid pace that measures of duplication landscapes may differ significantly even between closely related species (for example, **Figure S8**).

We note that one previous study (Abdullaev et al. 2021) reported that segmental duplication landscapes follow the phylogenetic drift hypothesis instead of the species-specific hypothesis supported by our results. However, their conclusion was based on only nine species and one network property (component size distribution). With our large dataset of 118 species, we found that the component size distribution also follows species-specific dynamics, consistent with the results based on our 14 measures **(see Supplementary Text S1)**. This result suggests that the results of (Abdullaev et al. 2021) may be limited in generalizability due to restricted species sampling and focus on a single network property.

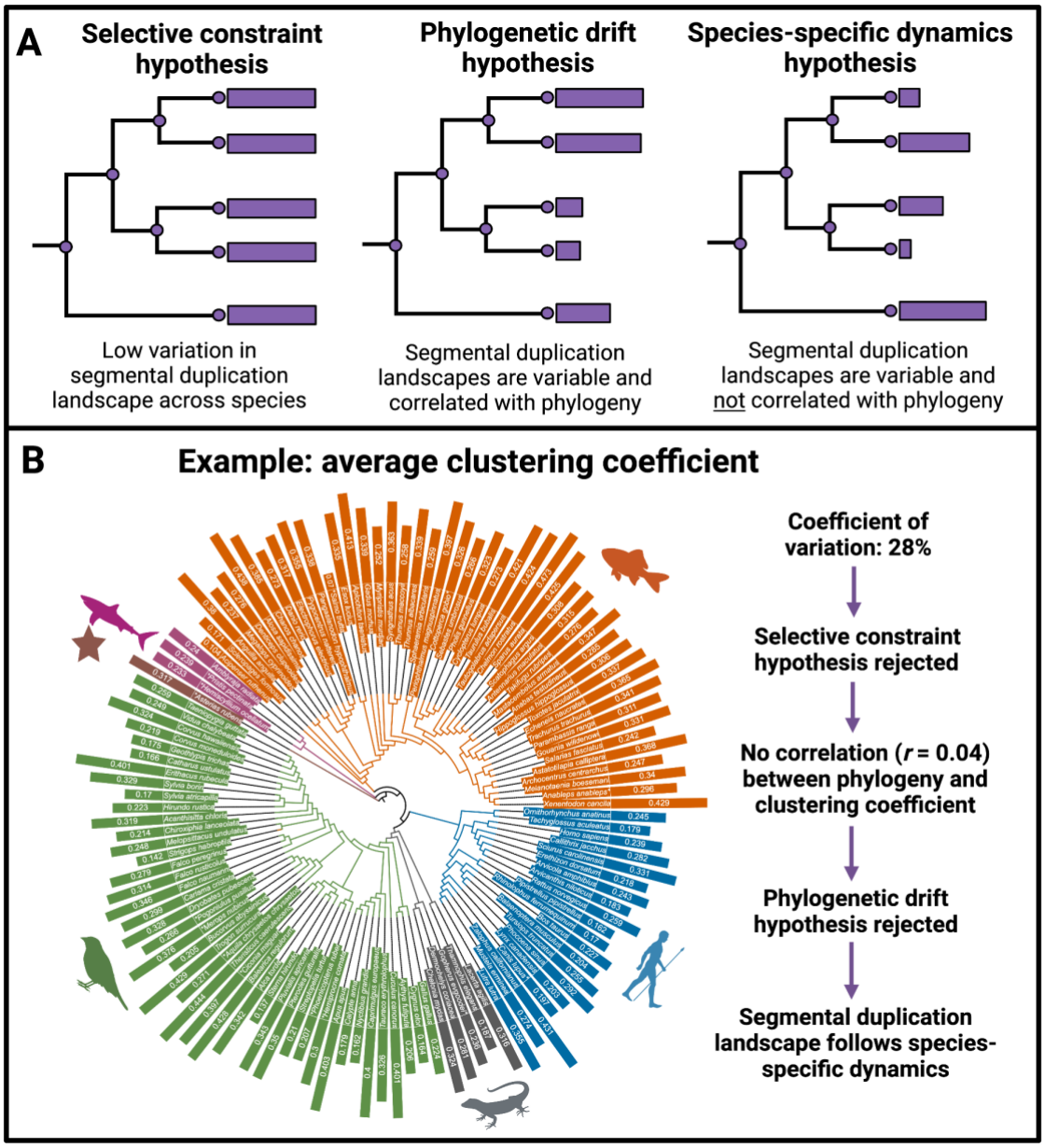

Figure 4. Hypotheses for the evolution of the segmental duplication landscape. A. Expectation for variation in the segmental duplication landscape under each hypothesis. The leaves in the trees in each schematic represent species. The length of the purple bar represents the value of a hypothetical measure of the segmental duplication for the corresponding species. The figure shows the expected variation in this measure under three hypotheses. **B.** A phylogenetic tree of 118 species annotated with the average clustering coefficient values. The height of the bar on each species indicates the value of the average clustering coefficient. Species' names appended with an asterisk mean that a close substitute for the species was used to generate the tree (see **Table S2**). The variation in the average clustering coefficient across species matches the expectation under the species-specific dynamics hypothesis.

Nevertheless, an interesting pattern stands out within mammals, but not across all species, in terms of the average similarity scores (**Figures 5A-B**). In particular, therian mammals (mammals with live birth) exhibit much higher average similarity scores than monotremes (platypus and echidna, the egg-laying mammals) (Hedge's $g = 6.45, p = 0.01$; **Figure 5C**), leading to a strong correlation between phylogeny and average similarity scores within

mammals $r = 0.85; p = 2.26 \times 10^{-53}$; **see Methods**). Because higher similarity scores generally indicate more recent duplications, this result suggests that monotreme genomes may contain older duplications and fewer recent duplications than therian genomes. However, we can break down the dissimilarity score (1 – similarity score) into two parts: mismatch error, which reflects the number of substitutions between duplicates, and gap error, caused by insertions or deletions. Our results show that the gap score differs significantly between monotremes and therian mammals (Hedges $g = 6.45$; Mann-Whitney U test, $p = 0.01$), while average mismatch scores are similar across both groups (Mann-Whitney U test, $p = 0.38$) **(Figures 5D-E)**. If monotreme duplications were indeed older, we would expect both their mismatch and gap scores to be higher. In contrast, the fact that only gap scores are elevated in monotremes suggests their genomes are more prone to insertions and deletions rather than simply containing older duplications. Two hypotheses could be tested to explain the higher gap score in monotremes: 1) larger density of processed pseudogenes in monotremes; and 2) a higher rate of indels in monotremes. Parsing this out would be an interesting topic for future research.

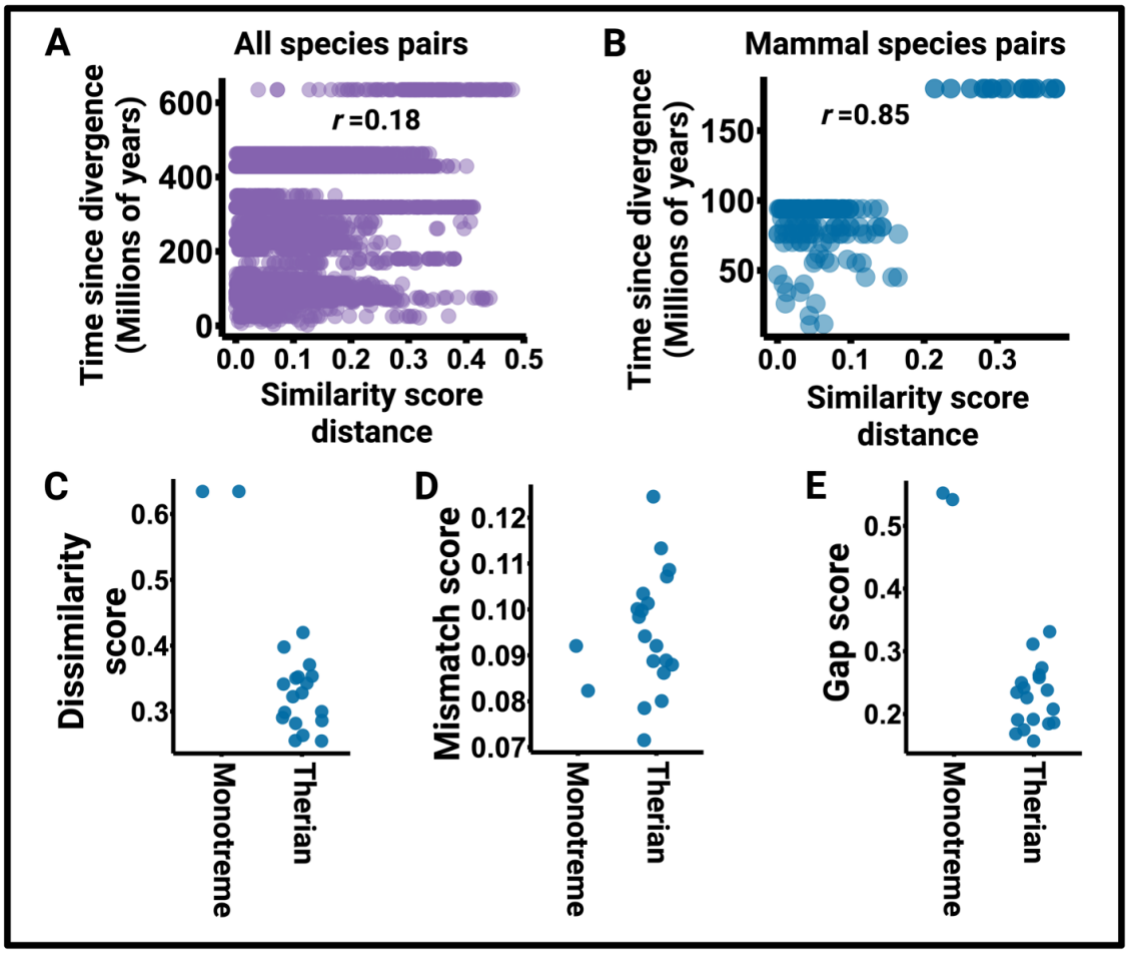

Figure 5. Insights from the average similarity scores for segmental duplications across species. A. Phylogenetic distances are not correlated with the average similarity score distances across pairs of species. **B.** Phylogenetic distances are strongly correlated with average similarity score distances across pairs of mammalian species. Monotreme-therian species pairs drive this correlation. In **A** and **B**, each point represents a species pair. **C-E.** Monotremes (echidna and platypus) have higher average dissimilarity scores (1 – similarity score) than therian mammals. This difference in average dissimilarity scores between therian mammals and monotremes is not driven by differences between mismatch scores but by differences in gap scores. In **C-E**, each point represents a mammalian species.

Functional enrichments

We conducted Gene Ontology (GO) enrichment analyses using g:Profiler for 11 species represented in both our dataset and g:Profiler annotations. These 11 species included four mammals (human, platypus, dolphin, and greater horseshoe bat), four ray-finned fishes (eastern happy, brown trout, lumpfish, and climbing perch), two birds (kakapo and golden eagle), and one reptile (Good's thornscrub tortoise). For each species, we identified the top 100 genes with the largest number of overlapping segmental duplications and tested these sets for enrichment of biological processes.

We did not detect any biological processes enriched across multiple species, suggesting that the types of genes amplified by segmental duplication are largely lineage-specific. Nonetheless, several species showed lineage-specific enrichments, including response to pheromones (p-value adjusted for multiple hypothesis testing, $p_{\mathrm{adj}} = 1.1 \times 10^{-11}$) and sensory perception of chemical stimulus ($p_{\mathrm{adj}} = 2.7 \times 10^{-8}$) in platypus, regulation of DNA-templated transcription ($p_{\mathrm{adj}} = 6.5 \times 10^{-11}$) in humans, synaptic signaling ($p_{\mathrm{adj}} = 0.0011$) in greater horseshoe bat, protein metabolic process ($p_{\mathrm{adj}} = 0.039$) in dolphin, multicellular organismal process ($p_{\mathrm{adj}} = 0.049$) in brown trout, and regulation of lymphangiogenesis ($p_{\mathrm{adj}} = 0.049$) in lumpfish.

