Quantifying the influence of genetic context on duplicated mammalian genes

Alexander S. Moffett[1,2,*], Andrea Falcón-Cortés[1,2,*], and Michele Di Pierro[1,2]

[1]Center for Theoretical Biological Physics, Northeastern University, Boston, Massachusetts, 02115, USA
[2]Department of Physics, Northeastern University, Boston, Massachusetts, 02115, USA
[*]Indicates co-first authors.

April 3, 2025

Abstract

Gene duplication is a fundamental part of evolutionary innovation. While single-gene duplications frequently exhibit asymmetric evolutionary rates between paralogs, the extent to which this applies to multi-gene duplications remains unclear. In this study, we investigate the role of genetic context in shaping evolutionary divergence within multi-gene duplications, leveraging microsynteny to differentiate source and target copies. Using a dataset of 193 mammalian genome assemblies and a bird outgroup, we systematically analyze patterns of sequence divergence between duplicated genes and reference orthologs. We find that target copies, those relocated to new genomic environments, exhibit elevated evolutionary rates compared to source copies in the ancestral location. This asymmetry is influenced by the distance between copies and the size of the target copy. We also demonstrate that the polarization of rate asymmetry in paralogs, the "choice" of the slowly evolving copy, is biased towards collective, block-wise polarization in multi-gene duplications. Our findings highlight the importance of genetic context in modulating post-duplication divergence, where differences in cis-regulatory elements and co-expressed gene clusters between source and target copies may be responsible. This study presents a large-scale test of asymmetric evolution in multi-gene duplications, offering new insight into how genome architecture shapes functional diversification of paralogs.

Gene duplications play a crucial role in evolution, providing frequent opportunities for functional innovation. If a gene duplication fixes in a population, both copies can face relaxed purifying selection, leading to loss of function in one of the copies [1, 2]. However, several mechanisms have been proposed to explain long-term maintenance of both copies, including neofunctionalization, subfunctionalization, escape from adaptive conflict, and selection for increased gene dosage [3]. Most of the possible fates of a duplication are expected to result in asymmetric evolution of the two copies, as has been widely observed [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]. While it is not immediately clear which of the two copies will experience an accelerated rate of evolution, previous work suggests that the copy in the ancestral genomic location tends to evolve at similar rate before and after the duplication, while the copy inserted into a new location often experiences an accelerated evolutionary rate. Adopting the terminology of Dewey [20], we refer to the copy in the original location as the source copy and the the relocated copy as the target copy. If the target copy is more likely to have a higher substitution rate than the source copy, we say that asymmetrically evolving paralogs are polarized. Polarization of paralogs has been shown in small collections (varying between one and six) of model mammalian species [8, 9, 10, 13, 15] and nematodes [12]. Depending on the specific duplication event, this increase of evolutionary rate in the source copy can be consistent with positive selection [9, 13] or relaxed negative selection.

Substitution rate asymmetry between paralogs is often location-dependent, highlighting differences in the regulatory environment of the source and target copies, where cis-regulatory elements and co-expressed genes can be drastically different depending on the location of a gene [19]. Polarization has also been found to correlate with local recombination rate in yeast [6], suggesting that the relevant genetic context cannot be reduced to only cis-regulatory elements. Regardless of the specific nature of genetic context, both the distance between copies and the size of duplicated regions are factors that might influence polarization of paralogs. Both of these attributes of duplications reflect the conservation of genetic context between paralogs, where closeby source and target copies are likely to directly share cis-regulatory elements while large duplications may carry cis-regulatory elements and co-expressed genes to the source location. Paralogs resulting from tandem duplications do not seem to evolve asymmetrically [8, 10]. In *C. elegans*, the genomic distance between target and source copies on the same chromosome has little effect on rate asymmetry, while relocation of the target copy to a different chromosome is associated with accelerated evolution [12]. In the same study, the length of the duplicated region was inversely related to rate asymmetry, so that smaller duplications tend to evolve more asymmetrically. Additionally, asymmetric evolution in multi-gene duplications is not well-characterized, despite the importance of segmental duplications [21, 22, 23, 24, 25, 26, 27, 28, 29, 30] and whole-genome duplications [31, 32, 33]. Paralogs resulting from whole-genome duplications, called “onhologs” or “homeologs”, have been found to evolve asymmetrically [7, 34], while a study on multi-gene duplications in *A. thaliana* did not find significant asymmetric evolution [35].

Despite considerable progress in characterizing the polarization of gene duplications, more work is needed in order to uncover quantitative relationships among inter-paralog distances, sizes of duplicated regions, and rate asymmetry and polarization. Here, we address this gap in the literature, analyzing gene duplication in a broad sample of mammals. We focus on multi-gene duplications, which allows us to reliably identify source and target copies and quantitatively analyze the effects of genetic context on polarization. We test the location-dependence of paralog polarization using multi-gene duplications found through microsynteny analysis. Using these duplications and estimates of their duplication time, we find strong evidence for polarization, where source copies tend to evolve slower than target copies. We observe decreasing polarization with the size of duplicated regions and a weak relationship between polarization and the genomic distance between the two copies on the same chromosome, while paralogs on different chromosomes exhibit the strongest polarization. Finally, we observe block-wise polarization for simultaneously duplicated genes. Altogether, these results highlight the importance of genetic context in the evolutionary fate of gene duplications.

We first identify microsynteny blocks, regions of conserved gene order, by analyzing the structure of gene homology across pairs of genomes. Microsynteny forms a basis for comparing genomes of distantly related species and accurately inferring orthology [36, 37, 38, 39, 40, 41, 42, 43, 44]. We use microsynteny to identify both source and target copies (Fig. 1a), where the largest microsynteny block that a gene belongs to (primary microsynteny) is likely the source copy. The correspondence between primary microsynteny and source copies follows from the likelihood that the flanking genes around the duplicated region have conserved order in species diverging before the duplication event, while the target copy is will not have this property.

To find microsynteny in mammals, we must first infer homology relationships between genes and then identify conserved gene ordering. We collected chromosome-length genome assemblies from the DNA Zoo [45] and Gencode [46] to assemble a dataset of 193 mammalian genome assemblies and a single bird outgroup (Fig. 1b). While past work on single-gene duplications has relied on a small number of genomes from model organisms, we perform a comprehensive test of location-dependent evolutionary asymmetry in a dense set of mammals. We used OrthoFinder 2 [47, 48] to identify orthogroups (sets of homologous genes across any number of species in the analysis) from translated amino acid sequences (Fig. S1). In order to remove the effects of simple tandem duplications on our results, we condensed each sequence of consecutive genes all in the same orthogroup into a single representative gene, which we also call tandem duplication equivalence classes (TDECs). This commonly used strategy [49, 50, 51] simplifies the identification of multi-gene duplications while retaining information about simple tandem duplications (where a single gene has been duplicated one or more times immediately adjacent to the original copy). Unless explicitly stated otherwise, we use “TDEC” and “gene” interchangeably for ease of reading.

