Evolvability through segmental duplication follows species-specific dynamics

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

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

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

† *Contributed equally*
* *Correspondence*

Abstract
Evolvability, the capacity of populations to respond to selection, is influenced by genomic structural properties such as segmental duplications. These duplications establish redundancy, creating adaptive potential by co-opting duplicated elements for new functions. Here, we identify segmental duplications in 117 vertebrate species and the starfish outgroup using long-read genome assemblies. We constructed independent segmental duplication networks for each species to represent the landscape of segmental duplications therein. We quantified the duplication landscapes across species using 13 network properties and tested three hypotheses for their evolution: 1) selective constraint, 2) phylogenetic drift, and 3) species-specific dynamics. Our results show that segmental duplication profiles across vertebrates are shaped by fast-evolving, species-specific dynamics, driving unique adaptive potentials. One notable exception of lineage-specific evolution is observed within mammals, where monotremes (platypus and echidna) exhibit distinct genomic structural features, characterized by elevated insertion and deletion rates compared to therian mammals. Analysis of genome size variation in vertebrates revealed no significant correlation with segmental duplication metrics. However, in ray-finned fish, lineage-specific events such as whole-genome duplications in some species (e.g., sterlet sturgeon and brown trout) influence genome size. These findings highlight the role of segmental duplications in species-specific adaptive potential and demonstrate the utility of network-based methods in genomic research. Our dataset and framework provide a valuable resource for understanding genomic evolvability, underscoring the integration of evolutionary biology and network analysis in studying genome evolution.

Introduction

In 1988, Dawkins noted that heritable organismal properties, such as bilateral symmetry and segmented body plan, might influence the propensity of populations to respond to selection, a concept termed “evolvability” (Dawkins 1988). He argued that these properties themselves are subject to evolution. One such property at the genome level is the amount of standing genetic variation in a population (Houle 1992; Tenaillon and Matic 2020; Hansen and Pélabon 2021), maintained either by drift or balancing selection (Aqil et al. 2023; Bitarello et al. 2023). For example, a previously neutral allele in a genetic polymorphism may confer an advantage if the environmental conditions change (Lewontin 1987; Hermisson and Pennings 2005). Greater standing variation makes it more likely that one or more segregating sites in the population will harbor an allele that is beneficial in the new environment. Additionally, the processes generating new genetic variation, such as the mutation and recombination rates, also influence evolvability. These processes regulate the introduction of new genetic material, affecting how effectively a population can respond to new environmental challenges.

Despite the conceptual maturation of evolvability (Nuño de la Rosa 2017; Villegas et al. 2021; Nuño de la Rosa 2023), empirical research has lagged behind (Hansen and Wagner 2023). There is a particular dearth of research on evolvability driven by genomic structural properties, such as segmental duplications. We refer to this kind of evolvability as “genomic structural evolvability.” We argue that segmental duplications play a crucial role in this mode of evolvability. In particular, segmental duplications often result in multiple copies of functional elements, such as genes and regulatory regions, creating redundancy (Ohno 1970). This redundancy creates potential for evolution, as mutations in these duplicated regions can lead to their co-option for new functions (Innan and Kondrashov 2010), increasing the likelihood of a population adapting to new selective pressures in changing environments. Indeed, the mutation rate is elevated in segmental duplicates relative to unique regions (Vollger et al. 2023), making it more likely that key mutations leading to novel functions for the duplicates will be acquired. Additionally, genomic regions rich in segmental duplication are prone to accrue even more duplications over time, especially through an increased rate of recombination errors (Kim et al. 2008; Levasseur and Pontarotti 2011). Collectively, we argue that the landscape of segmental duplications is a good indicator of genomic structural evolvability, and thus, we use the two terms interchangeably. Overall, studying the evolution of the segmental duplication landscape is imperative to understanding differential rates of evolution across species.

In this article, to gain insights into genomic structural evolvability, we study segmental duplication networks. First, we generated segmental duplication calls based on long-read genome assemblies of 117 vertebrate species and the starfish outgroup **(Figure 1A)**. For each species, we constructed a weighted undirected segmental duplication network, where nodes represent genomic regions and edges between two nodes indicate that they are duplicates. The weights on the edges reflect sequence similarity between duplicates **(Figure 1B-G; see Methods)**. We calculated 13 network properties, as well as the average similarity score (a non-network measure) across duplications for each species. We used these 14 measures collectively to quantify the segmental duplication landscapes across species. Previous work has shown that models involving gene duplication followed by divergence between duplicates yield realistic properties of gene regulation, protein interaction, and segmental duplication networks. One such property is the scale-free property (i.e., some genes are hubs with a large number of connections, whereas the majority of other genes have fewer connections) (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). Such biological networks have been exploited to understand gene function, regulation pathways, and disease mechanisms (Barabási and Oltvai 2004; Barabási et

al. 2010). 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), analyses of empirical segmental duplication data remain sparse (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. Our network-based analysis of segmental duplications not only fills this gap but also provides a framework for understanding genomic evolvability across vertebrates.

Here, we test three competing hypotheses for the evolution of genomic structural evolvability. The first hypothesis is the “selective constraint hypothesis”: the segmental duplication landscape shows little variation among vertebrates, suggesting strong selective constraints on genomic structural evolvability. However, this is unlikely, as selection operates on traits conferring immediate advantage, whereas evolvability is, by definition, the potential to evolve (Gould 2002). The second is the “phylogenetic drift hypothesis”: the segmental duplication landscape has drifted gradually since the more recent common ancestor of vertebrates. If this is true, the segmental duplication landscape would align with vertebrate phylogeny. This hypothesis is unlikely to hold, as the rate of segmental duplications depends on the existing segmental duplication landscape in a lineage (Kim et al. 2008; Levasseur and Pontarotti 2011). This contingency suggests the segmental duplication landscape will not coincide with phylogeny. The third hypothesis is the “species-specific dynamics hypothesis”: the segmental duplication landscape is highly variable and does not coincide with phylogeny. This would indicate that the segmental duplication landscape in vertebrates follows species-specific dynamics. This implies that each species harbors a unique potential to respond to selection based on its specific segmental duplication profile.