The case of the platypus is particularly striking: the enrichment is driven by the expansion of the vomeronasal pheromone receptor *V1R* gene family, consistent with previous findings (Grus et al. 2007; Zhou et al. 2021). Similarly, the functional enrichment in the greater horseshoe bat is driven by duplication of neural function-related genes, while in humans the functional enrichment of transcription regulation is driven by the expansion of zinc finger protein genes (**Figure 6**).

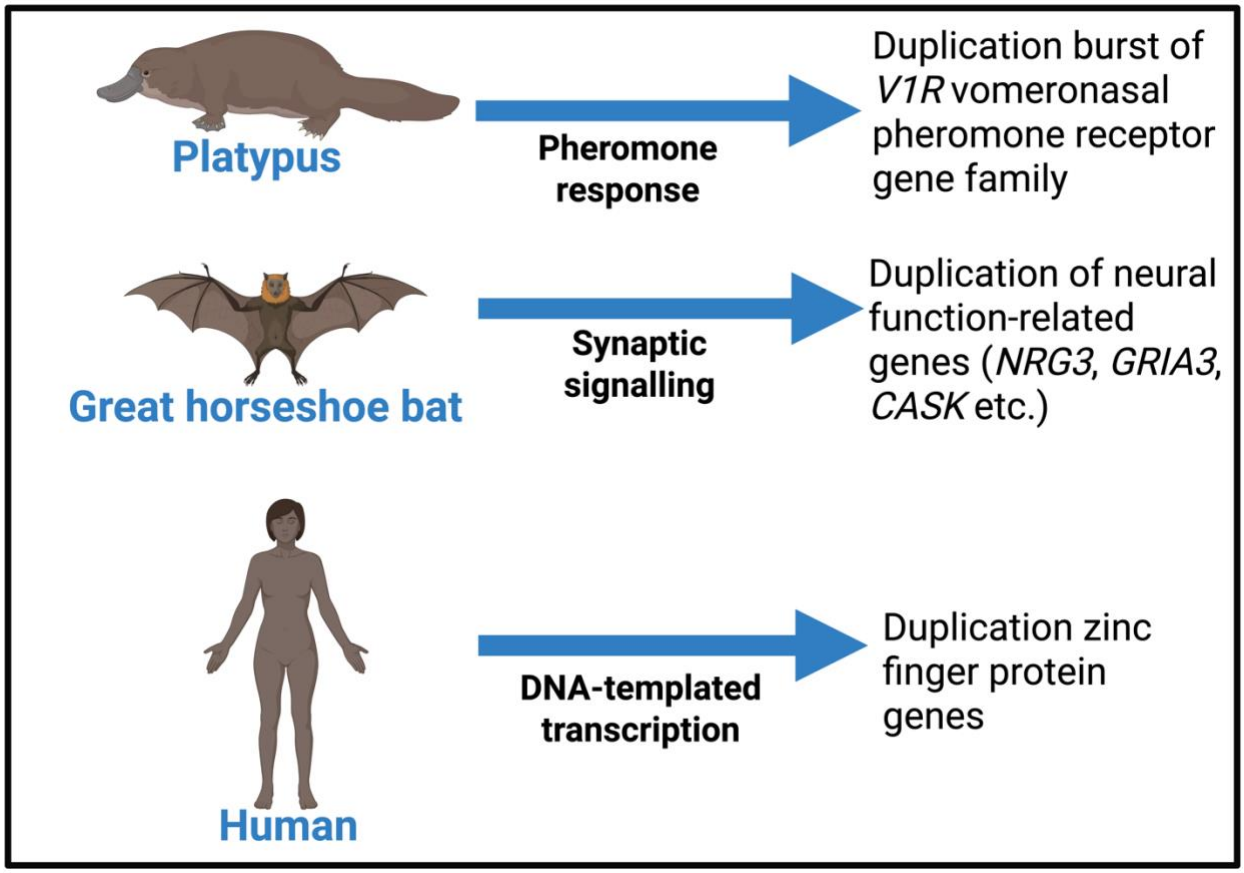

Figure 6. Functional enrichments of hyper-duplicated genes in the platypus, the great horseshoe bat, and humans.

Discussion

In this study, we generate segmental duplication calls from repeat-masked long-read-sequenced genomes of 117 vertebrate species and the starfish. We caution that like most pipelines for detecting segmental duplications, our approach is sensitive to repeat masking. While we have used the most comprehensive repeat libraries available to mask genomes and indirectly show in our analysis that repeat masking is adequate, minor effects of inconsistencies in repeat masking across species may nonetheless linger.

Based on these segmental duplication calls, we find that vertebrate genomes exhibit a higher a propensity for tandem duplications relative to interspersed duplications, mirroring previous findings in humans. However, in subtelomeric regions, birds and mammals show the opposite propensity towards interspersed duplications. Moreover, we find that tandem duplications tend to be larger than interspersed duplications in vertebrates.

We quantified segmental duplication landscapes based on network measures extracted from segmental duplication networks. Comparing the differences in these properties against phylogenetic distances between species, we find that vertebrate segmental duplication landscapes are rapidly evolving. The rapid evolution of duplication landscapes across species may confer unique adaptive potentials, enabling certain lineages to respond more effectively to selective pressures. This may, in part, explain why some lineages are virtual living fossils while others are more prone to undergoing adaptive radiation.

Finally, we show that hyper-duplicated genes in certain species are enriched for specific biological processes. For example, in the platypus, these genes are enriched for pheromone response, driven by expansion of the vomeronasal pheromone receptor *V1R* gene family. Such expansions highlight how segmental duplications can become molecular substrates for evolutionary novelty.

A strength of our study is the application of network-based methods to study genomic structure. Network properties capture relationships among duplicated genomic regions that are invisible to traditional analyses, offering deeper systems-level insights into the structural organization of genomes.

Future work could extend our network framework to a multi-layer segmental duplication network, where each species represents a layer. Here, inter-species edges can connect orthologous genomic regions between species. This fine-grained approach, analogous to multi-layer co-expression networks (Russell et al. 2023), would enable the identification of communities of duplicated genomic regions that persist across layers (species). It would allow for the systematic discovery of gene copy number expansions underlying both lineage-specific adaptation and convergent evolution. A practical starting point could focus on a smaller and more closely related group of species such as primates, rodents, or carnivores, for which genomic alignments are more reliable. Overall, although analyzing multi-layer segmental duplication networks would be computationally demanding, it would provide deeper comparative insights than analyses limited to metrics derived from single-species segmental duplication networks.

As long-read sequencing technologies continue to improve and become commonplace, network-based approaches will become even more powerful tools for characterizing complex genomic phenomena. Overall, our study highlights the importance of integrating network analysis with evolutionary biology to unravel the intricate dynamics of genome evolution.

Methods

Data

We downloaded FASTA files for all 150 curated assemblies from GenomeArk (https://www.genomeark.org/), which holds genomes sequenced by the vertebrate genome project (as of August 16th, 2022). These assemblies included sex chromosomes. In this study, sex chromosomes and autosomes were analyzed together. For humans, we used the hg38 reference assembly for better comparison to previous studies on duplications in humans.

Before looking for segmental duplications (SDs), i.e., duplications with a size of at least one kilobase, it is necessary to mask all short repeats within each genome as much as possible to avoid false SD positives and improve the running time of SD detection software. This is because BISER (Išerić et al. 2022) (the SD detection software we use) assumes that all short repeats are filtered out; if they are not, short tandem repeat clusters and transposable elements will be called as SDs despite not being biologically so. Indeed, if these repeats are not masked, short repeats or transposable elements may masquerade as segmental duplications. We used RepeatMasker (v4.1.3) for this purpose. Since masking was inadequate for many non-mammalian species using the default masking library (Dfam minimal), we employed the complete Dfam 3.6 library. Additionally, we augmented this masking using species-specific repeat libraries from RepBase (v10/26/2018). Indeed, for all mammals, all birds, all reptiles, half of the ray-finned species, and 2 out of 3 of cartilaginous fish species, we used species-specific libraries. For the starfish, one cartilaginous fish, and the other half of the ray-finned fishes, species-specific libraries were not available. So we used class-level repeat libraries in addition to the complete Dfam library to mask them. Thus, the repeat masking is not expected to be highly inconsistent across taxa. Masking took from a few hours to a few weeks (it was especially performance-intensive on *Actinopterygii* genomes), and masked on average 25% of each genome: from 2% (*Asteroidea*) to 82% (*Bal. musculus*). We note that at the time of our analysis (data access date August 16th, 2022), the starfish genome was the only invertebrate genome available in the GenomeArk database.

Once the genomes were masked, we used BISER v1.2 (Išerić et al. 2022) to find all SDs within them. BISER relies on the SD error model (Numanagic et al. 2018) that assumes that (1) sequence similarity of SDs is higher than 75% (in other words, the error is below 25%), (2) random point mutations account for 15% of the total error and that they follow Poisson error distribution and are independent of each other (Jain et al. 2020); and (3) large indels and block variants account for the remaining 10% of the total error. The higher error rate threshold enables BISER to detect older SDs that occurred before the primate split in evolutionary history (the standard cutoff error threshold of 10% (Bailey and Eichler 2006) is sufficient only for detecting primate SDs (Numanagic et al. 2018)).

BISER finds all SDs in a given genome using a *k*-mer-based sketch of Jaccard distance through MinHash (Broder 1997) and *k*-mer windowing (Jain et al. 2020) to approximate edit distance under the described SD error model. This model allows it to quickly calculate SDs in the whole genome and decompose them into core blocks (Jiang et al. 2007). However, like all SD detection tools, BISER is extremely sensitive to the repeat masking quality; in our runs, it took from 4 minutes (good quality masking) to more than 10 days (low-quality masking; median time: 10 minutes) to find all SDs in a given genome. Our pipeline ran successfully for 118 genomes. All analysis in this manuscript is based on those 118 genomes. 109 of the 118 assemblies are pseudo-haploid consensus-collapsed assemblies. The remaining 9 are true (maternal or paternal) haploid assemblies obtained from trio-binning.

Evolutionary distances

To get a phylogenetic tree for the 118 species analyzed here **(Figure 1A)**, we used a list of Latin binomials corresponding to these species and used it as input for TimeTree 5 (Kumar et al. 2022). On the tree, 14 species were replaced by close substitutes (**Table S2**). For our analysis, we used these close substitutes as proxies for the original species to measure phylogenetic distances. When using substitutes, we still use the names of the original species and append them with an asterisk. The output from TimeTree was a tree file in the Newick format with branch lengths in terms of the number of years. We used the `cophenetic.phylo()` function from the "ape" library in R (Paradis and Schliep 2018) to obtain evolutionary distances (in years) between each pair of species. The tree was visualized using the interactive Tree Of Life (iTOL) (Letunic and Bork 2021).

Network construction

We construct a weighted network for each species from the data described above. We construct the network by considering consecutive windows, sized $\Delta = 5000$ base pairs (5 kbp) on chromosomes **(Figure 7A)**. We use each such window as a node and connect pairs of nodes by weighted edges. We determine weighted edges as follows. In general, the duplication size does not coincide with Δ or multiples of it. For instance, assume that a region R_x of length 1.5Δ bp is duplicated in another region R_y of length 2.75Δ bp and that the similarity score between them is q. We assume that the starting coordinate of the region R_x corresponds to the starting coordinate of region R_y, and the same for the ending coordinate. Therefore, a window of size Δ in R_x corresponds to a subregion of R_y of size $2.75\Delta \times (1\Delta/1.5\Delta) = 1.83\Delta$ bp. For expository purposes, we also assume that R_x is contained in two nodes, n_1^x of size 1Δ bp and n_2^x of size 0.5Δ bp, and that R_y is contained in three nodes, n_1^y of size 0.85Δ bp, n_2^y of size 1Δ bp, and n_3^y of size 0.9Δ bp **(Figure 7B)**. We calculate the edges, and their weights as follows. We assume that the size of R_y is larger than R_x without loss of generality. First, we find the subregions of R_y that overlap the projection of the subregion of node n_i^x on R_y, for each ith node contained in R_x. This procedure determines the nodes in R_y that are adjacent to n_i^x by an edge. For example, in **Figure 7B**, node n_1^x is projected to the first 1.83Δ bp of R_y. Therefore, we connect n_1^x to n_1^y and n_2^y by an edge. Note that the entirety of n_1^y (of size 0.85Δ bp) overlaps with the projection of n_1^x and that the first 0.98Δ bp of n_2^y overlaps with the projection of n_1^x. Likewise, n_2^x is projected to the last 0.92Δ bp of R_y. Because n_3^y accounts for the last 0.9Δ bp of R_y, we connect n_2^x to each of n_2^y and n_3^y by an edge. Note that the last 0.02Δ bp of n_2^y overlaps with the projection of n_2^x and that the entirety of n_3^y (of size 0.9Δ bp) overlaps with the projection of n_2^x. In this manner, we define an edge between a node in R_x and a node in R_y if and only if there is a positive overlap between R_y and the projection of R_x on R_y. Second, we define the weight of the edge between n_i^x and n_j^y by

$$w_{i,j}^{x,y} = \frac{q \times (\text{size of part of } n_j^y \text{ covered by the projection of } n_i^x \text{ on } R_y)}{\Delta}. \quad (1)$$

The generated networks are undirected. We allow a node to participate in edges with multiple duplications and edges with other nodes on the same or different chromosomes, as shown in **Figures 7A-B**. We show a part of the network of *Homo sapiens* in **Figure 7C**. The first component of each node label represents the chromosome number, and the second component represents the starting coordinate of the node in kbp. We recall that the node spans from the starting position to the subsequent 5 kbp. For example, node (Chr 1, 23260) is located on

chromosome 1 and spans from 23260 kbp to 23264.999 kbp. We also show the degree distribution of three representative (chosen arbitrarily) networks from three taxonomic classes: ray-finned fish, mammals, and birds in **Figures 7D**, **E**, and **F**, respectively.