We then produced homology matrices between all unique pairs of species, including comparisons

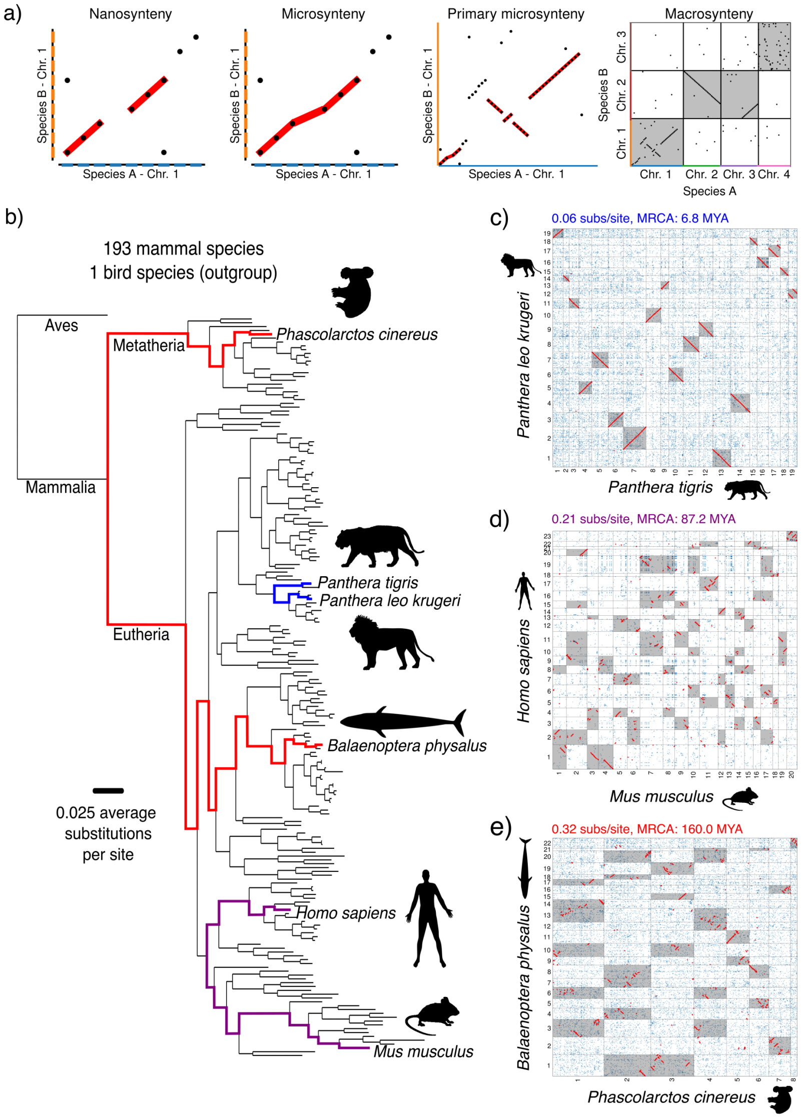


Figure 1: **Nanosynteny, microsynteny, and genetic context in mammals.** a) We define nanosynteny as perfectly preserved gene order, as observed in pairwise dot plots. Microsynteny refers to conserved gene order allowing for small lineage-specific duplications, deletions, translocations, and inversions. Both nanosynteny and microsynteny fit under the widely used definition of collinearity. Primary microsynteny is the largest microsynteny block containing a gene from one of the two genomes in a dot plot. This means that primary microsynteny is defined with respect to a focal and reference species. For example, we choose species A to the focal species and species B to be the reference species, and we can see a multi-gene duplication has occurred in the species A lineage. The genes in species B that are syntenic to the two duplicated regions of species B are contained in a larger synteny block (highlighted in red) and a smaller block. The larger synteny block, primary microsynteny, is likely the ancestral region that was duplicated, and is thus the original context of the duplicated genes before duplication. To distinguish nanosynteny and microsynteny from the traditional meaning of synteny, we use the term macrosynteny, meaning the conservation of genes in microsynteny between two chromosomes. b) The phylogeny of the 193 mammalian genome assemblies and single bird outgroup in our dataset, estimated with STAG [52] and STRIDE [53]. Branch lengths are the amino acid substitutions per site averaged over all orthologs. Three pairwise comparisons are highlighted, corresponding to dot plots between (in order of increasing time to MRCA) c) *Panthera leo krugeri* and *Panthera tigris* (0.06 average substitutions per site, MRCA 6.8 MYA) d) *Homo sapiens* and *Mus musculus* (0.21 average substitutions per site, MRCA 87.2 MYA) e) *Balaenoptera physalus* and *Phascolarctos cinereus* (0.32 average substitutions per site, MRCA 160 MYA). Microsynteny is highlighted in red, and macrosynteny between chromosomes is highlighted in grey. Note the decreasing lengths of microsyntenic regions with increasing time to MRCA, reflecting increased relative fragmentation of genomes with time.

between each genome and itself, totaling 18,915 comparisons. A homology matrix is a binary matrix H_{ij} representing a comparison between species A and B, where i runs over all genes in species A and j in species B. If genes i and j are homologous, then $H_{ij} = 1$, otherwise $H_{ij} = 0$. Homology matrices are visualized as so-called dot plots, which allow rapid inspection of gene order conservation (Fig. 1c-e).

Because of the large amount of data, microsynteny blocks must be inferred algorithmically from homology matrices. Although the problem has already been subject to investigation [49, 50], we developed a novel approach to inferring microsynteny tailored to to address the specific objectives of this study. Our algorithm, dubbed APES (**AP**ES **E**xplore **S**ynteny), identifies synteny at the gene level, which is significantly easier than trying to do the same at the level of nucleotide sequences. Additionally, instead of attempting to recover all synteny, APES is designed to yield a low false positive rate. To achieve this goal, APES relies on first identifying blocks of perfectly conserved gene order, which we call nanosynteny, and then assembling nearby syntenic genes to determine microsynteny blocks (Fig. 2). In this way, each microsynteny block contains at least one nanosynteny block. By randomly permuting homology matrices, we show that nanosynteny blocks containing 3 or more pairs of genes (each pair containing a single orthologous gene from each species in the comparison) are highly unlikely to arise by chance (Fig. 2b-c). The strong statistical significance of nanosynteny blocks with 3 or more putative orthologous gene pairs provides a solid foundation for construction of reliable microsynteny blocks. A complete description of the APES algorithm, together with source code, can be found in the Methods section.

Using APES, we identified microsynteny blocks in all species comparisons. In homology matrices, two or more microsynteny blocks containing the same sequence of three or more genes reliably indicate a multi-gene duplication. Next, we sought to determine the source and target copies in these duplications. While other methods exist [54], a straightforward approach to identifying the source copy is to look for the copy with the longest stretch of microsynteny (primary microsynteny) with the reference genome.

We then calculated the Levenshtein distance between each paralog in the source and target copies and the corresponding syntenic ortholog in the reference genome, yielding the number amino acid substitutions per site. Referring to the source paralog as A1, the target paralog as A2, and the syntenic reference ortholog as B1, the difference in distance $\Delta d_L = d_L(\mathrm{A1}, \mathrm{B1}) - d_L(\mathrm{A2}, \mathrm{B1})$ is the difference between Levenshtein distance between the paralog in the larger, primary microsynteny block (source) with the reference species from the Levenshtein distance for the paralog in the smaller microsynteny block (target). Δd_L is negative if the source paralog has experienced a slower substitution rate than the target paralog, positive if the source paralog has experienced a faster substitution rate, and zero if both paralogs have experienced the same substitution rate.