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 critical gaps in understanding how the segmental duplication landscape evolves and drives species-specific adaptive potential.

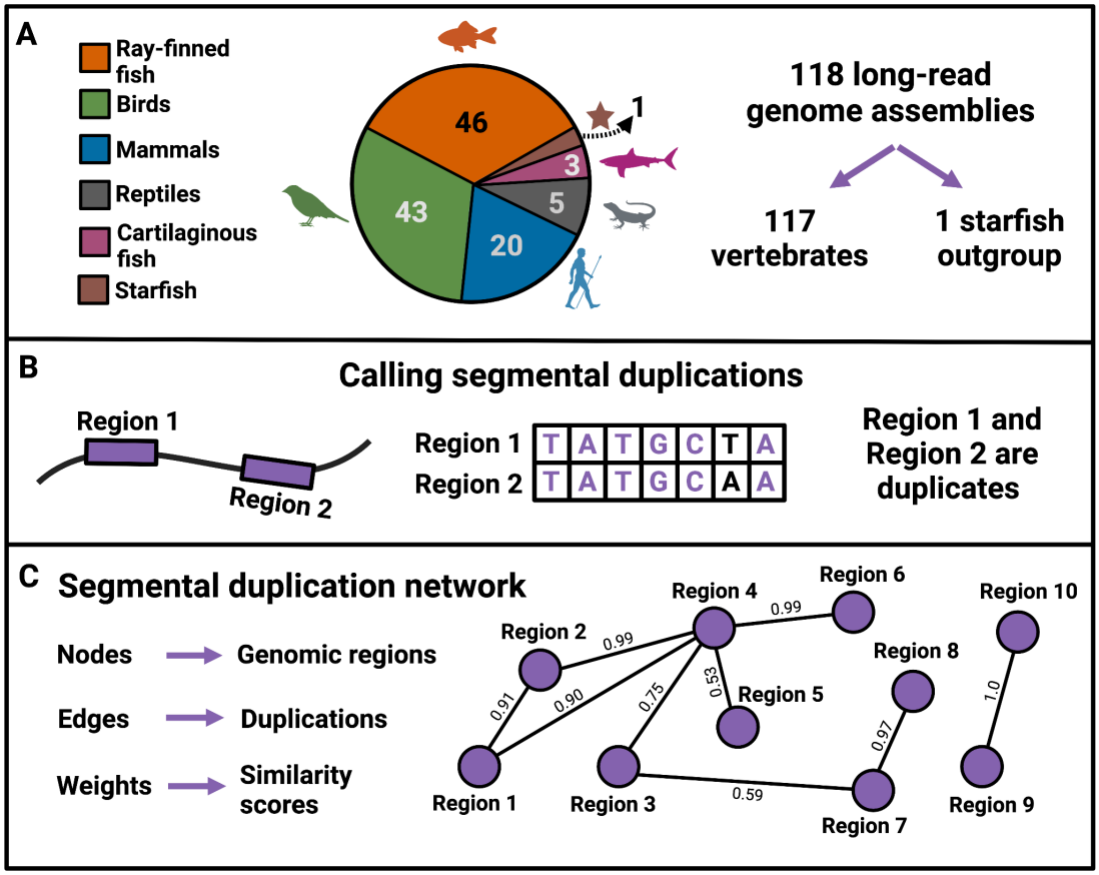

Figure 1. Methodological framework. A. The number of species in each taxonomic class is in our data. **B.** A schematic showing segmental duplication calling from genome sequences. The image on the left represents a chromosome with two genomic regions marked as Region 1 and Region 2. The middle image 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

Genomic structural evolvability follows species-specific dynamics

We conducted comparative analyses of segmental duplication networks across 118 genomes to investigate how genomic structural evolvability varies across vertebrate species. Specifically, we tested three hypotheses introduced earlier: (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 varies in a unique, species-specific manner **(Figure 2A)**.

To quantify the segmental duplication landscapes, we constructed a network for each species independently, where nodes represent genomic regions and edges signify duplications. Edge weights indicate similarity scores, representing how recently the duplication happened. 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 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; Table S2)** (Kumar et al. 2022). If most of these 14 measures meaningfully correlate with phylogeny, we favor the phylogenetic drift hypothesis; otherwise, we favor the species-specific dynamics hypothesis. Following Cohen's guidelines of Pearson's r (*Cohen 1988*), we considered a measure to be meaningfully correlated with phylogeny if $r > 0.3$. We found that none of the 14 measures of segmental duplication landscape meaningfully correlates with phylogeny across vertebrates (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 2B**. 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 S1**).

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)**.