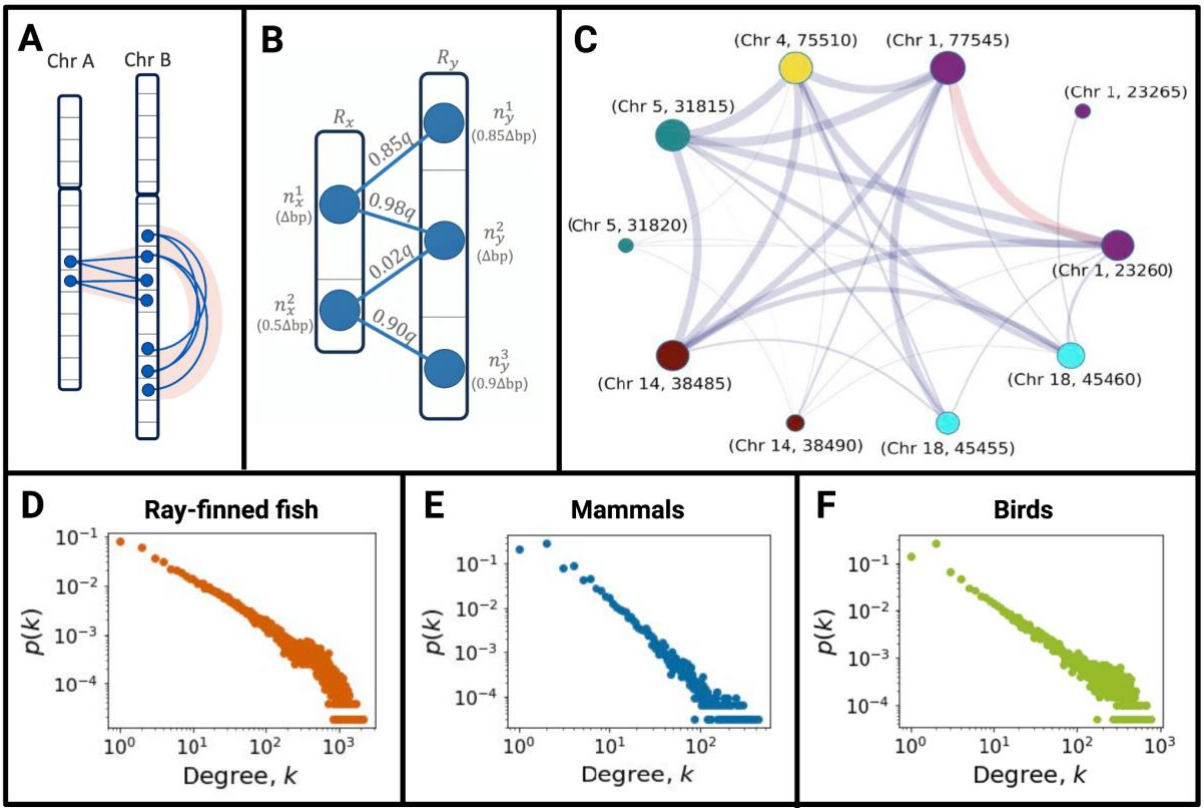

Figure 7. Network construction and degree distribution. A. Schematic for edge construction between two duplicated regions. **B.** Examples of edges for two chromosomes. The part of the network shown here is a magnification of edges between Chr A and Chr B in panel **A**. **C.** A part of the weighted segmental duplication network constructed for *Homo sapiens*. Different node colors represent different chromosomes. The number followed by the comma represents the starting coordinate (kbp) of the node on the chromosome **D-F.** The degree distribution of segmental duplication network from a representative species of ray-finned fish, mammals, and birds, respectively.

Network properties

To understand the underlying duplication patterns, their evolutionary significance, and their potential impact on genomic architecture across species, we computed the following network properties for each species.

We denote the adjacency matrix of the segmental duplication network by $A = (a_{ij})$, where the nodes are the genomic regions of size 5 kbp. The adjacency matrix represents whether each node pair is connected by an edge. We use the duplication similarity score as the weight of the edge between the ith and jth nodes, which we denote w_{ij}. Note that $a_{ij} = a_{ji} = 1$ if $w_{ij} = w_{ji} > 0$, $a_{ij} = a_{ji} = 0$ if $w_{ij} = w_{ji} = 0$, and by assumption that $a_{ii} = w_{ii} = 0$. The degree of the ith node is defined by

$$k_i = \sum_{j=1;\, j\neq i}^{N} a_{ij}, \tag{2}$$

where N is the number of nodes in the network. The degree represents the number of other nodes that the focal node is adjacent to by an edge. A high-degree node is a region that has been duplicated many times and, therefore, is suggested to be a hotspot of duplication activity.

The weighted degree, also called the node strength (Barrat et al. 2004), of the ith node is given by

$$s_i = \sum_{j=1;\, j\neq i}^{N} w_{ij}. \tag{3}$$

The node strength measures the cumulative similarity of all duplications for a focal genomic region across the genome. A node (region) with high strength is a region that has been duplicated many times and, on average, has high sequence similarity with its duplicates.

We have explained network metrics that are calculated for each node. We now explain other network metrics that are measured for each segmental duplication network, or equivalently, for each species. Some metrics use the node degree or the weighted degree of all the nodes in the segmental duplication network.

The mean degree is given by

$$\bar{k} = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}. \tag{4}$$

It is the average of the node degrees over all the nodes. It represents the average number of duplication events per genomic region. A higher mean degree indicates a more densely interconnected duplication network, possibly reflecting extensive segmental duplication events.

The mean node strength is given by

$$\bar{s} = \frac{1}{N}\sum_{i=1}^{N} s_i = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}. \tag{5}$$

It is a measure of the duplication similarity score per pair of genomic regions. Larger $\bar{s}$ values suggest stronger duplication relationships between genomic regions overall in the entire network.

We measure the heterogeneity of a distribution by the coefficient of variation (CV). The CV of the degree distribution and that of the strength distribution is given by

$$CV_k = \frac{\sigma_k}{\bar{k}} \tag{6}$$

and

$$CV_s = \frac{\sigma_s}{\bar{s}}, \tag{7}$$

respectively, where σ_k and σ_s are the sample standard deviations of the node's degree and strength, respectively. The CV of the degree distribution measures how much variation across the *N* nodes exists in terms of the number of duplication connections a node has. A large CV_k value indicates that some regions have markedly high degrees, representing duplication hubs, while other regions have small degrees. A small CV_k value suggests a relatively uniform duplication pattern across genomic regions. The CV of the node strength distribution measures the variation across the *N* nodes in terms of the cumulative duplication similarity score for a node. A large CV_s value suggests that some genomic regions have markedly strong duplication relationships with other regions. A small CV_s value indicates that duplication similarities are more evenly distributed across the genome.

The density of the network is defined by

$$\rho = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} \tag{8}$$

and represents the fraction of node pairs connected by an edge (Newman 2018; Wasserman 1994). High density means that segmental duplications are widespread, with many genomic region pairs sharing duplications. Low density suggests that duplications occur in a more isolated manner, possibly reflecting evolutionary constraints.

The weighted density of the network is given by

$$D = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \tag{9}$$

and represents the average edge weight over all node pairs, including those that are not adjacent by an edge (Wasserman 1994). The weighted density reflects the overall duplication similarity across the genome. A large D value suggests that duplications tend to be strong in terms of the similarity score and widespread. A small D value suggests that duplications are weaker or sparser.

The average clustering coefficient (Watts and Strogatz 1998) is defined by

$$C = \frac{1}{N} \sum_{i \in G} C_i, \tag{10}$$

where C_i is the local clustering coefficient for the ith node given by

$$C_i = \frac{2 \times (\text{Number of pairs of neighbors of node } i \text{ that are adjacent})}{k_i\,(k_i - 1)}. \tag{11}$$

The average clustering coefficient represents the propensity that two neighbors of a node are also directly connected. A high *C* value suggests that duplications tend to form connected clusters in the sense of triangles, possibly indicating duplication hotspots. A low *C* value suggests that duplications occur in a more random fashion.

In addition to these properties, we also measure the mean, CV, and skewness of the edge weight, w_{ij} (Watts and Strogatz 1998). The mean edge weight represents the average similarity score between duplicated regions per duplication event. A high mean edge weight suggests recent or highly conserved duplications. A low mean weight suggests that duplications are generally divergent, suggesting older duplication events or functional divergence. The CV of the edge weight measures variability in duplication similarity across the duplicated region pairs. A large value of the CV of the edge weight implies a mix of old and new duplications. A small value of the CV of the edge weight suggests that the duplication similarity is relatively uniform across duplicated region pairs, indicating consistent duplication mechanisms for different duplicated region pairs. The skewness for the edge weight is given by

$$\gamma = \frac{\frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}\left(w_{ij}-\frac{\bar{s}}{k}\right)^3}{\sigma_w^3}, \tag{12}$$

where σ_w is the sample standard deviation of the edge weight. The skewness of the edge weight measures the asymmetry in the distribution of the duplication similarity score (i.e., edge weight). Positive skewness ($\gamma > 0$) suggests that most duplication events have low similarity, with a few having exceptionally high similarity. Negative skewness ($\gamma < 0$) suggests that most duplications have high similarity, which may indicate genome regions with conserved segmental duplications. It should be noted that the CV also measures the extent of variation in the duplication similarity score; however, unlike the skewness, the CV does not measure the asymmetry.

We quantify the abundance and strength of cis-duplications relative to those of trans-duplications by measuring the so-called density ratio. To compute the density ratio for a given chromosome to other chromosomes, we first consider the density, ρ_{cis}, of the subnetwork that consists of the edges that connect two nodes in the same chromosome. Second, we compute the density, ρ_{trans}, of the subnetwork consisting of the edges connecting the node in the given chromosome to other chromosomes. Then, we compute the density ratio $\frac{\rho_{\text{cis}}}{\rho_{\text{trans}}}$ for the given chromosome. Finally, we calculate the mean of the density ratios over all chromosomes for a given species. We measure density ratios for both the unweighted network and the weighted network. A high unweighted density ratio implies that duplications mostly occur within the same chromosome. This case suggests localized duplication mechanisms such as tandem duplications or segmental duplications due to replication errors. A low unweighted density ratio implies that inter-chromosomal duplications are relatively frequent, which could be due to chromosomal translocations, rearrangements, or interspersed duplications. A high weighted density ratio signifies that intra-chromosomal duplications exhibit stronger similarity scores than inter-chromosomal ones overall. This case implies that intra-chromosomal duplications are either more conserved or more abundant than inter-chromosomal ones. A low weighted density ratio suggests that inter-chromosomal duplications have strong average similarity or are abundant.