As expected, we found that for focal genomes, a gene is very likely to be most similar to the corresponding gene in the reference genome contained in primary microsynteny (Fig. 3). We denote

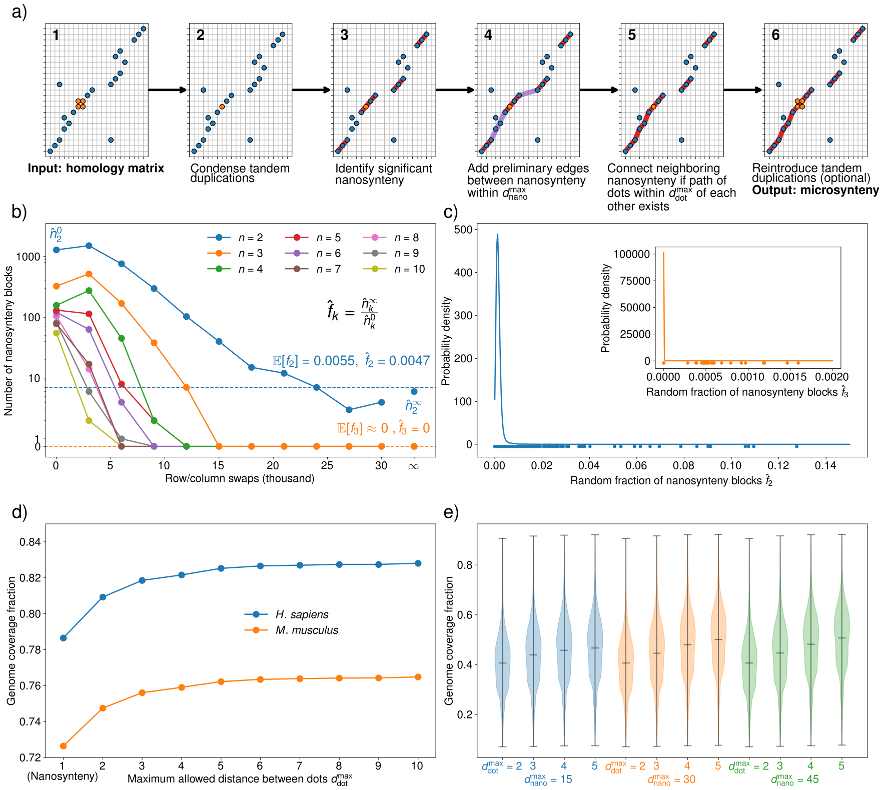


Figure 2: **Inferring statistically significant microsynteny.** a) A schematic of APES for identifying microsynteny blocks, which builds microsynteny blocks from statistically significant nanosynteny blocks in an easily interpretable manner. Step 1: The algorithm takes a homology matrix as input, as well as three hyperparameters: the maximum distance allowed between gene pairs ($d_{\mathrm{dot}}^{\mathrm{max}}$), the maximum distance allowed between nanosynteny blocks ($d_{\mathrm{nano}}^{\mathrm{max}}$), and the minimum anchoring nanosynteny block size. Step 2: We find nearest neighbor genes in both genomes which belong to the same orthogroup and condense them into single condensed genes for each set of neighboring genes with matching orthogroups. This prevents false positive microsynteny blocks within large tandem duplication blocks. Step 3: Next, we identify all nanosynteny from the homology matrix, retaining nanosynteny blocks containing more than the minimum allowed number of gene pairs. This minimum size parameter can be chosen according to a permutation test as detailed in b) and c). Step 4: We then add preliminary edges between nanosynteny blocks within $d_{\mathrm{nano}}^{\mathrm{max}}$ of each other. Step 5: For each pair of connected nanosynteny blocks, we check if there is a path of gene pairs within $d_{\mathrm{dot}}^{\mathrm{max}}$ of each other bridging the two nanosynteny blocks. If at least one such path exists, we combine the two nanosynteny blocks into a preliminary microsynteny block along with the shortest identified path of gene pairs. This step is then iterated until all preliminary edges between nanosynteny blocks are checked. Step 6: TDECs can optionally then be expanded into their constituent genes and integrated into microsynteny blocks. We do not expand TDECs in this article, so that microsynteny blocks are relationships between TDECs in pairs of genomes. b) A permutation test reveals that nanosynteny blocks with more than two gene pairs are unlikely to appear in a homology matrix by chance. We performed random swaps of rows and columns in the *H. sapiens*-*M. musculus* homology matrix, where each swap consisted of randomly choosing and switching two rows and two columns. We then found nanosynteny blocks in the permuted homology matrices. Nanosynteny blocks containing k gene pairs are shown in different colored lines for increasing numbers of random row/column swaps. $\hat{n}_2^0$ is the number of nanosynteny blocks containing $k = 2$ gene pairs for zero swaps, while $\hat{n}_2^\infty$ is the same number for 30,000 swaps, where the matrix is effectively random. The number $\hat{f}_k = \hat{n}_2^\infty / \hat{n}_2^0$ describes the fraction of nanosynteny blocks in the original homology matrix that occur due to random chance. For $k > 2$, the observed fraction is zero. We calculate the expected fractions $\mathbb{E}[f_k]$ from a percolation theoretic statistical model (see Supplemental Information and Fig. S2), finding a close match with the observed fractions. c) We performed the analysis in b) for homology matrices between all pairs of species in our dataset, and calculated $\hat{f}_2$ and $\hat{f}_3$ for each. We used kernel density estimates to represent the distributions of $\hat{f}_2$ and $\hat{f}_3$ (inset), with the dots below representing the value for each homology matrix. While $\hat{f}_2$ reached up to ~ 0.25 for some species pairs, $\hat{f}_3$ was below 0.003 for all species pairs. We chose a minimum nanosynteny block size of $k = 3$ gene pairs for all subsequent analysis. d) Microsynteny blocks are insensitive to maximum allowed distances between gene pairs of $d_{\mathrm{dot}}^{\mathrm{max}} = 3$ and larger. We identified synteny blocks in the *H. sapiens*-*M. musculus* homology matrix with a range of $d_{\mathrm{dot}}^{\mathrm{max}}$ values and measured the fraction of genes in each genome that were within a microsynteny block. By $d_{\mathrm{dot}}^{\mathrm{max}} = 3$, genome coverage fractions for both species plateau, indicating that the genes within microsynteny blocks change little for larger $d_{\mathrm{dot}}^{\mathrm{max}}$ values. e) We repeated the analysis in d) for all pairs of species in our dataset and used kernel density estimates to represent the distributions of genome coverage fractions. Again, by $d_{\mathrm{dot}}^{\mathrm{max}} = 3$ the distribution of genome coverage fractions is similar to the distribution for $d_{\mathrm{dot}}^{\mathrm{max}} = 10$, suggesting that the conclusion from d) applies to the dataset as a whole.

f as the fraction of genes in the reference genome for which the homologous focal gene in primary microsynteny is the closest in Levenshtein distance ($\Delta d_L < 0$) calculated over all cases where $\Delta d_L \neq 0$. Thus $f = P_{AB}(\Delta d_L < 0 | \Delta d_L \neq 0)$ is defined for an ordered pair of species, where switching species A with species B between the roles of focal and reference species can change f. We define $f^{\star}$ as the same fraction, except calculated only for reference genes belonging to a multi-gene duplication in the focal genome. The fraction $f^{\star}$ is designed to indicate how often genes in source copies of multi-gene duplications evolve slower than genes in target copies. We find that f is large across all species comparisons, with the majority of species comparisons yielding f near 0.9 (Fig. 3a-g). Using only nanosynteny yields very similar results (Fig. S3). This result supports the widespread use of synteny in determining orthology. When we restrict our analysis to genes in multi-gene duplications, we find a much broader distribution of $f^{\star}$ peaked around 0.8 but with very few values less than 0.5 (Fig. 3a-g). While this result seems to suggest asymmetric evolution of copies after multi-gene duplications, a more careful analysis of phylogenetic relationships between focal and reference species is needed.