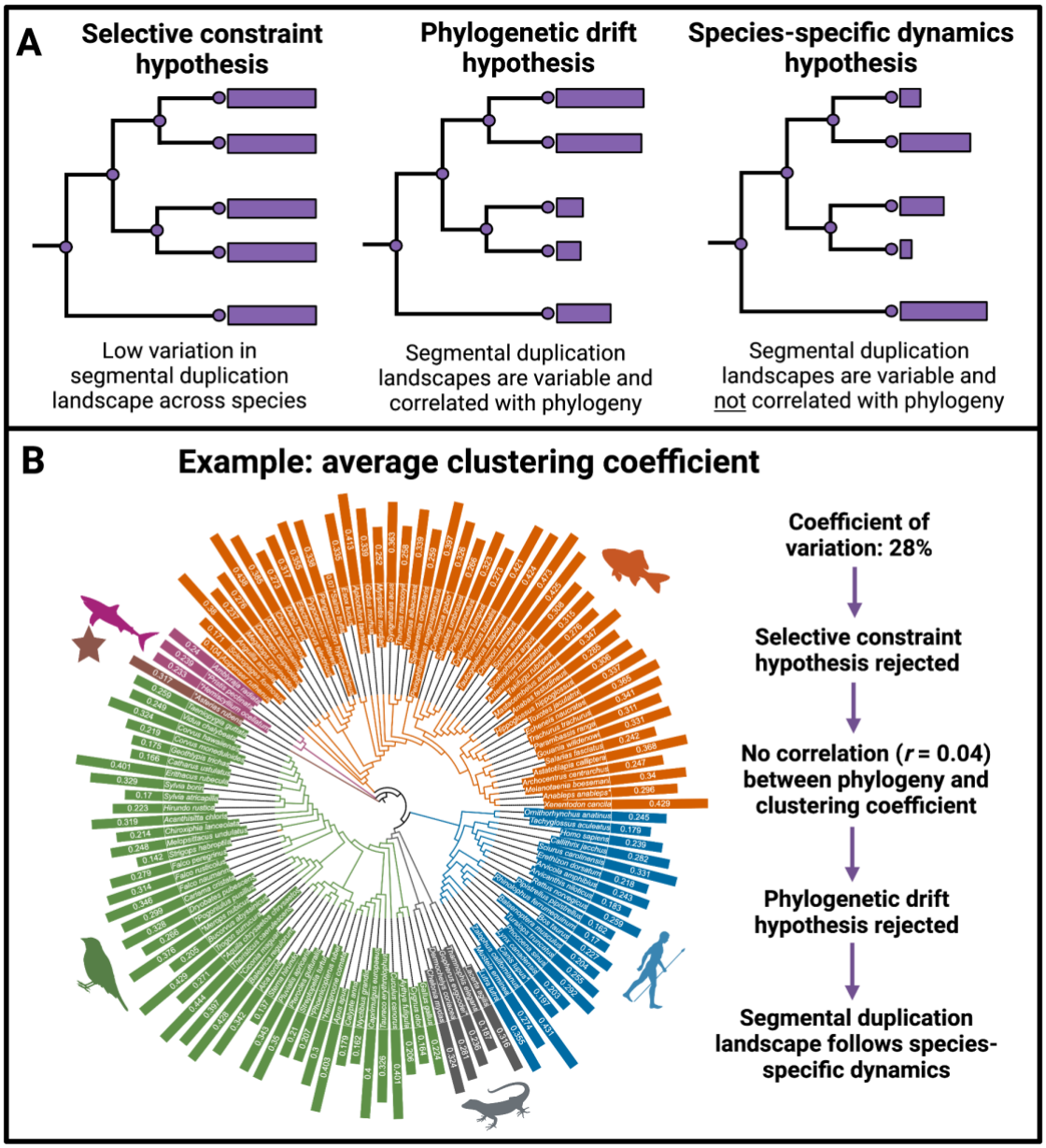

Figure 2. Hypotheses for the evolution of segmental duplication landscape A. Expectation for variation in segmental duplication landscape under each hypothesis. The leaves in the trees in each schematic represent species. The length of the purple bar represents the value of a hypothetical measure of the segmental duplication for the corresponding species. The figure shows the expectation for the 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 S1**). 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 3A-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 3C)**, 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 3D-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. This finding implies that monotreme genomes evolve through distinct structural processes compared to therian mammals, with a greater tendency to accumulate insertions and deletions.

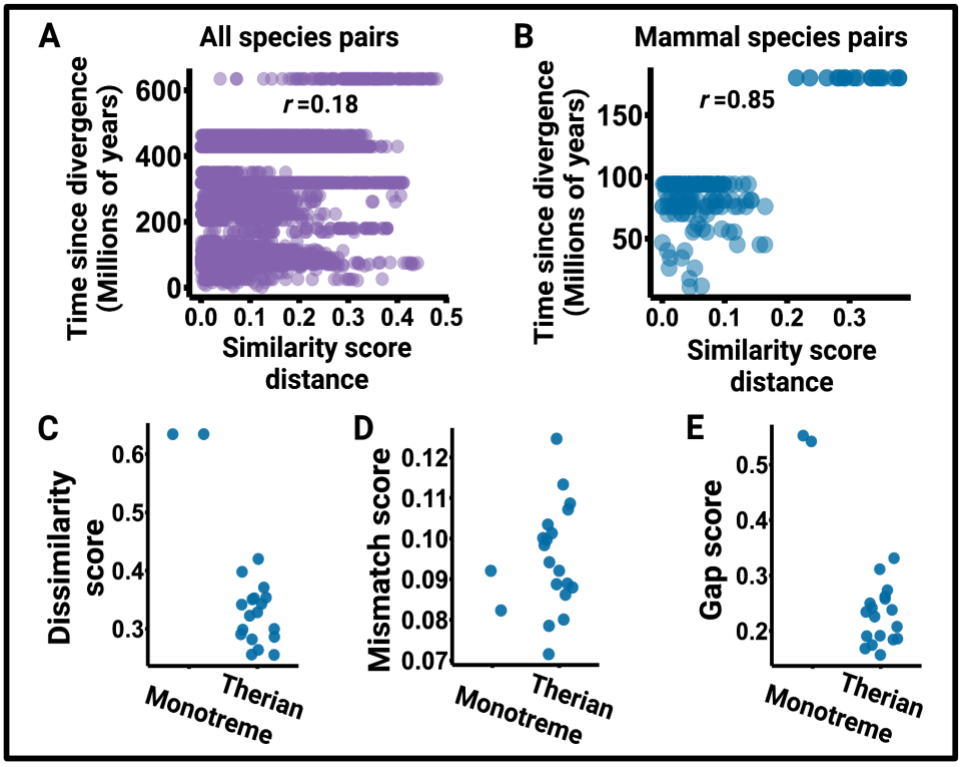

Figure 3. Insights from the average similarity scores for segmental duplications across species. A. Phylogenetic distances are not meaningfully correlated with the average similarity score distances across pairs of species. **B.** Phylogenetic distances are strongly correlated with average similarity 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.