We also explore how the node strength depends on the node degree for each network in a potentially nonlinear manner. We fit the following relation (Barrat et al. 2004):

$$\ln\bar{s} = \alpha + \beta \ln k, \tag{13}$$

where k is the node's degree and $\bar{s}$ is the node strength averaged over all the nodes whose degree is k. To estimate α and β, we fitted a line to the set of points $\{\ln k, \ln\bar{s}\}$ using sklearn

(Pedregosa et al. 2011) in Python. Equation (13) implies that $\bar{s}$ is proportional to k^{β}. If $\beta > 1$, then hubs, i.e., large-degree nodes, tend to have stronger duplications, implying that duplication hubs preferentially duplicate with highly similar regions. If $\beta = 1$, there is no such correlation between the node's degree and edge weight, implying that the duplication strength is independent of how many duplications a node has experienced. Various empirical weighted networks yield $\beta > 1$ (Barrat et al. 2004).

To calculate the network properties, we use NetworkX version 2.8.5 in Python (Hagberg et al. 2008).

Comparison with phylogenetic tree

We calculate all the 14 properties for each species in the dataset. For a given property, which is a scalar, we quantify the dissimilarity between two species p and p' in terms of the property by

$$d_{\Xi}(p,p') = |\Xi_p - \Xi_{p'}|, \tag{14}$$

where Ξ_p and $\Xi_{p'}$ represent the network property for species p and p', respectively. We compare $d_{\Xi}(p,p')$ to the phylogenetic distance between p and p' to explore which network-based measure may be associated with the phylogeny. We measure the correlation between the phylogenetic distance and $d_{\Xi}(p,p')$ by regarding each pair of species as a sample.

Here, we note an important concern. Our correlation analysis involves 6,962 species pairs (118 x 117/2). It is well-recognized in statistics that with such a large number of data points, hypothesis testing could yield "significant" results even for trivial effects (Berkson 1938; Cohen 1994). This phenomenon ("crud factor") (Clayton 2021) cautions against overinterpreting p-values in large datasets. Therefore, rather than relying solely on p-value thresholds, we consider the two variables to be correlated only if $|r| > 0.3$, the conventional threshold for medium effect size (Cohen 1988) that is routinely used in biology (Fredlund et al. 2012; Friedman and Alm 2012; Pai et al. 2019; Polk et al. 2025).

For the comparison in taxonomic class level, we omitted the analysis of the starfish, cartilaginous fish, and reptile classes because each of them had at most five species.

Data Availability

The segmental duplication calls obtained using BISER, for the 118 genomes analyzed in this study have been deposited on FigShare. In addition to the files containing segmental duplication calls, the file “Species_and_Assembly_Info.csv,” containing taxonomic details of species, genome sizes, qualities of assembly, and the link to download the relevant FASTA (genomic sequence) files, is also available on FigShare. This data can be accessed using the following link:https://figshare.com/articles/dataset/Segmental_Duplication_Calls_BISER_Vertebrates_LongReadGenomes/27859545.

Conflict of Interest

The authors declare no competing interests.

Acknowledgment

I.N. was supported by the National Science and Engineering Council of Canada Discovery Grant (RGPIN-04973), Canada Research Chairs Program, Canada Foundation for Innovation’s John R. Evans Leaders Fund (CFI JELF) and B.C. Knowledge Development Fund (BCKDF). N.M. also acknowledges support from the Japan Science and Technology Agency (JST) Moonshot R&D (under grant no. JPMJMS2021), the National Science Foundation (under grant no.2052720), and JSPS KAKENHI (under grant nos.JP 23H03414, 24K14840, and W24K030130). O.G. acknowledges support from the National Science Foundation (under grant nos.2049947 and 2123284). O.G. and N.M. acknowledge support from the National Institute of General Medical Sciences (under grant no.1R01GM148973-01). The funders had no role in study design, data collection and analysis, publication decisions, or manuscript preparation.

References

Abdullaev ET, Haridoss DA, Arndt PF. 2023. Reconstruction of segmental duplication rates and associated genomic features by network analysis. Available from: https://www.biorxiv.org/content/10.1101/2023.03.18.533287.abstract

Abdullaev ET, Umarova IR, Arndt PF. 2021. Modelling segmental duplications in the human genome. *BMC Genomics* [Internet] 22:496. Available from: https://link.springer.com/article/10.1186/s12864-021-07789-7

Alexandrou MA, Swartz BA, Matzke NJ, Oakley TH. 2013. Genome duplication and multiple evolutionary origins of complex migratory behavior in Salmonidae. *Mol. Phylogenet. Evol.* [Internet] 69:514–523. Available from: https://www.sciencedirect.com/science/article/pii/S1055790313003035

Allendorf FW, Bassham S, Cresko WA, Limborg MT, Seeb LW, Seeb JE. 2015. Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *J. Hered.* [Internet] 106:217–227. Available from: https://academic.oup.com/jhered/article/106/3/217/2960044

Al-Sallami HS, Cheah SL, Han SY, Liew J, Lim J, Ng MA, Solanki H, Soo RJ, Tan V, Duffull SB. 2014. Between-subject variability: should high be the new normal? *Eur. J. Clin. Pharmacol.* [Internet] 70:1403–1404. Available from: https://idp.springer.com/authorize/casa?redirect_uri=https://link.springer.com/article/10.1007/s00228-014-1740-8&casa_token=DwuQCxpEU7kAAAAA:mEGZow3G9evNQz18MDGKCsk6MPeyRx8atE5-bPyyyBRBmp9C3B14bH2rov4GXiqek7I9aDDiQ0-FXfoh

Aqil A, Gokcumen O. 2025. Read mismapping from segmental duplications drives spurious trans associations. *bioRxiv* [Internet]:2025.10.08.681258. Available from: http://dx.doi.org/10.1101/2025.10.08.681258v1

Aqil A, Li Y, Wang Z, Islam S, Russell M, Kallak TK, Saitou M, Gokcumen O, Masuda N. 2025. Switch-like gene expression modulates disease risk. *Nat. Commun.* [Internet] 16:5323. Available from: http://dx.doi.org/10.1038/s41467-025-60513-x

Banse P, Elena SF, Beslon G. 2024. Innovation in viruses: fitness valley crossing, neutral landscapes, or just duplications? *Virus Evol.* [Internet] 10. Available from: http://dx.doi.org/10.1093/ve/veae078

Bergman CM, Quesneville H, Anxolabéhère D, Ashburner M. 2006. Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome. *Genome Biol.* [Internet] 7:R112. Available from: https://link.springer.com/article/10.1186/gb-2006-7-11-r112

Berkson J. 1938. Some difficulties of interpretation encountered in the application of the chi-square test. *J. Am. Stat. Assoc.* [Internet] 33:526–536. Available from: http://dx.doi.org/10.1080/01621459.1938.10502329

Bi X, Wang K, Yang L, Pan H, Jiang Haifeng, Wei Q, Fang M, Yu H, Zhu C, Cai Y, et al. 2021. Tracing the genetic footprints of vertebrate landing in non-teleost ray-finned fishes. *Cell* [Internet] 184:1377-1391.e14. Available from: https://www.cell.com/cell/fulltext/S0092-

86742100089-1

Brunet FG, Roest Crollius H, Paris M, Aury J-M, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol. Biol. Evol.* [Internet] 23:1808–1816. Available from: https://dx.doi.org/10.1093/molbev/msl049

Campbell MA, Hale MC, McKinney GJ, Nichols KM, Pearse DE. 2019. Long-term conservation of ohnologs through partial tetrasomy following whole-genome duplication in salmonidae. *G3 (Bethesda)* [Internet] 9:2017–2028. Available from: https://academic.oup.com/g3journal/article-abstract/9/6/2017/6028030

Carroll SB. 2005. Evolution at two levels: on genes and form. *PLoS Biol.* [Internet] 3:e245. Available from: http://dx.doi.org/10.1371/journal.pbio.0030245

Cheng F, Wu J, Cai X, Liang J, Freeling M, Wang X. 2018. Gene retention, fractionation and subgenome differences in polyploid plants. *Nat. Plants* [Internet] 4:258–268. Available from: http://dx.doi.org/10.1038/s41477-018-0136-7

Cheng P, Huang Y, Du H, Li C, Lv Y, Ruan R, Ye H, Bian C, You X, Xu J, et al. 2019. Draft genome and complete Hox-cluster characterization of the sterlet (Acipenser ruthenus). *Front. Genet.* [Internet] 10:776. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2019.00776/full

Clayton A. 2021. Bernoulli's fallacy: Statistical illogic and the crisis of modern science. Available from: https://books.google.com/books?hl=en&lr=&id=HzQGEAAAQBAJ&oi=fnd&pg=PT9&dq=bernoulli+fallacy&ots=PEt5m1aogj&sig=fikjfkKu_9OEL99DOnuv97WkO4s

Cohen J. 1988. Statistical power analysis for the behavioral sciences. 2nd ed. London, England: Routledge Available from: https://www.taylorfrancis.com/books/mono/10.4324/9780203771587/statistical-power-analysis-behavioral-sciences-jacob-cohen

Cohen J. 1994. The earth is round (p < .05). *American Psychologist* [Internet] 49:997–1003. Available from: http://dx.doi.org/10.1037/0003-066X.49.12.997

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.* [Internet] 4:841–852. Available from: http://dx.doi.org/10.1038/s41559-020-1166-x

Elliott TA, Gregory TR. 2015. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* [Internet] 370:20140331. Available from: https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2014.0331

Fiston-Lavier A-S, Anxolabehere D, Quesneville H. 2007. A model of segmental duplication formation in Drosophila melanogaster. *Genome Res.* [Internet] 17:1458–1470. Available from: http://dx.doi.org/10.1101/gr.6208307

Francoeur ER, Audano PA, Ferraj A, Balachandran P, Beck CR. 2025. Segmental duplication-mediated rearrangements alter the landscape of mouse genomes. *bioRxivorg* [Internet].

Available from: http://dx.doi.org/10.1101/2025.07.18.665526

Fredlund E, Staaf J, Rantala JK, Kallioniemi O, Borg A, Ringnér M. 2012. The gene expression landscape of breast cancer is shaped by tumor protein p53 status and epithelial-mesenchymal transition. *Breast Cancer Res.* [Internet] 14:R113. Available from: http://dx.doi.org/10.1186/bcr3236

Friedman J, Alm EJ. 2012. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* [Internet] 8:e1002687. Available from: http://dx.doi.org/10.1371/journal.pcbi.1002687

Gregory TR, Andrews CB, McGuire JA, Witt CC. 2009. The smallest avian genomes are found in hummingbirds. *Proc. Biol. Sci.* [Internet] 276:3753–3757. Available from: https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2009.1004?casa_token=v8p1VPnti-kAAAAA:bpqQIoI7C8uxRAtruWSedYAQ9I2XPS8fCHWE-QNnOIfnzNAaG2ofZ5v5UtvpHaBrB_xDmiCsOPJednxK&casa_token=O1ye1q9c1QkAAAAA:g5N_EImZHN-0uPYTQ61OHbz2kfqEXUoO53L2ysnsBzsHiQiu8E8n3I3BX98BQd8cGTa3-gK4k2uOHuZ8

Grus WE, Shi P, Zhang J. 2007. Largest vertebrate vomeronasal type 1 receptor gene repertoire in the semiaquatic platypus. *Mol. Biol. Evol.* [Internet] 24:2153–2157. Available from: http://dx.doi.org/10.1093/MOLBEV/MSM157

Haldane JBS. 1933. The part played by recurrent mutation in evolution. *Am. Nat.* [Internet] 67:5–19. Available from: http://dx.doi.org/10.1086/280465

Hughes AL, Hughes MK. 1995. Small genomes for better flyers. *Nature* [Internet] 377:391. Available from: https://ui.adsabs.harvard.edu/abs/1995Natur.377..391H/abstract

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* [Internet] 11:97–108. Available from: http://dx.doi.org/10.1038/nrg2689

Išerić H, Alkan C, Hach F, Numanagić I. 2022. Fast characterization of segmental duplication structure in multiple genome assemblies. *Algorithms Mol. Biol.* [Internet] 17:4. Available from: https://link.springer.com/article/10.1186/s13015-022-00210-2

Kapusta A, Suh A, Feschotte C. 2017. Dynamics of genome size evolution in birds and mammals. *Proc. Natl. Acad. Sci. U. S. A.* [Internet] 114:E1460–E1469. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.1616702114