In order to understand the significance of our results, we divided the data into before- and after-speciation duplications, as these two cases are qualitatively different. When the duplication occured before speciation, as in the example shown in Fig. 4a, genes A1 and A2 are paralogs, while A1 and B1 are orthologs, and A2 and B1 are paralogs. This means that the divergence times between A1 and B1 and A2 and B1 are different, obscuring differences in substitution rate. In this case, primary microsynteny does not reliably correlate with the source copy (Fig. 4b). However, when the duplication occurred after speciation as shown in Fig. 4d, differences in substitutions between A1 and B1 compared to A2 and B1 are only due to differences in substitution rate and random chance, not divergence time. Primary microsynteny can be used for after-speciation duplications to identify the source and target copies (Fig. 4e).

We took a conservative approach to estimating the duplication time for each multi-gene duplication through a consensus of two independent methods. Our intent was to minimize the number of times we underestimated the duplication times, to ensure a low false positive rate in determining gene triplets resulting from after-speciation duplications in the focal genome in comparison with an ortholog from the reference genome. In the first method, we gathered the duplication times inferred by DLCpar [55] for each paralogous pair in the multi-gene duplication. Within each multi-gene duplication, we then searched for the paralogous pair inferred to arise from a well-supported duplication (50% or more of species that should have inherited the duplication in the absence of deletions have both copies of the gene) with the earliest (largest, in MYA) duplication time. This duplication time was assigned to the entire multi-gene duplication. In the case where none of the paralogous pairs were inferred to be the result of a well-supported duplication, we simply chose the earliest duplication time to represent the multi-gene duplication. Our second method involved searching for the presence of two or more copies

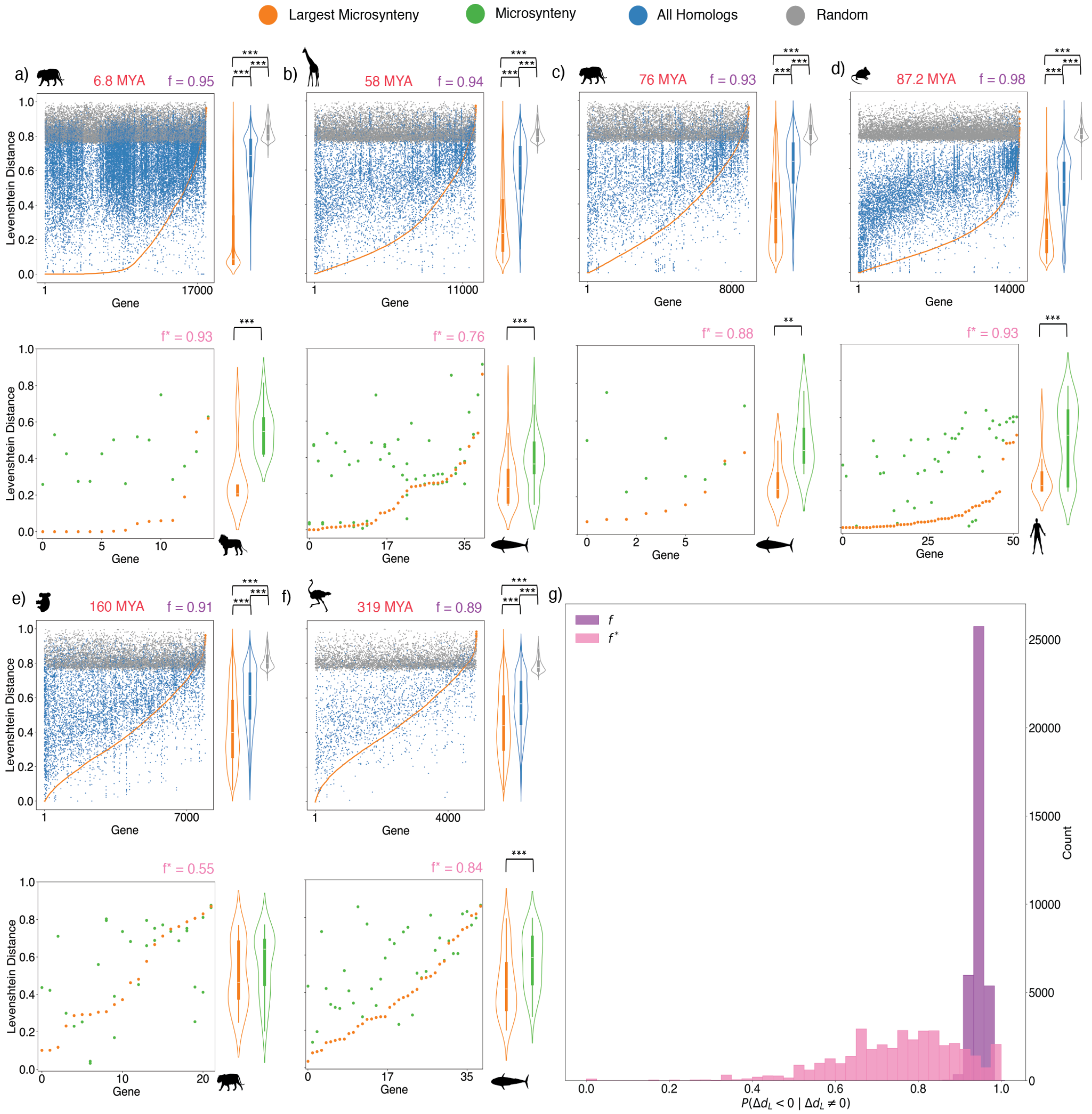


Figure 3: **Microsynteny reliably predicts the most similar homolog in pairwise genome comparisons.** For each pairwise genome comparison, we designated one species as "reference" and the other as "focal". We show example pairwise comparisons, written with the reference species first and focal species second, ordered by increasing time to MRCA: a) *Panthera leo krugeri* vs. *Panthera tigris* b) *Balaenoptera physalus* vs. *Giraffa camelopardalis* c) *Balaenoptera physalus* vs. *Panthera tigris* d) *Homo sapiens* vs. *Mus musculus* e) *Panthera tigris* vs. *Phascolarctos cinereus* f) *Balaenoptera physalus* vs. *Struthio camelus*. The upper plots of a)-f) have the indices of all genes with an orthogroup assignment from the reference species shown on the x-axis. Each point on the plot is a homologous (blue and orange) or randomly chosen gene (gray) in the focal species. Orange points with the same x-value represent the unique homolog of a reference species gene among all of its focal species homologs, belonging to the largest microsynteny block. The y-value of each point is the Levenshtein distance between the reference gene and a homolog or randomly chosen gene in the focal species. The value of f is the fraction of reference genes for which the corresponding homolog in the largest microsynteny block has the lowest Levenshtein distance of all homologs. All distributions of Levenshtein distances for the largest-microsyteny, homolog, and random genes were significantly different from each other (two-sample t-test, $p < 0.05$). The lower a)-f) show similar results, except that only reference genes that belong to at least two microsynteny blocks in the comparison with the focal species are shown. $f^\star$ is defined similarly to f but for the restricted set of genes shown in the lower plots. All distributions were significantly different (two-sample t-test, $p < 0.05$), except in e), where $f^\star = 0.55$. g) The distributions of f and $f^\star$ for all 37,442 ordered pairs of species, excluding comparisons of genomes to themselves. The distribution of f is sharply peaked around 0.9, demonstrating the utility of synteny in identifying the closest homolog between species. On the other hand, the $f^\star$ distribution is far more broad, reflecting the need for a more in-depth analysis of $f^\star$.

of the regions resulting from a multi-gene duplication in each other genome. We then determined the most recent common ancestor of all species in which we found two or more copies of the duplicated regions. The time to this ancestral species should be an upper bound on the duplication time if we assume there cannot be independent multi-gene duplication events of the same region. As a consensus rule, we took the larger of the two calculated duplication times. This process identified multi-gene duplication timings with a fairly uniform distribution (Fig. S4).