Segmental duplication landscape does not predict genome size

Ancient whole genome duplications (WGD), followed by differential retention of duplicated material, have been proposed to contribute to present-day genome size variation across the tree of life (Bennett and Leitch 2005; Gregory 2005; Kraaijeveld 2010). Additionally, true (localized) segmental duplications also increase genome size, by definition. Since both localized duplication and remnants of ancient whole genome duplications are captured by our segmental duplication calls, we asked whether segmental duplications can explain genome size variation among vertebrates. We measured three aspects of segmental duplications for each species: 1) the number of duplications, 2) mean duplication size, and 3) the fraction of the genome they affect. We found no significant correlation between genome size and any of these measures **(Figures 4A-C)**. Based on these results, we conclude that neither ancient whole genome duplications nor true segmental duplications are primary contributors to genome size variation in vertebrates. Instead, this conclusion aligns with previous studies identifying transposable elements, as opposed to segmental duplications, as major determinants of genome size in vertebrates (Kidwell 2002; Sun et al. 2012; Elliott and Gregory 2015; Gao et al. 2016; Kapusta et al. 2017; Sotero-Caio et al. 2017; Biscotti et al. 2019), and mirrors similar findings in plants (Lee and Kim 2014).

Two ray-finned fish species, the sterlet sturgeon and the brown trout, stand out as exceptions, with unusually large genomes and high fractions of their genomes covered by segmental duplications compared to other ray-finned fish in our dataset **(Figure 4D)**. 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 strong segmental duplication signal associated with 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 as detectable 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 segmental duplications.

As a byproduct of our primary analysis on genome size variation, we also reproduced previous findings (Hughes and Hughes 1995; Gregory et al. 2009; Kapusta et al. 2017) that bird genomes are significantly smaller than those of other amniotes (Cohen's $D = 4.5; p < 2.2 \times 10^{-16}$; **Figure 4E**). This reduction in genome size correlates with smaller red blood cells, which may have evolved to meet the high metabolic demands of flight (Szarski 1983; Hughes and Hughes 1995; Gregory 2002; Andrews et al. 2009; Wright et al. 2014). That the smallest mammalian genome in our dataset (1.76 billion base pairs) belongs to the common pipistrelle microbat (*Pipistrellus pipistrellus*) supports previous studies suggesting that flight exerts selective pressure for genome reduction in bats, too (Smith and Gregory 2009). This observation is also consistent with a previous finding of molecular convergence between bats and birds to adapt to flight (Matsuda and Makino 2024).