Kim PM, Lam HYK, Urban AE, Korbel JO, Affourtit J, Grubert F, Chen X, Weissman S, Snyder M, Gerstein MB. 2008. Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Res.* [Internet] 18:1865–1874. Available from: https://genome.cshlp.org/content/18/12/1865.short

Krieger J, Fuerst PA. 2002. Evidence for a slowed rate of molecular evolution in the order acipenseriformes. *Mol. Biol. Evol.* [Internet] 19:891–897. Available from: https://academic.oup.com/mbe/article-abstract/19/6/891/1094801

Kumar S, Chauhan MP, Tomar A, Kasana RK. 2018. Coefficient of variation (GCV & PCV),

heritability and genetic advance analysis for yield contributing characters in rice (Oryza sativa L.). *Journal of Pharmacognosy and Phytochemistry* [Internet] 7:2161–2164. Available from: https://www.phytojournal.com/archives?year=2018&vol=7&issue=3&ArticleId=4509

Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. 2022. TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol. Biol. Evol.* [Internet] 39. Available from: http://dx.doi.org/10.1093/molbev/msac174

Langham RJ, Walsh J, Dunn M, Ko C, Goff SA, Freeling M. 2004. Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics* [Internet] 166:935–945. Available from: https://dx.doi.org/10.1093/genetics/166.2.935

Li Y, Zhang J. 2025. On the probability of reaching high peaks in fitness landscapes by adaptive walks. *Mol. Biol. Evol.* [Internet] 42:msaf066. Available from: https://dx.doi.org/10.1093/molbev/msaf066

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* [Internet] 533:200–205. Available from: https://www.nature.com/articles/nature17164

Lin X-H, Zhang T-W. 2010. Structure and dynamics of artificial regulatory networks evolved by segmental duplication and divergence model. *Int. J. Autom. Comput.* [Internet] 7:105–114. Available from: https://link.springer.com/article/10.1007/s11633-010-0105-8

Lin Y-L, Gokcumen O. 2019. Fine-Scale Characterization of Genomic Structural Variation in the Human Genome Reveals Adaptive and Biomedically Relevant Hotspots. *Genome Biol. Evol.* [Internet] 11:1136–1151. Available from: http://dx.doi.org/10.1093/gbe/evz058

Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ. 2005. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* [Internet] 437:94–100. Available from: http://dx.doi.org/10.1038/nature04029

Lowe S, Browne M, Boudjelas S, De Poorter M. 2000. 100 of the world's worst invasive alien species: a selection from the global invasive species database. academia.edu Available from: https://www.academia.edu/download/33655728/100_world_worst_invasive_alien_species_English.pdf

Lü J, Wang P. 2020. Modeling and analysis of bio-molecular networks. 1st ed. Singapore, Singapore: Springer Available from: https://link.springer.com/content/pdf/10.1007/978-981-15-9144-0.pdf

Macqueen DJ, Johnston IA. 2014. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proc. Biol. Sci.* [Internet] 281:20132881. Available from: https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2013.2881

Moffett AS, Falcón-Cortés A, Di Pierro M. 2025. Quantifying the influence of genetic context on duplicated mammalian genes. *bioRxivorg* [Internet]. Available from: http://dx.doi.org/10.1101/2025.04.03.647042

Ohno S. 1970. Evolution by gene duplication. 1970th ed. Berlin, Germany: Springer Available from: https://link.springer.com/book/10.1007/978-3-642-86659-3

907 Osmanski AB, Paulat NS, Korstian J, Grimshaw JR, Halsey M, Sullivan KAM, Moreno-Santillán
908 DD, Crookshanks C, Roberts J, Garcia C, et al. 2023. Insights into mammalian TE diversity
909 through the curation of 248 genome assemblies. *Science* [Internet] 380:eabn1430. Available
910 from: http://dx.doi.org/10.1126/science.abn1430

911 Pai S, Hui S, Isserlin R, Shah MA, Kaka H, Bader GD. 2019. netDx: interpretable patient
912 classification using integrated patient similarity networks. *Mol. Syst. Biol.* [Internet] 15:e8497.
913 Available from: http://dx.doi.org/10.15252/msb.20188497

914 Pan D, Zhang L. 2008. Tandemly arrayed genes in vertebrate genomes. *Comp. Funct.*
915 *Genomics* [Internet] 2008:545269. Available from: http://dx.doi.org/10.1155/2008/545269

916 Pastor-Satorras R, Smith E, Solé RV. 2003. Evolving protein interaction networks through gene
917 duplication. *J. Theor. Biol.* [Internet] 222:199–210. Available from:
918 https://www.sciencedirect.com/science/article/pii/S0022519303000286

919 Polk EA, Aqil A, Collins JB, Tanny RE, Andersen EC, Gokcumen O, Ferkey DM. 2025. Wild
920 strains reveal natural variation in C. elegans avoidance behaviors. *G3 (Bethesda)* [Internet].
921 Available from: http://dx.doi.org/10.1093/g3journal/jkaf243/8281743

922 Redmond AK, Casey D, Gundappa MK, Macqueen DJ, McLysaght A. 2023. Independent
923 rediploidization masks shared whole genome duplication in the sturgeon-paddlefish ancestor.
924 *Nat. Commun.* [Internet] 14:2879. Available from: https://www.nature.com/articles/s41467-023-
925 38714-z

926 Russell M, Aqil A, Saitou M, Gokcumen O, Masuda N. 2023. Gene communities in co-
927 expression networks across different tissues. *PLoS Comput. Biol.* [Internet] 19:e1011616.
928 Available from: http://dx.doi.org/10.1371/journal.pcbi.1011616

929 Solé RV, Pastor-Satorras R, Smith E, Kepler TB. 2002. A model of large-scale proteome
930 evolution. *Adv. Complex Syst.* [Internet] 05:43–54. Available from:
931 https://www.worldscientific.com/doi/abs/10.1142/S021952590200047X

932 Vázquez A, Flammini A, Maritan A, Vespignani A. 2003. Modeling of protein interaction
933 networks. *Complexus* [Internet] 1:38–44. Available from: https://karger.com/cpu/article-
934 abstract/1/1/38/74295

935 Vollger MR, Dishuck PC, Harvey WT, DeWitt WS, Guitart X, Goldberg ME, Rozanski AN, Lucas
936 J, Asri M, Human Pangenome Reference Consortium, et al. 2023. Increased mutation and gene
937 conversion within human segmental duplications. *Nature* [Internet] 617:325–334. Available from:
938 https://www.nature.com/articles/s41586-023-05895-y

939 Vollger MR, Dishuck PC, Sorensen M, Welch AE, Dang V, Dougherty ML, Graves-Lindsay TA,
940 Wilson RK, Chaisson MJP, Eichler EE. 2019. Long-read sequence and assembly of segmental
941 duplications. *Nat. Methods* [Internet] 16:88–94. Available from:
942 http://dx.doi.org/10.1038/s41592-018-0236-3

943 Wang J, Zhang G, Sun C, Chang L, Wang Y, Yang X, Chen G, Itgen MW, Haley A, Tang J, et
944 al. 2025. DNA gains and losses in gigantic genomes do not track differences in transposable
945 element-host silencing interactions. *Commun. Biol.* [Internet] 8:704. Available from:
946 http://dx.doi.org/10.1038/s42003-025-08127-3

947 Zhang L, Lu HHS, Chung W-Y, Yang J, Li W-H. 2005. Patterns of segmental duplication in the
948 human genome. *Mol. Biol. Evol.* [Internet] 22:135–141. Available from:
949 https://dx.doi.org/10.1093/molbev/msh262

950 Zhou Y, Shearwin-Whyatt L, Li J, Song Z, Hayakawa T, Stevens D, Fenelon J, Peel E, Cheng
951 Y, Pajpach F, et al. 2021. Platypus and echidna genomes reveal mammalian biology and
952 evolution. *Nature* [Internet] 592:756–762. Available from: http://dx.doi.org/10.1038/s41586-020-
953 03039-0

954 Zhu W, Wang Z, Li H, Li P, Ni L, Jiao L, Ren Y, You P. 2022. A chromosome-level genome of
955 Brachymystax tsinlingensis provides resources and insights into salmonids evolution. *G3*
956 [Internet] 12:jkac162. Available from: https://academic.oup.com/g3journal/article-
957 pdf/doi/10.1093/g3journal/jkac162/45183708/jkac162.pdf

958

Supplementary Information

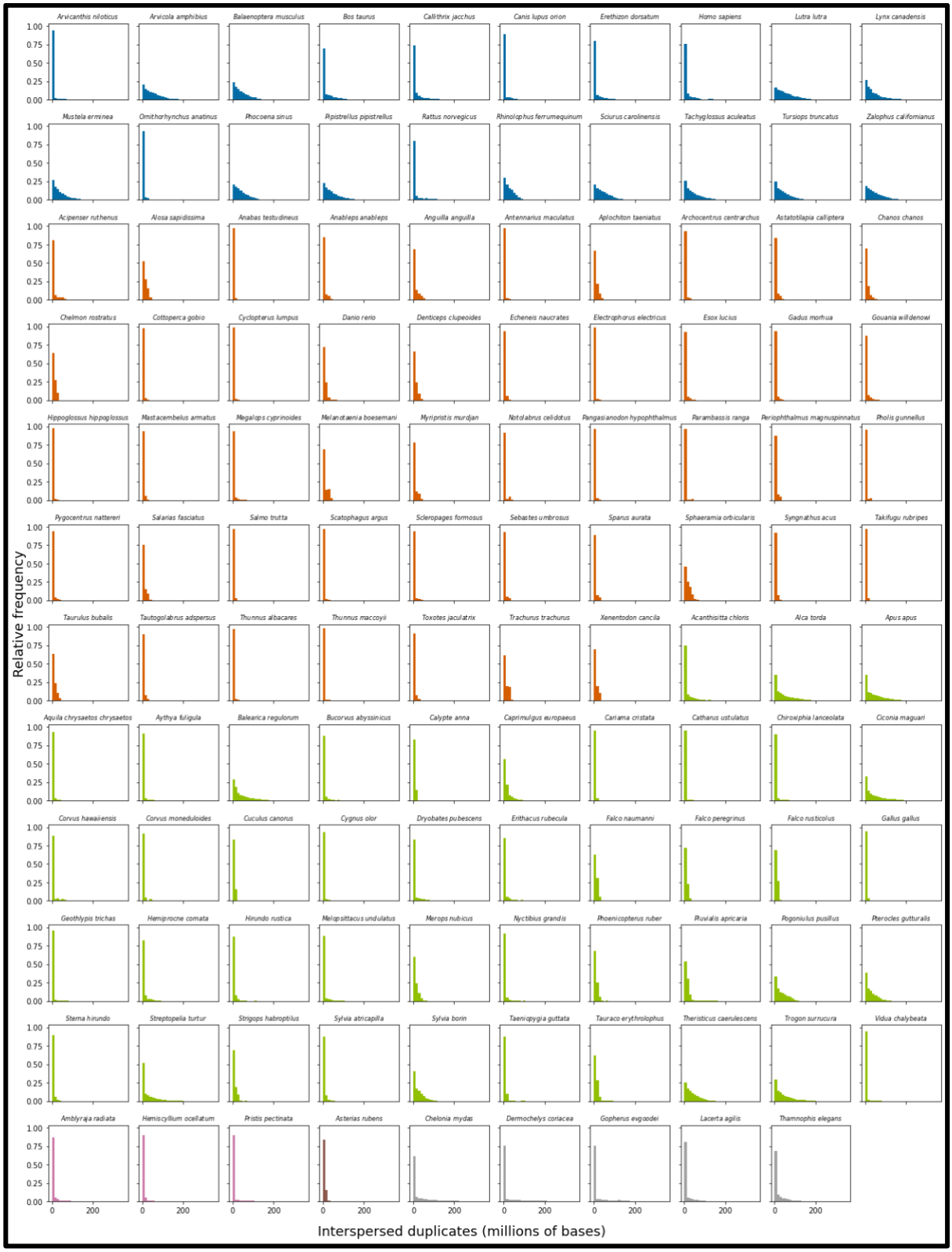

Figure S1. Distribution of distances between pairs of intrachromosomal duplicates. Each subplot represents distance distribution for one species. The histograms are color-coded by taxonomic class. The observed skewed

964 distributions indicate that most duplicate pairs are located relatively close to each other, although some are separated
965 by large genomic distances.