Because after-speciation duplications directly reflect the asymmetry of evolutionary rates between the source and target copies, we use this information to characterize the effect of genetic context in the evolutionary rate of otherwise apparently identical copies at the time of duplication. For the after-speciation subset, we estimated the substitution rates of the source (r_{source}) and target (r_{target}) copies. Of these 22,342 gene triplets, 26% had no difference in rate ($r_{\text{source}} = r_{\text{target}}$). Another 30% of gene triplets had a higher substitution rate for source copy ($r_{\text{source}} > r_{\text{target}}$) while 44% had a lower substitution rate for source copy ($r_{\text{source}} < r_{\text{target}}$) (Fig. 4g). This amounted to a probability of $P(\Delta d_L < 0|\Delta d_L \neq 0) = 0.593$ overall, so that roughly 60% of all cases where there is rate asymmetry between paralogs in multi-gene duplications, the target copy evolves faster than the source copy. All results were similar when we replaced microsynteny analysis with nanosynteny (Fig. S5). Note that the probability $P(\Delta d_L < 0|\Delta d_L \neq 0)$ pools all species pairs together, in contrast to $P_{AB}(\Delta d_L < 0|\Delta d_L \neq 0)$ which is defined with respect to an ordered pair of species.

We did not find a strong relationship between rate asymmetry and duplication time. Visually, $P(\Delta d_L < 0|\Delta d_L \neq 0)$ had little dependence on duplication time (Fig. 4h), while the correlation between the sign of Δd_L and duplication time was very weak ($\rho = -0.042$, $p = 4.35 \cdot 10^{-10}$). This result is inconsistent with previous findings of accelerated rates in young duplications [13, 15]. However, we did not focus our efforts on estimating fine-grained, branch-specific rates and further work is needed to examine the issue of rate acceleration and deceleration after duplication in mammals.

Our analysis to this point implicitly considers each gene triplet to be independent from all others. This assumption can fail for two reasons. First, two gene gene triplets that only differ by the reference species ortholog clearly are the result of the same duplication event. Second, a pair of paralogs in one species and a pair of paralogs in another species can arise from a single duplication event followed by speciation, so that a gene triplet containing one such pair of paralogs and a gene triplet containing the other are not independent. We used several approaches to address the impact of this interdependence on our results. First, by performing bootstrap subsampling from the set of species, we can effectively reduce the impact of dependence by decreasing the probability of choosing two species that share two or more copies of a genomic region resulting from a duplication. Bootstrap subsampling yielded a low probability of finding $P(\Delta d_L < 0|\Delta d_L \neq 0) < 0.5$ even when using only a quarter of the species from our dataset (Fig. 4i). This shows that our results are robust to significantly reductions in the dataset,

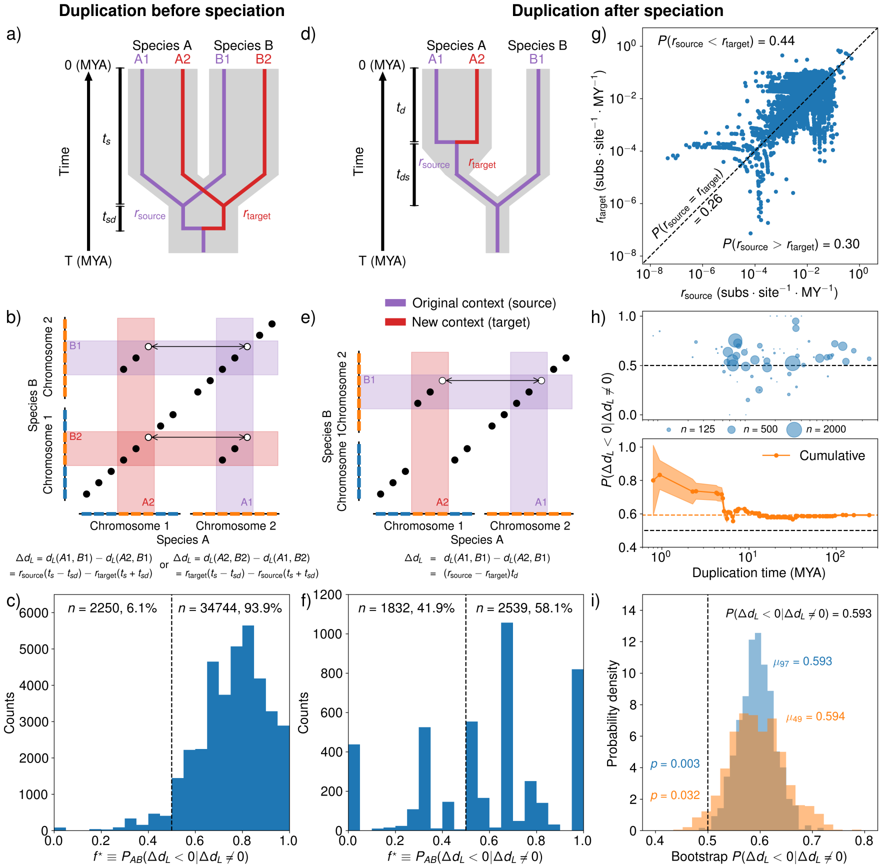


Figure 4: **Target gene copies within multi-gene duplications tend to evolve faster source copies.** a) Illustration of gene duplication before speciation in a two-species comparison. In this example, there are two genes from some orthogroup in both species A and species B. The large gray tree in the background is the species tree, while the thin tree in the foreground is the gene tree. Species A and species B are separated by a speciation event t_s MYA, which was preceded by a duplication event in the focal orthogroup $t_s + t_{sd}$ MYA. The purple branch of the gene tree represents the gene lineage descended from the source copy, while the red branch is the target copy lineage. b) A homology matrix (i.e. dot plot) showing evidence of a multi-gene duplication, where each gene within the duplicated region has a gene tree as shown in a). In this case, where the multi-gene duplication occured before speciation, we cannot determine the original context of genes B1 or B2 from primary microsynteny alone. If we take the largest synteny block containing B1, we recover its original context in chromosome 2, while following the same procedure for B2 leads us to misidentify its original context as being within chromosome 1. c) The distribution of $f^{\star}$ for before-speciation duplications. The $f^{\star}$ distribution is markedly skewed to the right, with 93.9% of $f^{\star}$ values lying above 0.5. This is likely due to gene divergence times, rather than differences in evolutionary rates. Genes A1 and B1 (A2 and B2) in a) and b) diverged more recently than genes A2 and B1 (A1 and B2), so that a greater number of substitutions per site between A2 and B1 (A1 and B2) than between A1 and B1 (A2 and B2) is expected even if $r_{\text{source}} = r_{\text{target}}$. d) If we instead compare paralogs resulting from a duplication t_d MYA with an orthologous gene in a species that diverged $t_d + t_{ds}$ MYA, before the duplication occurred, we can isolate the effects of differential evolutionary rates. e) By calculating the substitutions per site between A1 and B1 and between A2 and B1, we can infer the rate of amino acid substitution. In this case, primary microsynteny is associated with the original context while the smaller microsynteny block contains the target copy of the duplication. f) When $f^{\star}$ is calculated for after-speciation comparisons, we still see 58.1% of species pairs with $f^{\star} > 0.5$, suggesting that target copies of duplications may evolve faster than source copies. g) Inferred evolutionary rates for genes in target copies and their paralogs in source copies. For 44% of gene triplets examined, the rate of source copy substitutions was smaller than the rate of target copy substitutions ($r_{\text{source}} < r_{\text{target}}$). On the other hand, 30% of gene triplets yielded as larger source substitution rate ($r_{\text{source}} > r_{\text{target}}$) and 26% of paralogs evolved at identical rates ($r_{\text{source}} = r_{\text{target}}$). Some gene triplets with near-zero r_{source} or r_{target} are not shown. This demonstrates the asymmetry of paralog evolution after duplication, and the polarization substitution rates so that target copies tend to evolve faster. h) For duplications separated by inferred age, no clear trend emerged between rate asymmetry $P(\Delta d_L < 0 | \Delta d_L \neq 0)$ and age (top). We also calculated cumulative $P(\Delta d_L < 0 | \Delta d_L \neq 0)$ for all duplications of age t MYA and younger, with little relationship between age and asymmetry (bottom). Including all gene triplets, we found that 59.3% of gene triplets with non-identical paralogs had an elevated evolutionary rate in the target copy compared with the source copy. i) To see how robust our results were to changes in the included species, we performed bootstrap sampling of species in our analysis without replacement, with either 97 species or 49 species per sample. For both sample sizes, we drew 1,000 samples and calculated $P(\Delta d_L < 0 | \Delta d_L \neq 0)$ using only gene triplets from the sample species. Even with a quarter of the original dataset, the probability of finding $P(\Delta d_L < 0 | \Delta d_L \neq 0) \leq 0.5$ remained low ($p = 0.032$). This indicates that our results are not due to the exact set of species we chose to include in our dataset.