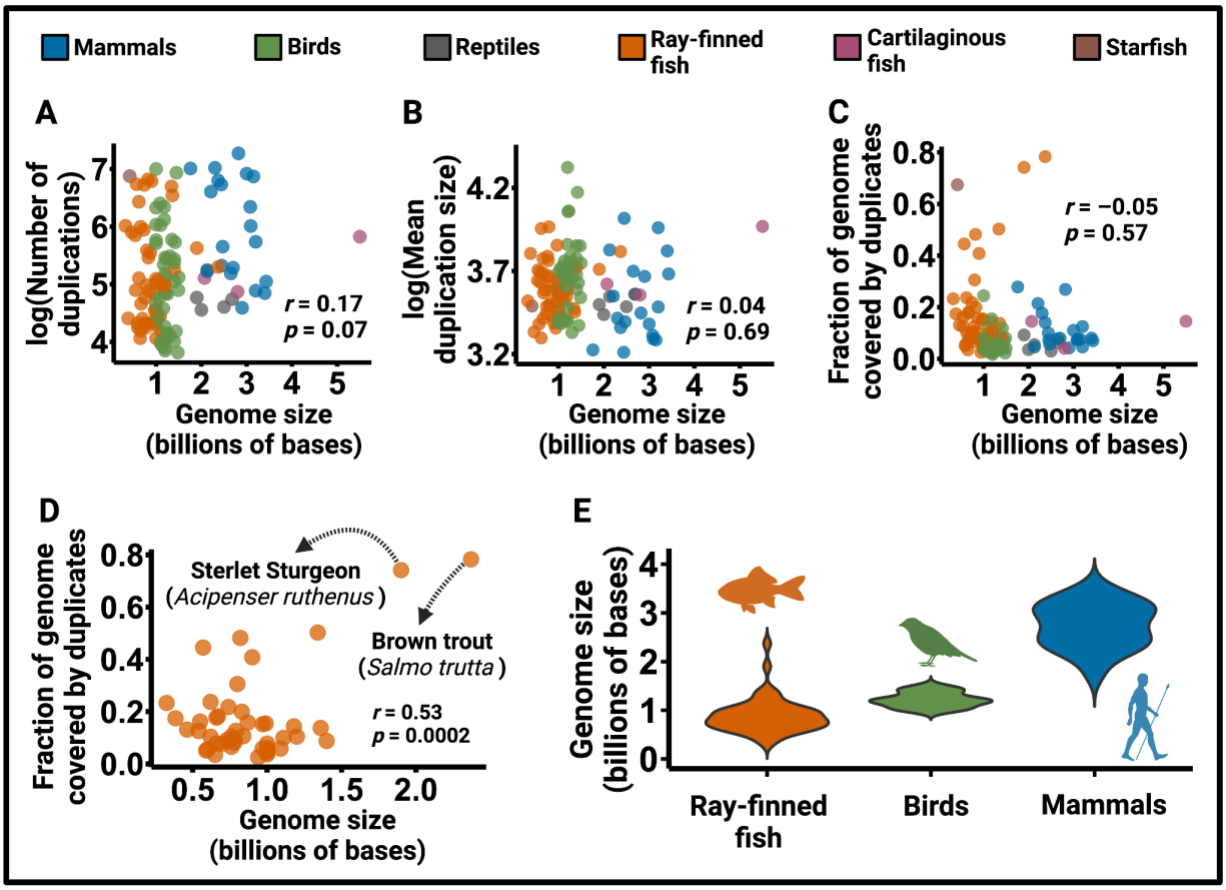

Figure 4. Genome size variation across vertebrates and the starfish outgroup. A-C. Genome size variation among vertebrates cannot be explained by the number of duplications, the size of duplications, or the fraction of genomes covered by segmental duplications. Each point in these figures represents a species. **D.** 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. **E.** Distribution of genome sizes among ray-finned fish, birds, and mammals.

Discussion

In this study, we used properties derived from segmental duplication networks across 118 species to quantify and analyze genomic structural evolvability across vertebrates. In particular, we tested three hypotheses concerning the evolution of segmental duplication landscapes: (1) the selective constraint hypothesis, (2) the phylogenetic drift hypothesis, and (3) the species-specific dynamics hypothesis. Our analysis reveals that segmental duplication profiles are predominantly shaped by fast-evolving species-specific dynamics that obscure remnants of phylogenetic signals from shared ancestry. Rapid divergence in 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.

Interestingly, our study uncovered a lineage-specific pattern within mammals. In particular, monotremes have lower average similarity scores, driven by more alignment gaps between duplicates, than therian mammals. Our results indicate this is likely due to a higher propensity among monotreme genomes to accrue deletions and insertions. This exception provides an exciting avenue for future research into the mechanisms driving the high rates of deletions and insertions in monotremes.

We also found no significant relationship between segmental duplication and genome size variation. Instead, genome size appears to be driven primarily by other factors, such as transposable elements. Nonetheless, there are exceptions specific to ray-finned fishes (the sterlet sturgeon and brown trout) that harbor large genomes disproportionately influenced by segmental duplications. These species exemplify how remnants from whole-genome duplications, coupled with lineage-specific mechanisms like slow molecular evolution or tetrasomic inheritance, can amplify the apparent contribution of segmental duplications to genome size.

One of the key contributions of this study is the generation of segmental duplication calls based on long-read assemblies from 118 species across six taxonomic classes. By enabling large-scale, cross-species analyses of duplication dynamics, this dataset offers a valuable resource to researchers exploring the evolution of segmental duplications both across vertebrates and within specific clades.

A significant 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 insights into the structural organization of genomes. 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 assembles from GenomeArk (https://www.genomeark.org/), which holds genomes sequenced by the vertebrate genome project (as of August 16th, 2022). Before looking for segmental duplications (SDs)—duplications of size 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 will be called as SDs despite not being biologically so. Indeed, if these repeats are not masked, short tandem repeats may masquerade as segmental duplications. We used RepeatMasker (v4.1.3) for this purpose. Since masking was inadequate for many non-mammalian species with the default masking library (Dfam minimal), we used the complete Dfam 3.6 library augmented with RepBase (v10/26/2018) for better masking performance. 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% of the genome (*Asteroidea*) to 82% (*Bal. musculus*).

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—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.

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 5A)**. 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 5B)**. 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 5B**, 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 5A-B**. We show a part of the network of *Homo Sapiens* in **Figure 5C**. 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 5D**, **E**, and **F**, respectively.