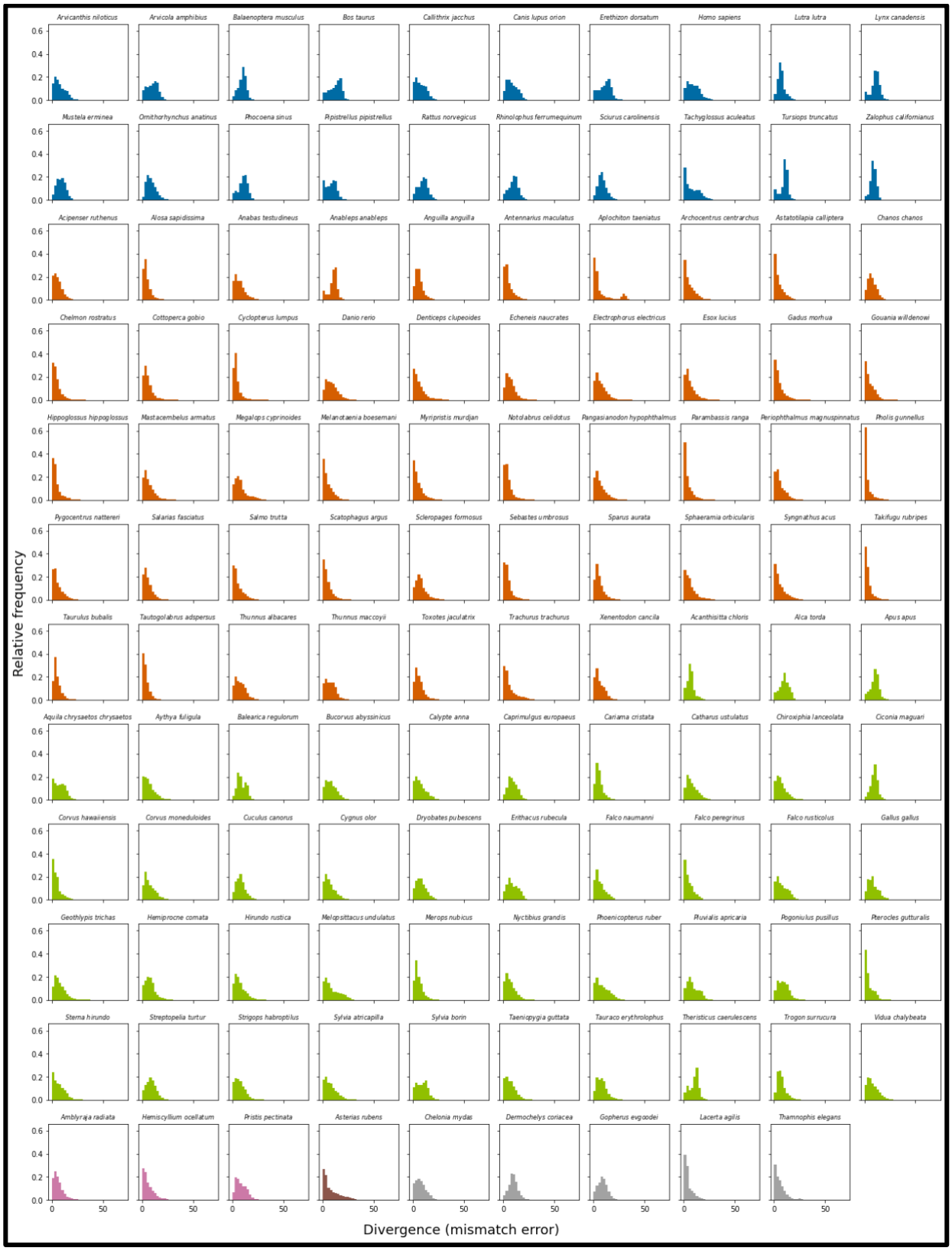

Figure S2. Divergence distributions of segmental duplications. We measure divergence as the mismatch score between partner duplicates. The histograms are color-coded by taxonomic class. We observe a general trend toward

duplicates being more similar and younger. We note that this pattern is less pronounced for mammalian species than other species..

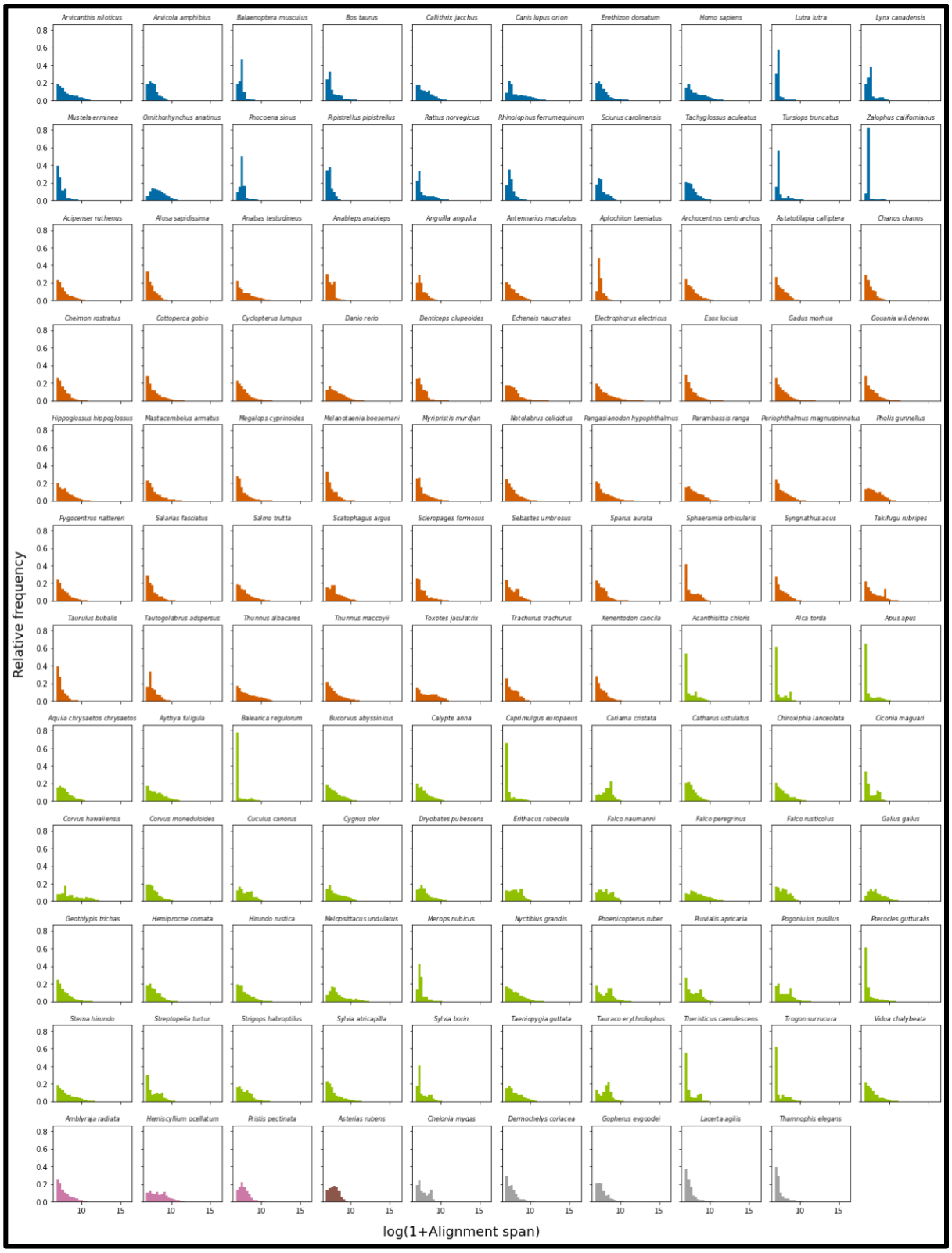

Figure S3. Size distributions of segmental duplications. The size (represented by the natural log of alignment span + 1) distribution of duplicate pairs for each species, color-coded by taxonomic classes. The observed distributions are typically right skewed, indicating a high proportion of shorter duplications with a long tail of large duplications, consistent across most species.

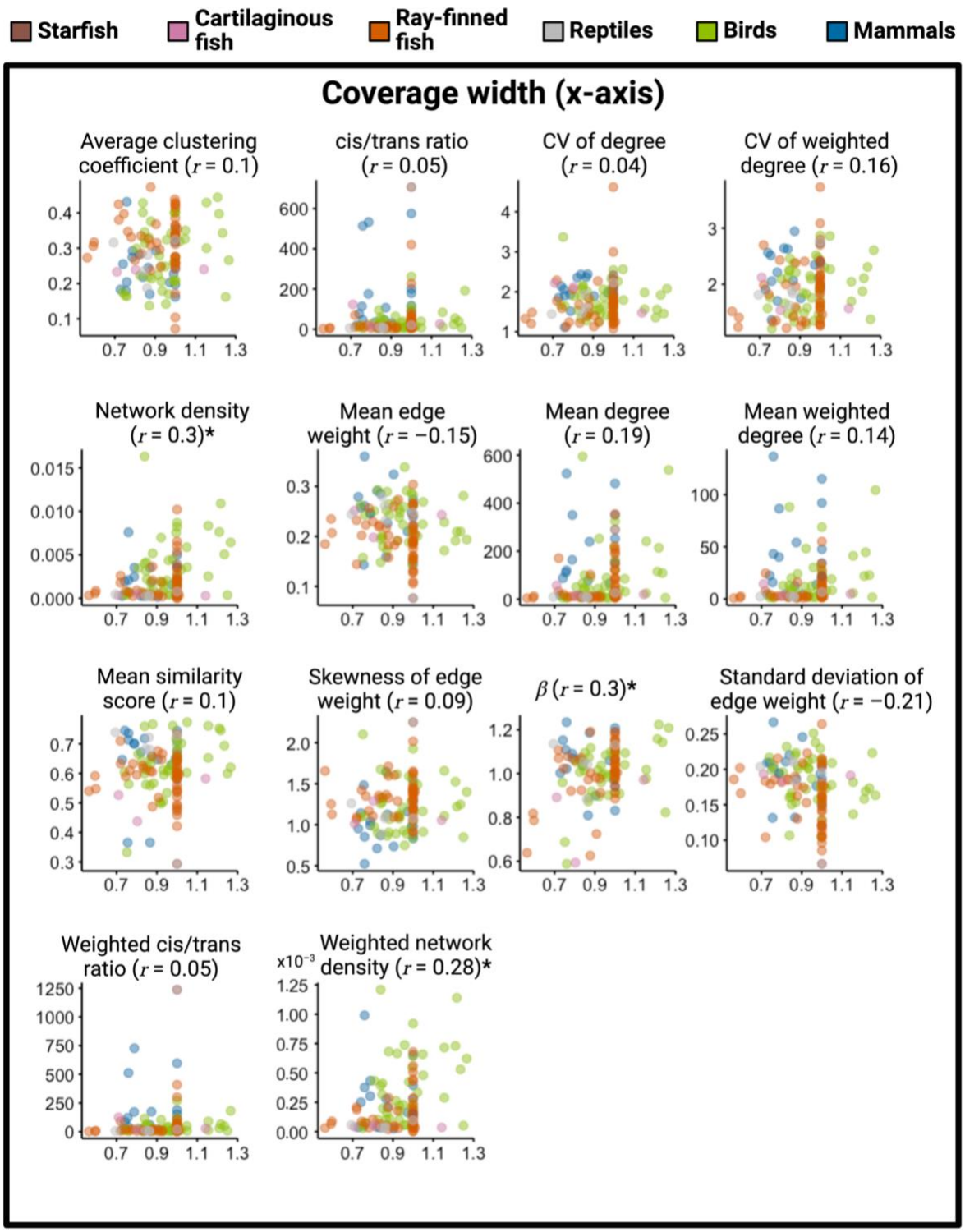

Figure S4. The relationship between coverage width and segmental duplication landscape metrics. The horizontal axis represents the coverage width. The vertical axis represents the segmental duplication landscape property whose name is shown on the top of each figure panel. A dot represents a species, color-coded by taxonomy class shown in the legend. Widths are greater than 1 when the assembled genome is larger than the estimated genome size. The Pearson correlation coefficient is denoted by r. Landscape properties for which the correlation has a raw p-value < 0.01 are appended by *.