suggesting that the apparent polarization in rate asymmetry is not a consequence of interdependent data or of the specific choice of species for analysis. We also performed our analysis while averaging over all relevant reference species for each pair of paralogs (i.e. for those species that diverged from the lineage of the focal species before the duplication event creating the paralogs of interest), both for all pairs of paralogs and for only those resulting from lineage-specific duplications. Averaging over references removes the possibility of comparing two gene triplets that are identical except for the reference ortholog, while examining terminal duplications averaged over references ensures that each gene triplet is independent. Additionally using a null model of amino acid substitution to account for differences likely to occur purely by random chance when $r_{\text{source}} = r_{\text{target}}$, we find strong polarization in statistically significant cases for both non-terminal and terminal duplications (Fig. S6).

Next, we sought to examine how the differences in the regulatory environments of the source and target copies impact the polarity of asymmetric evolution. We expect that the larger that a duplicated region is, the greater the chance is that the regulatory environment around a gene is carried to the target location in the genome. Consequently, we expected to see polarization decrease with the size of the duplication if asymmetry in regulatory environment between the source and target copies contributes to polarization. Accordingly, we found that polarity decreased with larger target copies, suggesting that larger duplications lead to lower polarity (Fig. 5a). When the target copy is small, r_{source} is much more likely to be greater than r_{target} than for large target copies. Notably, in the largest class of duplications, we still found polarization, suggesting that regulatory environments can extend past ~ 1 Mb or that other factors influence polarity. Such long-range regulatory activity is consistent with the largest enhancer-to-transcription start site distances found in the human genome [56]. Overall, the inverse relationship we found between polarization and duplication size supports the hypothesis that larger duplications are more likely to carry cis-regulatory elements and important co-expressed genes to the target genomic location, creating a similar regulatory environment to the source location and increasing purifying selection on target copy genes.

We also expected polarity to increase with the genomic distance between copies, reasoning that distant target copies would experience different regulatory environments from the source copy while nearby target copies would experience similar regulatory environments. The dependence of polarization on genomic distance between the two copies is less clear, with polarity appearing more or less insensitive for intrachromosomal duplications (Fig. 5b). However, polarity was maximal for copies on different chromosomes, consistent with previously reported results in *C. elegens* [12]. While increasing distance between copies on the same chromosome appears to have a moderate influence on polarity, the novel regulatory environment faced by target copies on different chromosomes from their source copies may explain the polarity found in their evolutionary rates.

While paralogous pairs within large duplicated regions clearly demonstrate polarity, we sought to