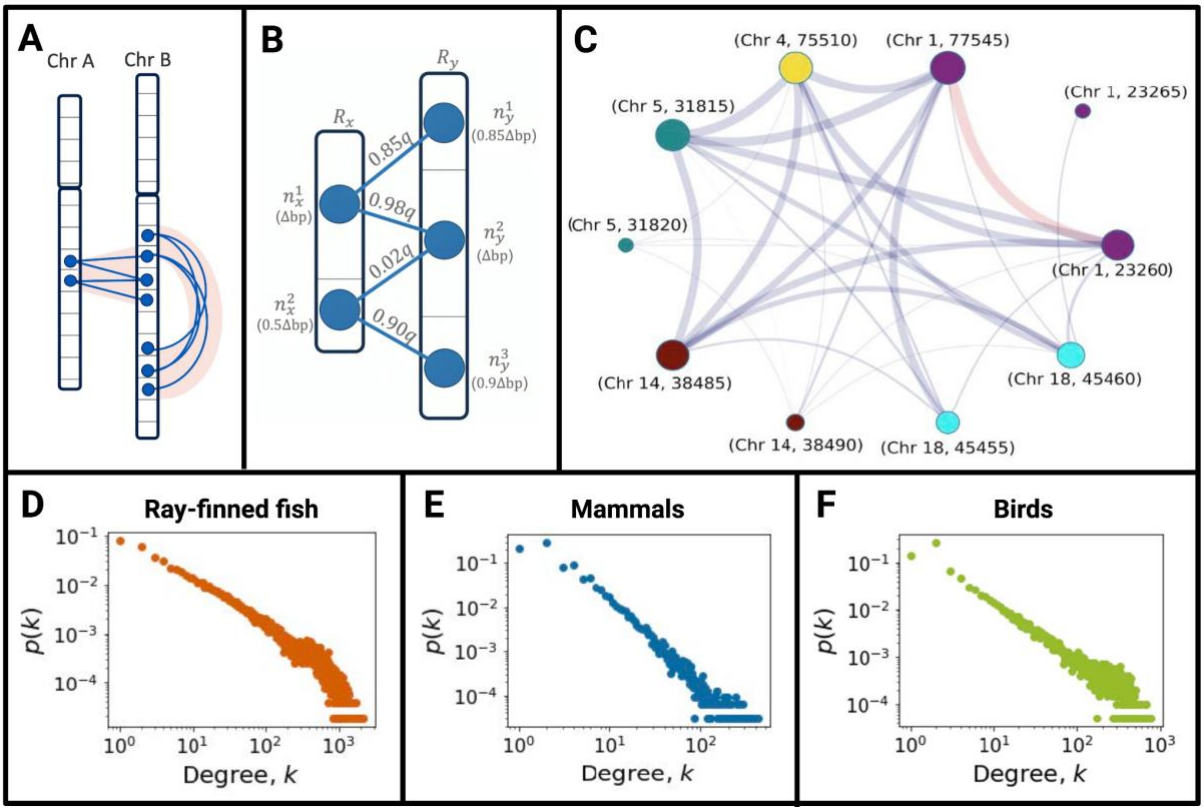

Figure 5. Network construction and degree distribution. A. Schematic for edge construction between two duplicated regions. **B.** Examples of edges for two chromosomes. The part of the network shown here is a magnification of edges between Chr A and Chr B in panel **A**. **C.** A part of the weighted 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 network by $A = (a_{ij})$ and the weight of the edge between the ith and jth nodes by 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 $a_{ij} = w_{ij} = 0$. 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 edges that a node has. 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 mean degree is given by

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

and represents the average number of duplications per node. The mean node strength is given by

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

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 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). 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). The 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

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

In addition to these properties, we also measure the mean, CV, and skewness of the edge weight, w_{ij} (Watts and Strogatz 1998). The skewness for the edge weight is given by

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

where σ_w is the sample standard deviation for the edge weight.

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.

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):

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

where k_u is the unique degrees and $\bar{s}_u$ is the node strength averaged over all the nodes whose degree is k_u. To estimate the parameters, we fitted the a line to the set of points $\{(\ln k_u, \ln\bar{s}_u)\}$ using sklearn (Pedregosa et al. 2011) in Python. Because Eq. (13) implies $\bar{s}_u$ is proportional to k_u^{β}, if $\beta > 1$, hubs, i.e., large-degree nodes, tend to have edges with large weights. If $\beta = 1$, there is no such correlation between the node's degree and edge weight. Various empirical weighted networks yield $\beta > 1$ (Barrat et al. 2004).

In addition to the 13 network measures (mean and CV of the degree distribution, mean and CV of the node strength distribution, network density, weighted network density, average clustering coefficient, ratios of densities on the same chromosome to other chromosomes for both weighted and unweighted networks, β, mean, CV, and skewness of the edge weight distributions), we also computed the average similarity score across duplications for each species. 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

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

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. 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. *bioRxiv* 2023.03.18.533287.

Abdullaev ET, Umarova IR, Arndt PF. 2021. Modelling segmental duplications in the human genome. *BMC Genomics* 22:496.

Alexandrou MA, Swartz BA, Matzke NJ, Oakley TH. 2013. Genome duplication and multiple evolutionary origins of complex migratory behavior in Salmonidae. *Mol. Phylogenet. Evol.* 69:514–523.

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.* 106:217–227.

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.* 70:1403–1404.

Andrews CB, Mackenzie SA, Gregory TR. 2009. Genome size and wing parameters in passerine birds. *Proc. Biol. Sci.* 276:55–61.

Aqil A, Speidel L, Pavlidis P, Gokcumen O. 2023. Balancing selection on genomic deletion polymorphisms in humans. *eLife* 12:e79111.

Bailey JA, Eichler EE. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat. Rev. Genet.* 7:552–564.

Barabási A, Gulbahce N, Loscalzo J. 2010. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* 12:56–68.

Barabási A-L, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5:101–113.

Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A. 2004. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. U.S.A.* 101:3747–3752.

Bennett MD, Leitch IJ. 2005. Genome Size Evolution in Plants. In: Gregory TR, editor. The Evolution of the Genome. Elsevier. p. 89–162: Burlington, MA.

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.* 7:R112.

Biscotti MA, Carducci F, Olmo E, Canapa A. 2019. Vertebrate genome size and the impact of transposable elements in genome evolution. In: Evolution, Origin of Life, Concepts and Methods. Cham: Springer International Publishing. p. 233–251.

Bitarello BD, Brandt DYC, Meyer D, Andrés AM. 2023. Inferring balancing selection from genome-scale data. *Genome Biol. Evol.* 15:evad032.

Bi X, Wang K, Yang L, Pan H, Jiang H, 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* 184(5):1377–1391.

Broder AZ. 1997. On the resemblance and containment of documents. In: Proceedings. Compression and Complexity of Sequences 1997 (Cat. No.97TB100171). IEEE Comput. Soc. p. 21–29.

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)* 9:2017–2028.

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.* 10:776.

Cohen J. 1988. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Routledge: London, England.

Dawkins R. 1988. The evolution of evolvability. In: Langton CG, editor. Artificial life: the proceedings of an interdisciplinary workshop on the synthesis and simulation of living systems, held September 1987 in Los Alamos, New Mexico. Proceedings volume in the Santa Fe Institute studies in the sciences of complexity. Vol. 6. Redwood City, CA: Addison-Wesley; 1989. p. 201–220.

Du K, Stöck M, Kneitz S, Klopp C, Woltering JM, Adolfi MC, Feron R, Prokopov D, Makunin A, Kichigin I, et al. 2020. The sterlet sturgeon genome sequence and the mechanisms of

segmental rediploidization. *Nat. Ecol. Evol.* 4:841–852.

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.* 370:20140331.