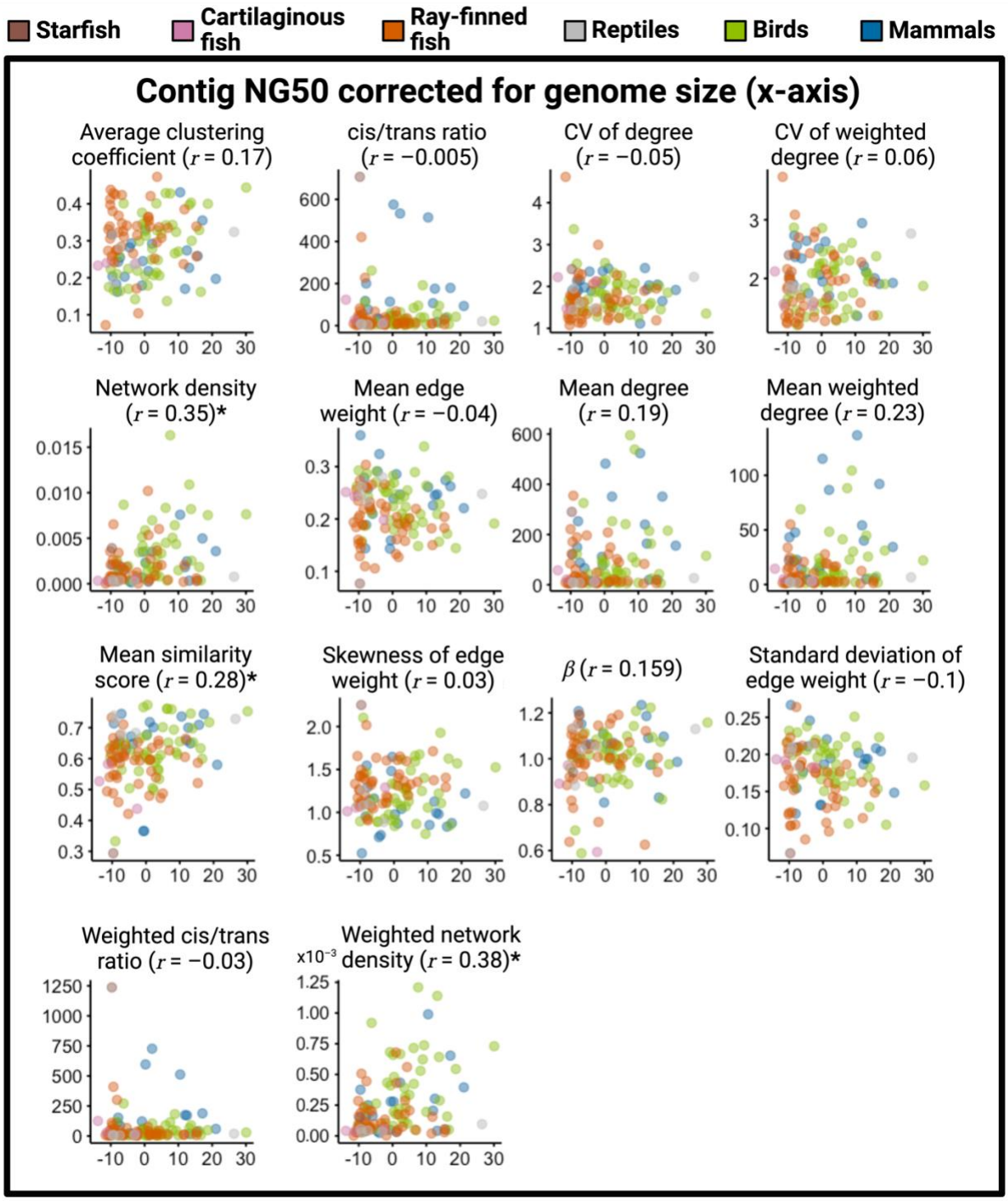

Figure S5. The relationship between contig NG50 corrected for genome size and segmental duplication landscape metrics. The horizontal axis represents the residuals from the linear model. The vertical axis represents the segmental duplication landscape property whose name is shown on the top of each figure panel. A dot represents a species, color-coded by taxonomy class shown in the legend. The Pearson correlation coefficient is denoted by r. Landscape properties for which the correlation has a raw p-value < 0.01 are appended by *.

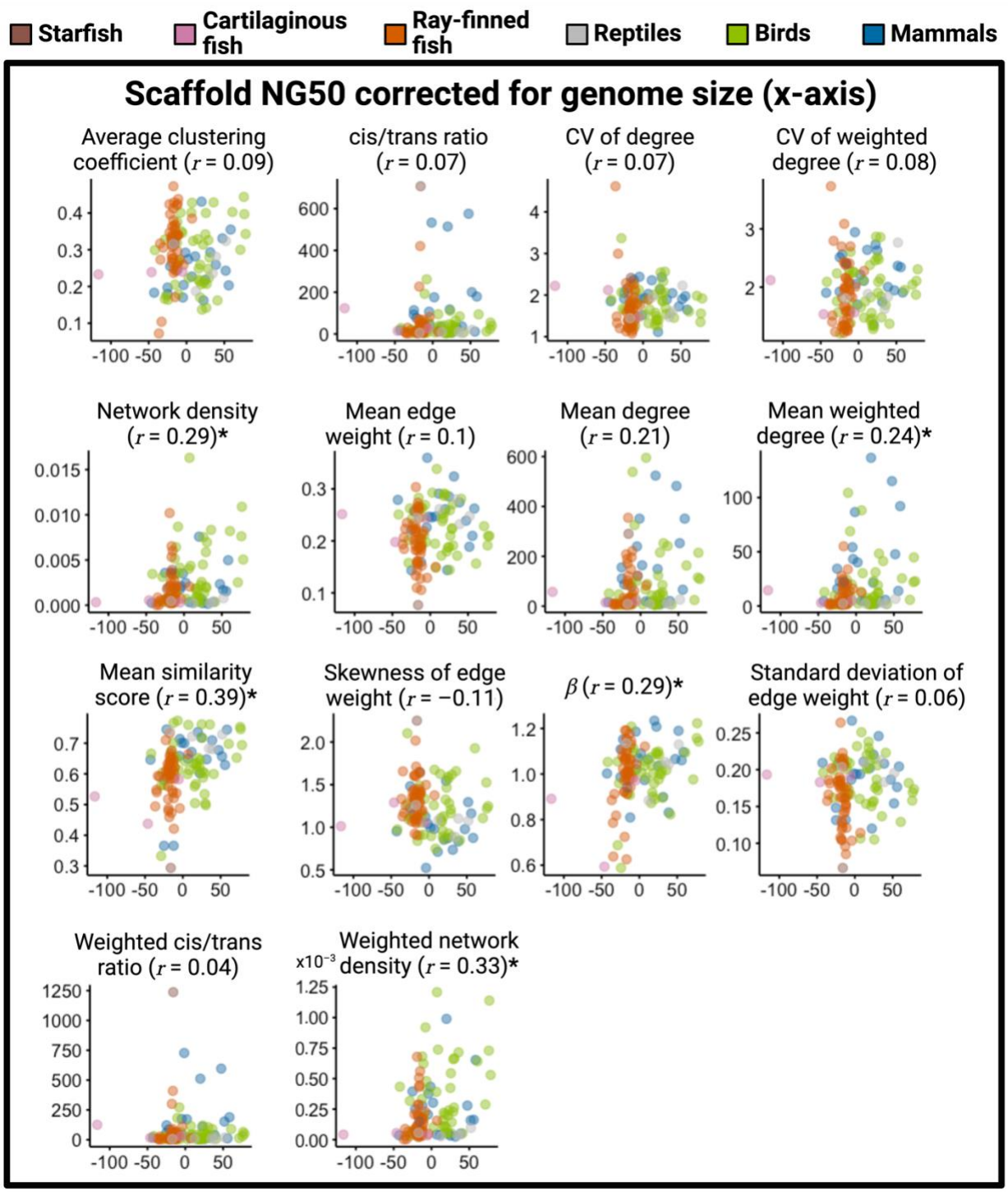

Figure S6. The relationship between scaffold NG50 corrected for genome size and segmental duplication landscape metrics. The horizontal axis represents the residuals from the linear model. The vertical axis represents the segmental duplication landscape property whose name is shown on the top of each figure panel. A dot represents a species, color-coded by taxonomy class shown in the legend. The Pearson correlation coefficient is denoted by r. Landscape properties for which the correlation has a raw p-value < 0.01 are appended by *.

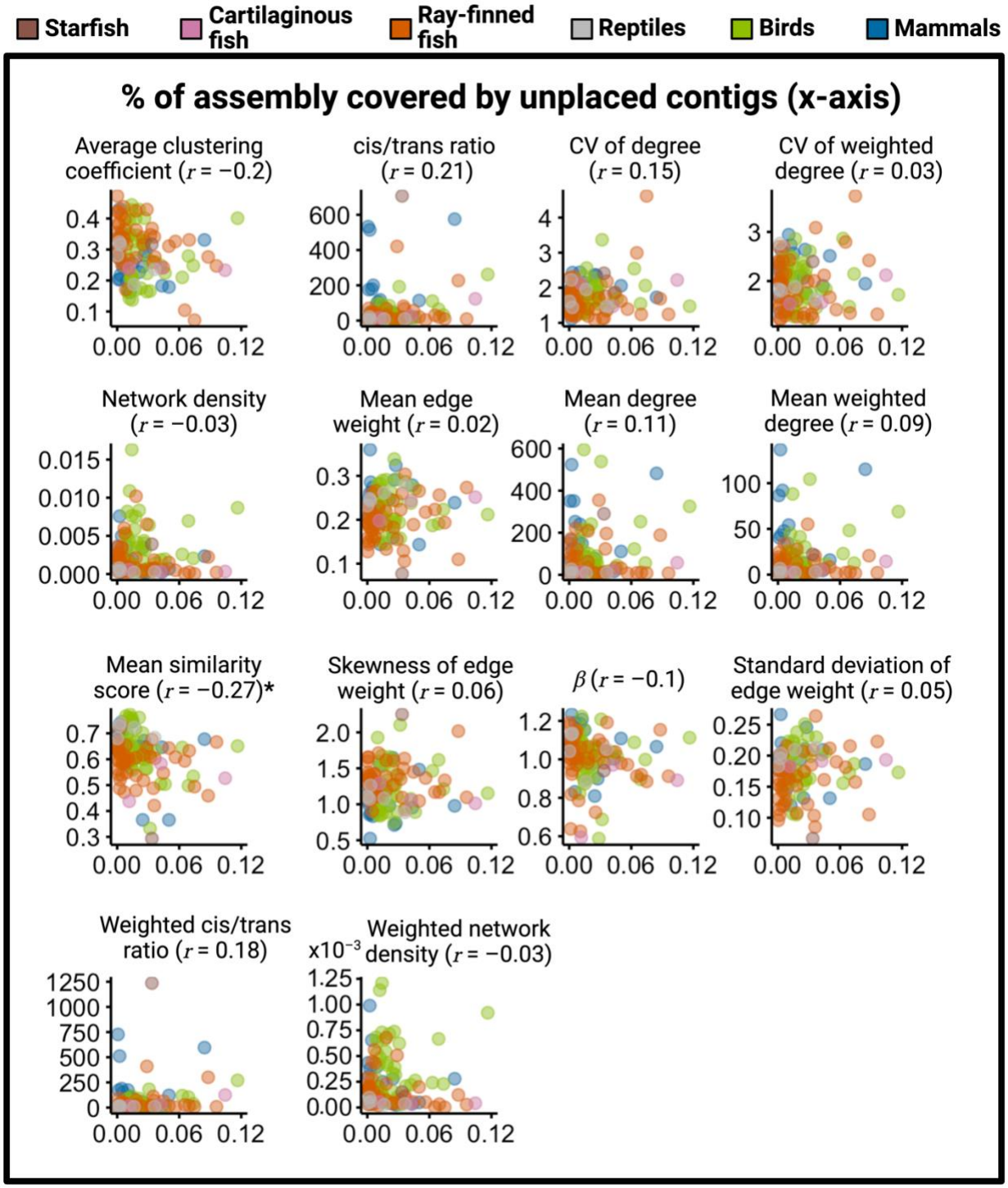

Figure S7. The relationship between the percentage of genome assembly covered by unplaced contigs and segmental duplication landscape metrics. The horizontal axis represents the unplaced contig coverage. The vertical axis represents the segmental duplication landscape property whose name is shown on the top of each figure panel. A dot represents a species, color-coded by taxonomy class shown in the legend. The Pearson correlation coefficient is denoted by r. Landscape properties for which the correlation has a raw p-value < 0.01 are appended by *.