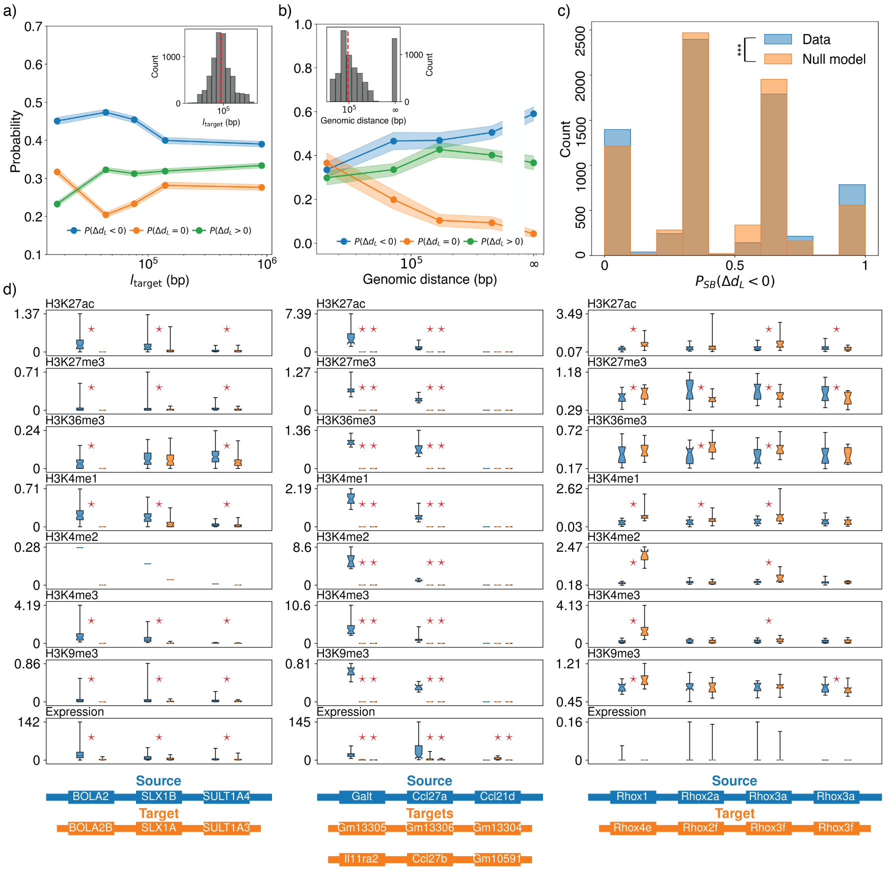


Figure 5: **The fate of multi-gene duplications is influenced by the size of the duplicated region and the genomic distance between copies, and is shared for genes within a copy.** a) The probability of the source copy evolving slower than the target copy, $P(\Delta d_L < 0)$, falls with larger target copy size in basepairs, l_{target}, while the probability of the source copy evolving faster, $P(\Delta d_L > 0)$, increases. The difference between the two probabilities represents the polarity of asymmetric paralog evolution, such that when $P(\Delta d_L < 0) < P(\Delta d_L > 0)$, the source copy is more likely to be the slowly evolving copy. Inset: the distribution of target copy sizes resulting from multi-gene duplications. b) When the same probabilities are examined as a function of the genomic distance between copies, both $P(\Delta d_L < 0)$ and $P(\Delta d_L > 0)$ increase with increased genomic distance. Again, the difference between the two probabilities indicates the polarity of asymmetric evolution. While there is not as clear of a trend as with l_{target}, at very small genomic distances $P(\Delta d_L < 0)$ and $P(\Delta d_L > 0)$ are very similar while at some larger distances $P(\Delta d_L < 0)$ and $P(\Delta d_L > 0)$ diverge. When the two copies are on different chromosomes (genomic distance of ∞), the difference between $P(\Delta d_L < 0)$ and $P(\Delta d_L > 0)$ is greatest. Note the monotonic decrease in $P(\Delta d_L = 0)$ with genomic distance. c) The probability of the source copy evolving slower than the target copy within a randomly chosen multi-gene duplication ($P_{SB}(\Delta d_L)$, SB for synteny block) is significantly different from a null model produced by randomly resorting paralogous genes into new multi-gene duplications. If all paralogous gene pairs within a large duplicated region were polarized in evolutionary rate towards either the target or source copy, we would expect to see only values of $P_{SB}(\Delta d_L) = 0$ or $P_{SB}(\Delta d_L) = 1$. When compared with an ensemble of gene triplets randomly assigned to multi-gene duplications, the probability of observing our original data was low ($p < 0.05$), with excess counts at $P_{SB}(\Delta d_L) = 0$ and $P_{SB}(\Delta d_L) = 1$. This suggests that the fate of multi-gene duplications is determined collectively more often than expected by random chance. All results in a)-c) were similar when only considering nanosynteny (Figs. S7 & S8). d) Analysis of ChIP-seq data for epigenetic markings associated with expression and RNA-seq data suggests that both may have a role in determining the polarity of asymmetric evolution in multi-gene duplications. While, only three sets of regions in the human and mouse genomes satisfied our criteria for analysis of ChIP-seq and RNA-seq data, we found significant differences (Wilcoxon signed-rank test, $p < 0.05$) across most paralogs in both epigenetic markings and expression for the two sets of regions on autosomal chromosomes in human (first column) and mouse (second column). For the large duplicated region in the mouse X chromosome (third column) there were fewer significant differences for epigenetic markings and no significant differences in expression. We examined individual genes in this analysis, rather than TDECs.

determine whether paralogous pairs created simultaneously in a single multi-gene duplication event share polarities. Since the genes within a single duplicated region share a common regulatory environment, we expect pairs of paralogs created in a single multi-gene duplication to be have correlated polarities, so that the “choice” of the slower and faster evolving paralog occurs in a block-wise manner. To test this expectation, we defined the within-synteny block probability $P_{SB}(\Delta d_L < 0)$ as the fraction of gene triplets within a duplicated region with $\Delta d_L < 0$. If all genes on in the source copy have evolved at a slower rate than their paralogs on the target copy $P_{SB}(\Delta d_L < 0)$ will be one, while $P_{SB}(\Delta d_L < 0)$ will be zero if the source copy genes have evolved faster than, or at the same as, their target copy paralogs. Intermediate probabilities reflect mixed polarities within duplicated regions. Using a null model that randomly assigns gene triplets to empirically determined duplicated regions, simulating what we would expect if the polarity of paralogous gene pairs was independently determined, we found that the $P_{SB}(\Delta d_L < 0)$ distribution displayed significant enrichment of $P_{SB}(\Delta d_L < 0) = 0$ and $P_{SB}(\Delta d_L < 0) = 1$. This is strong evidence that polarity determination in multi-gene duplications tends to occur in a block-wise manner rather than independently for each paralogous gene pair, suggesting that the fate of a duplicated gene can be influenced by the other genes and non-coding elements simultaneously duplicated.

Finally, we explored the role of differential epigenetic markings and gene expression on asymmetric evolution of paralogs. We restricted our analysis of epigenetic markings to histone modifications with known associations with gene expression [57, 58, 59, 60] (Table S1). Out of all multi-gene duplications in human and mouse, the two species for which we found experimental data, only three could be reliably placed in the after-speciation category. The single multi-gene human duplication had significant differences in both epigenetic markings and expression levels between the source and target copies, with higher markings and expression levels in the source copy (Fig. 5d). The distributions of expression levels and epigenetic markings represented in Fig. 5d are over all tissue types for which data was available. Thus, a broad distribution (with a wide range of expression or epigenetic marking levels) indicates variability across tissue types while a narrow distribution corresponds to homogeneity across tissues. The first set of duplicated mouse regions had two target copies associated with the source copy, and similarly had significant differences between the source copy and both targets. The second multi-gene mouse duplication was on the X chromosome, in contrast to the previous two duplications which had both copies on autosomal chromosomes. The source and target copies had fewer significant differences in epigenetic markings, while markings had higher levels in the target copy. Gene expression was not significantly different between any pair of paralogs. In total, these results are consistent with some role of epigenetic markings and gene expression in polarization. However, more work is needed to explore the molecular underpinnings of asymmetric evolution in multi-gene duplications.

Our findings provide substantial evidence for asymmetry and polarization in the amino acid substitution rate of source and target copies resulting from multi-gene duplications in a large dataset of mammals. In agreement with previous work on single-gene duplications, our results suggest an important role for location in polarizing substition rates of paralogs: the source copy in the ancestral genomic location tends to evolve slower than the target copy in the new locus created by the duplication event [8, 9, 10, 12, 13, 15]. By focusing on multi-gene duplications, we were able to demonstrate that polarization decreases with the size of the duplicated region, yet remains even for ~ 1 Mb duplications. Moreover, we demonstrate that the asymmetry in substitution rate tends to occur in a block-wise manner, so that genes in one copy tend to all evolve slower or faster than their paralogs in the other copy.

The statistical methods we used in synteny detection are central to our approach. By ensuring that nanosynteny is highly unlikely to have appeared simply due to random chance, and using statistically significant nanosynteny blocks as building blocks for microsynteny, we were able to identify reliable regions of conserved gene order across a large, representative set of mammals. APES allowed for easy interpretation of synteny blocks, where commonly used dynamic programming algorithms can yield conflicting synteny blocks without any notion of statistical significance.

Future work will also need to explicitly address the role of the three-dimensional organization of the genome and the locations of cis-regulatory elements around the source and target copies in determining the polarity of asymmetric evolution. As we have taken a phenomenological, mechanism-agnostic approach, it also remains to explore which models of duplication retention [3] determine substitution rate polarity in multi-gene duplications. Also, our focus on multi-gene duplications with microsynteny across species likely enriched our dataset for sets of genes with selection on their order, so that our data may not be representative of all multi-gene duplications in mammals.

Our efforts provide insight into how genes within multi-gene duplications evolve in mammals, a topic that has received far less attention than for single-gene duplications. If pairs of paralogs simultaneously produced by a single duplication event tend to polarize together as we have shown, the genes and non-coding DNA elements duplicated along with a gene have a significant impact on evolution of paralogs. On average, the impact of flanking regions around a paralog in the target region on polarization declines with the size of the target copy. Altogether, this suggests a key role for the genetic context carried along with target duplicated genes, which can decrease the frequency of polarization but cannot fully recreate the source regulatory environment.

Methods

Genome assemblies and annotations

We downloaded MAKER2 [61] genome annotations of 191 mammals and 1 bird (as an outgroup) from the DNA Zoo [45], along with the GENCODE human genome annotation v43 and mouse genome annotation v32 [46].

Orthogroup detection

We used OrthoFinder 2 [47, 48] to determine orthogroups, using DIAMOND [62] to identify putative homology relationships between genes. DIAMOND alignment scores are used as the edges of an orthogroup graph, with genes as nodes. The Markov Cluster Algorithm (MCL) [63] is then used to find communities in the orthogroup graph, yielding disjoint orthogroups. OrthoFinder also constructs a rooted species tree [53, 52] and rooted gene trees for each orthogroup, reconciled with the species tree through DLCpar [55]. DLCpar finds parsimonious gene tree reconciliations considering gene duplication, gene loss, and deep coalescence.