Gao B, Shen D, Xue S, Chen C, Cui H, Song C. 2016. The contribution of transposable elements to size variations between four teleost genomes. *Mob. DNA* 7:4.

Gould SJ. 2002. The Structure of Evolutionary Theory. Harvard University Press: Cambridge, MA.

Gregory TR. 2002. A bird's-eye view of the C-value enigma: genome size, cell size, and metabolic rate in the class Aves. *Evolution* 56:121–130.

Gregory TR. 2005. Genome Size Evolution in Animals. In: Gregory TR, editor. The Evolution of the Genome. Elsevier. p. 3–87: Burlington, MA.

Gregory TR, Andrews CB, McGuire JA, Witt CC. 2009. The smallest avian genomes are found in hummingbirds. *Proc. Biol. Sci.* 276:3753–3757.

Hagberg AA, Schult DA, Swart PJ. 2008. Exploring network structure, dynamics, and function using Networks. In: Varoquaux G, Vaught T, Millman J, editors. *Proceedings of the 7th Python in Science Conference.* Pasadena, CA, USA. p. 11–15.

Hansen TF, Pélabon C. 2021. Evolvability: A quantitative-genetics perspective. *Annu. Rev. Ecol. Evol. Syst.* 52:153–175.

Hansen TF, Wagner GP. 2023. The evolution of evolvability. The evolution of evolvability. In T. F. Hansen, D. Houle, N. Pavličev, & C. Pélabon (Eds.), Evolvability: A Unifying Concept in Evolutionary Biology? The MIT Press. p. 121–146: Cambridge, MA.

Hermisson J, Pennings PS. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation: Molecular population genetics of adaptation from standing genetic variation. *Genetics* 169:2335–2352.

Houle D. 1992. Comparing evolvability and variability of quantitative traits. *Genetics* 130:195–204.

Hughes AL, Hughes MK. 1995. Small genomes for better flyers. *Nature* 377:391.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* 11:97–108.

Išerić H, Alkan C, Hach F, Numanagić I. 2022. Fast characterization of segmental duplication structure in multiple genome assemblies. *Algorithms Mol. Biol.* 17:4.

Jain C, Rhie A, Zhang H, Chu C, Walenz BP, Koren S, Phillippy AM. 2020. Weighted minimizer sampling improves long read mapping. *Bioinformatics* 36:i111–i118.

Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, Pevzner PA, Eichler EE. 2007. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat. Genet.* 39:1361–1368.

Kapusta A, Suh A, Feschotte C. 2017. Dynamics of genome size evolution in birds and mammals. *Proc. Natl. Acad. Sci. U.S.A.* 114:E1460–E1469.

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

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.* 18:1865–1874.

Kraaijeveld K. 2010. Genome size and species diversification. *Evol. Biol.* 37:227–233.

Krieger J, Fuerst PA. 2002. Evidence for a slowed rate of molecular evolution in the order acipenseriformes. *Mol. Biol. Evol.* 19:891–897.

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* 7:2161–2164.

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.* 39(8):msac174.

Lee S-I, Kim N-S. 2014. Transposable elements and genome size variations in plants. *Genomics Inform.* 12:87–97.

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

Levasseur A, Pontarotti P. 2011. The role of duplications in the evolution of genomes highlights the need for evolutionary-based approaches in comparative genomics. *Biol. Direct* 6:11.

Lewontin RC. 1987. Polymorphism and heterosis: Old wine in new bottles and vice versa. *J. Hist. Biol.* 20:337–349.

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

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.* 7:105–114.

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 (Vol. 12). Invasive Species Specialist Group Auckland.

Lü J, Wang P. 2020. Modeling and Analysis of Bio-Molecular Networks. 1st ed. Springer: Singapore.

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.* 281:20132881.

Matsuda Y, Makino T. 2024. Comparative genomics reveals convergent signals associated with the high metabolism and longevity in birds and bats. *Proc. Biol. Sci.* 291:20241068.

Newman M. 2018. Networks. 2nd Ed. Oxford University Press: London, England.

Numanagic I, Gökkaya AS, Zhang L, Berger B, Alkan C, Hach F. 2018. Fast characterization of segmental duplications in genome assemblies. *Bioinformatics* 34:i706–i714.

Nuño de la Rosa L. 2017. Computing the extended synthesis: Mapping the dynamics and conceptual structure of the evolvability research front. *J. Exp. Zool. B Mol. Dev. Evol.* 328:395–411.

Nuño de la Rosa L. 2023. A history of evolvability: Reconstructing and explaining the origination of a research agenda. In T. F. Hansen, D. Houle, N. Pavličev, & C. Pélabon (Eds.), Evolvability: A Unifying Concept in Evolutionary Biology? The MIT Press. p. 11–34: Cambridge, MA.

Ohno S. 1970. Evolution by Gene Duplication. Springer: Berlin, Germany.

Paradis E, Schliep K. 2018. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35:526–528.

Pastor-Satorras R, Smith E, Solé RV. 2003. Evolving protein interaction networks through gene duplication. *J. Theor. Biol.* 222:199–210.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Duchesnay. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12:2825–2830.

Redmond AK, Casey D, Gundappa MK, Macqueen DJ, McLysaght A. 2023. Independent rediploidization masks shared whole genome duplication in the sturgeon-paddlefish ancestor. *Nat. Commun.* 14:2879.

Smith JDL, Gregory TR. 2009. The genome sizes of megabats (Chiroptera: Pteropodidae) are remarkably constrained. *Biol. Lett.* 5:347–351.

Solé RV, Pastor-Satorras R, Smith E, Kepler TB. 2002. A model of large-scale proteome evolution. *Adv. Complex Syst.* 05:43–54.

Sotero-Caio CG, Platt RN 2nd, Suh A, Ray DA. 2017. Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biol. Evol.* 9:161–177.