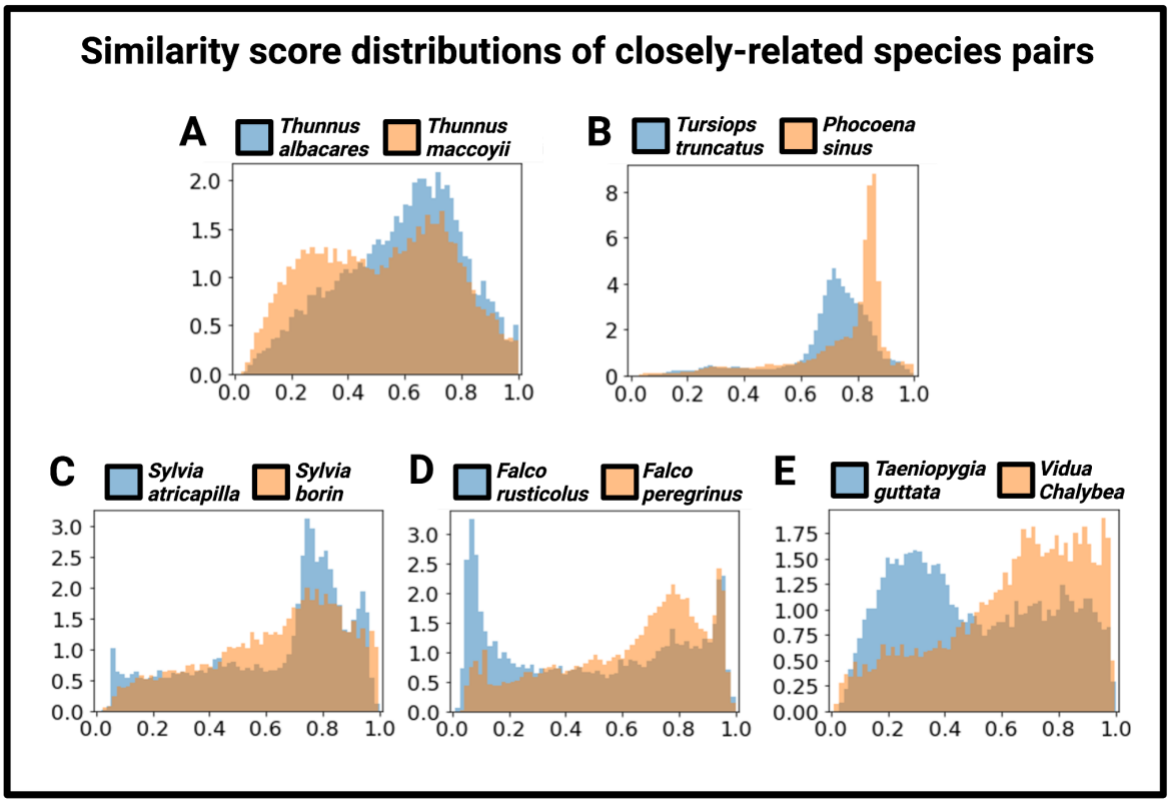

Figure S8. Similarity score distributions for closely-related species pairs. Five pairs of species that are phylogenetically close but have highly variable similarity distributions. The x-axis represents the similarity score and the y-axis represents density. **A.** Ray-finned fish. **B.** Mammals. **C-E.** Birds.

Table S1. CV for each of the 14 measures across species. For example, the "CV of node degree'', shown in the third row of the table, is a scalar measure calculated for each species. Its CV calculated based on all the ray-finned fish species is equal to 0.3427.

	Ray-finned fish	Mammals	Birds	All
Mean degree	1.3511	0.9542	1.4328	1.3952
Mean node strength	1.1969	1.0221	1.2969	1.4603
CV of node degree	0.3427	0.1585	0.2316	0.2689
CV of node strength	0.2945	0.1787	0.2259	0.2499
Network density	1.2066	0.9287	0.9382	1.1542
Weighted density	1.0628	0.9983	0.8133	1.0733
Clustering coefficient	0.2456	0.2653	0.3109	0.2841
Density ratio (cis to trans)	1.9270	1.1943	1.3387	1.9222
Weighted density ratio (cis to trans)	1.8682	1.2938	1.2629	2.2291
β	0.1412	0.1075	0.1186	0.1316
Mean edge weight	0.2236	0.2236	0.1994	0.2316
Std of edge weight	0.2347	0.1927	0.1798	0.2159
Skewness of edge weight	0.1655	0.2667	0.2510	0.2385
Mean similarity score	0.1140	0.1598	0.1363	0.1496

1010 **Table S2. The list of species for which we use a close substitution as proxies in place of the original species.**

Original species name	Substitute species name
Asterias rubens	*Leptasterias muelleri*
Pristis pectinata	*Pristis clavata*
Hemiscyllium ocellatum	*Chiloscyllium punctatum*
Salmo trutta	*Salmo salar*
Cottoperca gobio	*Cottoperca*
Anableps anableps	*Anableps*
Canis lupus orion	*Canis lupus*
Hemiprocne comata	*Hemiprocne*
Ciconia maguari	*Ciconia ciconia*
Phoenicopterus ruber	*Phoenicopterus*
Merops nubicus	*Merops apiaster*
Pogoniulus pusillus	*Pogoniulus*
Aquila chrysaetos chrysaetos	*Aquila chrysaetos*
Gopherus evgoodei	*Gopherus agassizii*
Notolabrus celidotus	*Not substitution found*

1011
1012

Table S3. **The correlation coefficient between the phylogenetic distance between two species and the distance in terms of each of the 14 measures.** The corresponding p-value is shown in the parentheses.

	Ray-finned fish	Mammals	Birds	All
Mean degree	0 (0.9883)	−0.12 (0.900)	0.02 (0.4989)	0.02 (0.0641)
Mean node strength	0 (0.9070)	−0.11 (0.1204)	−0.01 (0.7426)	0.01 (0.5584)
CV of node degree	0.20 (1.0863×10^{-10})	−0.02 (0.8186)	−0.09 (0.0103)	0.07 (1.4593×10^{-08})
CV of node strength	0.06 (0.0368)	−0.10 (0.1842)	0.12 (0.0002)	0.02 (0.0490)
Network density	−0.11 (0.0006)	−0.03 (0.7043)	0.07 (0.0475)	−0.04 (0.0006)
Weighted density	−0.12 (0.0001)	−0.08 (0.2447)	0.01 (0.6822)	−0.03 (0.0085)
Clustering coefficient	0.35 (6.1552×10^{-32})	−0.10 (0.2046)	0.08 (0.0128)	0.04 (0.0029)
Density ratio (cis to trans)	−0.02 (0.5592)	−0.13 (0.0670)	−0.08 (0.0178)	0.18 (6.3287×10^{-54})
Weighted density ratio (cis to trans)	0.01 (0.7203)	−0.14 (0.0629)	−0.08 (0.0149)	0.23 (2.0282×10^{-84})
β	0.06 (0.0708)	0.18 (0.0109)	0.09 (0.0067)	0.04 (0.0005)
Mean edge weight	−0.05 (0.1308)	0.41 (3.5282×10^{-16})	−0.08 (0.0163)	0.12 (1.1336×10^{-23})
Std of edge weight	−0.06 (0.0746)	0.34 (1.1546×10^{-16})	0.02 (0.6459)	0.11 (3.7016×10^{-21})
Skewness of edge weight	0.04 (0.2123)	0.48 (2.9343×10^{-12})	−0.09 (0.0065)	0.12 (5.0953×10^{-25})
Mean similarity score	−0.06 (0.0412)	0.85 (2.2588×10^{-53})	0.01 (0.8447)	0.18 (1.6309×10^{-47})

1013

Table S4. Effect sizes (Cohen's d) and p-values from the Mann-Whitney U test comparing species pairs with phylogenetic divergence less than 100 million years to pairs with greater divergence.

	Ray-finned fish	Mammals	Birds	All
Mean degree	0.0108 (0.6688)	0.0269 (0.6532)	−0.0103 (0.0218)	0.0789 (5.1899×10^{-12})
Mean node strength	−0.0221 (0.7859)	0.0395 (0.4435)	−0.0083 (0.0012)	0.1015 (9.8273×10^{-39})
CV of node degree	0.0501 (0.0681)	0.0351(0.7788)	−0.0009 (0.0088)	−0.1275 (1.5750×10^{-39})
CV of node strength	0.0514 (0.1737)	0.0152 (0.8508)	0.0669 (5.4327×10^{-10})	−0.1084 (1.1182×10^{-19})
Network density	0.0858 (0.7273)	0.0497 (0.4979)	−0.0497 (3.1993×10^{-05})	0.1306 (3.6272×10^{-87})
Weighted density	0.0997 (0.2668)	0.0595 (0.4162)	−0.0548 (1.1159×10^{-06})	0.1713 (5.7651x$\times 10^{-125}$)
Clustering coefficient	−0.0199 (0.3565)	0.0558 (0.4022)	−0.0024 (0.9342)	0.0007 (0.0357)
Density ratio (cis to trans)	-0.0535 (0.9065)	0.0465 (0.6276)	0.0301(0.0646)	−0.0841(0.0738)
Weighted density ratio (cis to trans)	−0.0474 (0.9843)	0.0433 (0.4041)	0.0219(0.2666)	−0.0958 (0.0002)
β	−0.0556 (0.0371)	0.0171 (0.6979)	−0.0312 (0.0536)	−0.0358 (0.0005)
Mean edge weight	0.0679 (0.1105)	−0.0011 (0.7281)	−0.0065 (0.9372)	−0.0456 (0.0002)
Std of edge weight	0.0631 (0.2097)	0.0047 (0.9468)	−0.0012 (0.1727)	−0.1383 (1.6960×10^{-39})
Skewness of edge weight	0.0365 (0.1249)	0.0078 (0.9221)	0.0136 (0.0099)	0.0265 (0.0763)
Mean similarity score	−0.0129 (0.4847)	0.0503 (0.3211)	−0.0412 (7.3731×10^{-05})	−0.0384 (9.3814×10^{-10})

Supplementary Text S1

In (Abdullaev et al. 2021), the authors examined segmental duplication networks of nine species. They found that the component size distribution of the network was more similar within the six mammal species than between the mammal and non-mammal species. We computed the property that they used (i.e., the ordered sequence of the component size of the segmental duplication network) for each of our 118 species and computed the so-called Bray-Curtis dissimilarity (BCD) between each pair of species using the same method as theirs (Abdullaev et al. 2021). Then, as we did for our 14 quantities in the main analysis, we calculated the Pearson correlation coefficient between the phylogenetic distance between pairs of species and the BCD. The correlation was $r = -0.18$ ($p = 5.1905 \times 10^{-09}$), $r = 0.23$ ($p = 0.0015$), $r = 0.05$ ($p = 0.1635$), and $r = 0.14$ ($p = 4.5435 \times 10^{-33}$) for ray-finned fish, mammals, birds, and all species, respectively. Although the *p* values are mostly small due to large sample sizes, r (as a measure of the effect size) is small. This result contrasts with those in (Abdullaev et al. 2021), which suggested that the component size distribution has an evolutionary signal, i.e., it is more conserved within mammals than between mammals and non-mammals. There are at least two possible reasons for the difference between their results and ours. First, we mainly compared within each of the three clades and used as many as 118 species. In contrast, their mammal versus non-mammal comparison is only based on six mammal species and three non-mammal species. Furthermore, one of the three non-mammal species used was *Caenorhabditis elegans*, which is evolutionarily farther than among the three clades we used. Second, the two studies used different computational methods to construct segmental duplication networks. Across-species comparison of segmental duplication networks aiming to improve the accuracy of segmental duplication calls and explicitly considering gene orthologs warrant future work.