Sun C, Shepard DB, Chong RA, López Arriaza J, Hall K, Castoe TA, Feschotte C, Pollock DD, Mueller RL. 2012. LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. *Genome Biol. Evol.* 4:168–183.

Szarski H. 1983. Cell size and the concept of wasteful and frugal evolutionary strategies. *J. Theor. Biol.* 105:201–209.

Tenaillon O, Matic I. 2020. The Impact of Neutral Mutations on Genome Evolvability. *Current Biology* 30:R527–R534.

Vázquez A, Flammini A, Maritan A, Vespignani A. 2003. Modeling of protein interaction

networks. *Complexus* 1:38–44.

Villegas C, Love AC, Nuño de la Rosa L, Brigandt I, Wagner GP. 2021. Conceptual Roles of Evolvability across Evolutionary Biology: Between Diversity and Unification. *PhilSci Archive*. 2021;19517.

Vollger MR, Dishuck PC, Harvey WT, DeWitt WS, Guitart X, Goldberg ME, Rozanski AN, Lucas J, Asri M, Human Pangenome Reference Consortium, et al. 2023. Increased mutation and gene conversion within human segmental duplications. *Nature* 617:325–334.

Vollger MR, Dishuck PC, Sorensen M, Welch AE, Dang V, Dougherty ML, Graves-Lindsay TA, Wilson RK, Chaisson MJP, Eichler EE. 2019. Long-read sequence and assembly of segmental duplications. *Nat. Methods* 16:88–94.

Wasserman S., and K. Faust. 1994. Social Network Analysis. Cambridge University Press: Cambridge.

Watts DJ, Strogatz SH. 1998. Collective dynamics of “small-world” networks. *Nature* 393:440–442.

Wright NA, Gregory TR, Witt CC. 2014. Metabolic “engines” of flight drive genome size reduction in birds. *Proc. Biol. Sci.* 281:20132780.

Zhu W, Wang Z, Li H, Li P, Ni L, Jiao L, Ren Y, You P. 2022. A chromosome-level genome of Brachymystax tsinlingensis provides resources and insights into salmonids evolution. *G3* 12:jkac162.

Supplementary Information

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

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

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	0.04 (0.2123)	0.48 (2.9343×10^{-12})	−0.09 (0.0065)	0.12 (5.0953×10^{-25})

weight				
Mean similarity score	−0.06 (0.0412)	0.85 (2.2588× 10^{-53})	0.01 (0.8447)	0.18 (1.6309× 10^{-47})

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× 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})

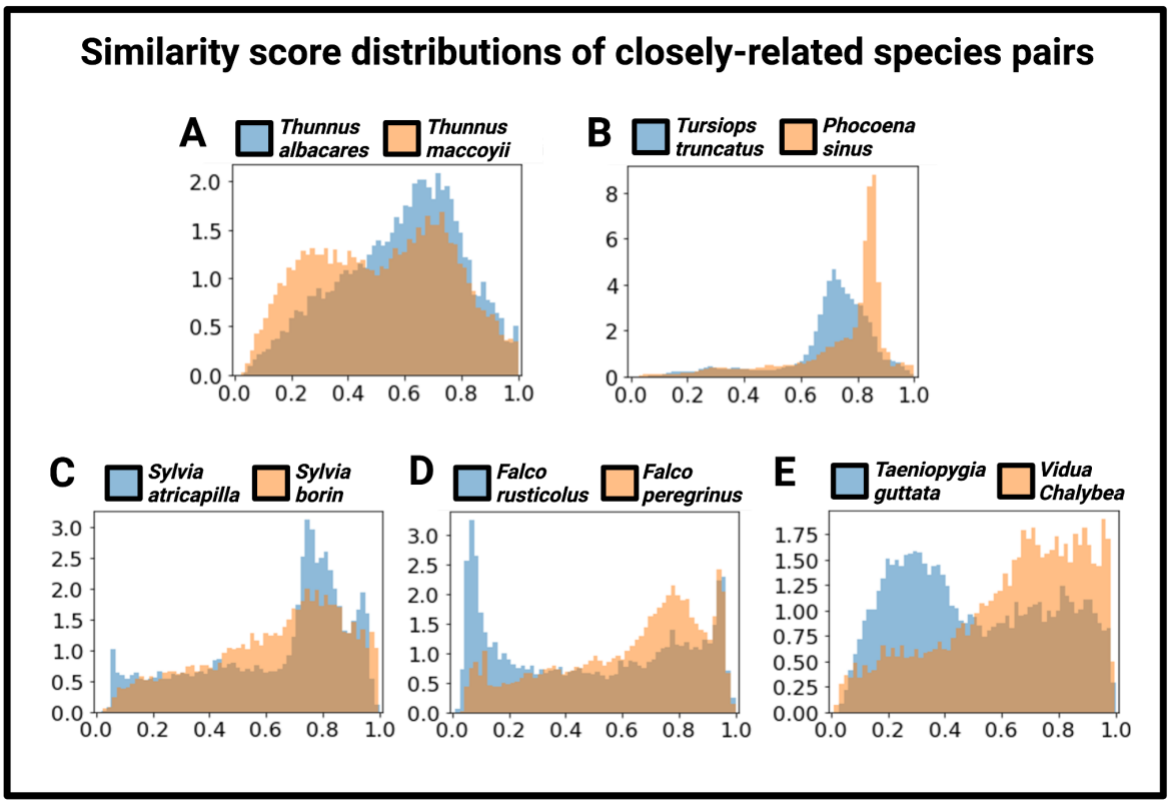

Figure S1. Similarity score distributions for closely-related species pairs. Five pairs of species that are phylogenetically close but have highly variable similarity distributions. The x-axis represents the similarity score and the y-axis represents density. **A.** Ray-finned fish. **B.** Mammals. **C-E.** Birds.

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.