Microsynteny inference with APES

We developed an algorithm called APES (**AP**ES **E**xplore **S**ynteny) in order to reliably infer microsynteny with a low chance of yielding false positive results. APES infers microsynteny in each pair of species by first finding regions of perfectly conserved gene order (nanosynteny), and subsequently piecing together statistically significant nanosynteny into microsynteny. Nanosynteny blocks were considered statistically significant if they contained three or more TDEC pairs in order to ensure that the expected fraction of nanosynteny blocks found purely due to chance were less than 0.05 for all genome comparisons (Fig. 2). In order to avoid including regions only representing more complex tandem duplications, we filtered out nanosynteny blocks containing fewer than three unique orthogroups. We then joined nanosynteny blocks based on two hyperparameter choices, controlling the maximum distance between nanosynteny blocks in both genomes, $d_{\mathrm{nano}}^{\mathrm{max}}$, and the maximum distance of gene pairs from each other in the path between the nanosynteny blocks, $d_{\mathrm{dot}}^{\mathrm{max}}$. See Fig. 2 and the Supplemental Information for details. Python code is available at https://github.com/DiPierroLab/GenomicTools.

Finding multi-gene duplications and associated timings

We identified multi-gene duplications in each genome using within-genome microsynteny. We applied a series of filtering steps to the within-genome microsynteny blocks as follows:

1. We first excluded all microsynteny blocks representing alignment of each gene in a chromosome with itself.
2. We next removed any remaining microsynteny blocks that contain the same gene twice.
3. Finally, we removed overlaps between microsynteny blocks. If any two microsynteny blocks were on the same two chromosomes, the overlap is between them is defined as the smaller number of shared genes in both blocks between the first and second genome (see Supplemental Information

for details). The larger (measured in number of TDECs) of two overlapping blocks was retained, while the smaller block was removed.

All analysis of multi-gene duplications was performed on the remaining pairs of microsyntenic regions in each genome. We remained intentionally agnostic as to the mechanism producing large duplicate regions, though these could be a result of segmental duplications, whole-genome duplications with subsequent chromosomal rearrangements, or even complex tandem duplications.

In order to infer the time of each multi-gene duplication event, we used the species tree and reconciled gene trees produced by OrthoFinder. Because OrthoFinder does not take synteny into account, the reconciliation process is independent for each orthogroup. However, in multi-gene duplications we must introduce the constraint that all paralogous pairs were produced simultaneously in a single duplication event. While reconciliation methods exist for this problem [64, 65], we took a simpler approach designed to reduce the probability of underestimating duplication times (measured in millions of years from the present). We used a consensus of two independent methods. First, we used the reconciled gene trees from OrthoFinder. Each paralogous pair is assigned to a duplication event at a node in the species tree. We will refer to this as the duplication node. Each duplication node has an associated support, defined as the fraction of descendant leaves (i.e. extant animal genomes in the dataset) of the duplication node that have retained both paralogs resulting from the duplication event. We defined the duplication node of a multi-gene duplication as the earliest (closest to the root node), well-supported (≥ 0.5) duplication node out of all the paralogous pairs therein. If none were well-supported, then we simply took the earliest duplication node. Second, we searched for the sequence of orthogroups found in the multi-gene duplication in all genomes. If two or more copies of those orthogroup sequences were found in a species we added it to a list of multiple-copy species. The duplication node was then simply the MRCA of all multiple-copy species in the species tree.

We then took the earlier of the two duplication nodes from our two methods as the consensus duplication node. Using divergence times from TimeTree 5 [66], the duplication time is defined as the time halfway between the divergence time associated with the duplication node and its parent species tree node.

Substitution rate inference

We estimate amino acid substitution rates, rather than estimating synonymous and non-synonymous nucleotide substitution rates. Using amino acid substitution rates is justified in our analysis because of the large divergence times between the species in our dataset (such that non-synonymous nucleotide substitutions are in danger of saturation), and because we have reasonable estimates of divergence times between paralogs from OrthoFinder 2 and TimeTree. We calculate substitution rates by dividing

Levenshtein distances by duplication times. See the Supplemental Information for details.

Epigenetics and gene expression data

We downloaded human and mouse histone ChIP-seq data from the ENCODE project [67, 68, 69] for the markings in Table S1. For each marking, we calculated the average fold change over control signal over each relevant gene for each tissue type available. This left us with an ensemble of marking signals for each human and mouse gene. We downloaded human and mouse gene expression data for the tissues listed below from the ENCODE project using RNAget [70].

Acknowledgments

We thank the members of the Nuclear Physics working group in the Center for Theoretical Biological Physics for valuable discussion. We are especially grateful to Erez Aiden and Olga Dudchenko for their incisive input. This research was supported by the National Science Foundation through the awards PHY-2412651 and PHY-2019745, and by the National Institute of General Medical Sciences of the NIH under award R35-GM146852. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

[1] Michael Lynch and John S Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, 2000.

[2] Michael Lynch and Vaishali Katju. The altered evolutionary trajectories of gene duplicates. *TRENDS in Genetics*, 20(11):544–549, 2004.

[3] Hideki Innan and Fyodor Kondrashov. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, 11(2):97–108, 2010.

[4] Gavin C Conant and Andreas Wagner. Asymmetric sequence divergence of duplicate genes. *Genome Research*, 13(9):2052–2058, 2003.

[5] Peng Zhang, Zhenglong Gu, and Wen-Hsiung Li. Different evolutionary patterns between young duplicate genes in the human genome. *Genome Biology*, 4:1–6, 2003.

[6] Ze Zhang and Hirohisa Kishino. Genomic background predicts the fate of duplicated genes: evidence from the yeast genome. *Genetics*, 166(4):1995–1999, 2004.

[7] Guillaume Blanc and Kenneth H Wolfe. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *The Plant Cell*, 16(7):1679–1691, 2004.

[8] Brian P Cusack and Kenneth H Wolfe. Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. *Molecular Biology and Evolution*, 24(3):679–686, 2007.

[9] Mira V Han, Jeffery P Demuth, Casey L McGrath, Claudio Casola, and Matthew W Hahn. Adaptive evolution of young gene duplicates in mammals. *Genome Research*, 19(5):859–867, 2009.

[10] Jin Jun, Paul Ryvkin, Edward Hemphill, and Craig Nelson. Duplication mechanism and disruptions in flanking regions determine the fate of mammalian gene duplicates. *Journal of Computational Biology*, 16(9):1253–1266, 2009.

[11] Alexander Y Panchin, Mikhail S Gelfand, Vasily E Ramensky, and Irena I Artamonova. Asymmetric and non-uniform evolution of recently duplicated human genes. *Biology Direct*, 5:1–12, 2010.

[12] Vaishali Katju. To the beat of a different drum: determinants implicated in the asymmetric sequence divergence of Caenorhabditis elegans paralogs. *BMC Evolutionary Biology*, 13:1–15, 2013.

[13] Cinta Pegueroles, Steve Laurie, and M Mar Albà. Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. *Molecular Biology and Evolution*, 30(8):1830–1842, 2013.

[14] Lev Y Yampolsky and Michael A Bouzinier. Faster evolving drosophila paralogs lose expression rate and ubiquity and accumulate more non-synonymous snps. *Biology Direct*, 9:1–16, 2014.

[15] Oriol Pich i Roselló and Fyodor A Kondrashov. Long-term asymmetrical acceleration of protein evolution after gene duplication. *Genome Biology and Evolution*, 6(8):1949–1955, 2014.

[16] Peter WH Holland, Ferdinand Marlétaz, Ignacio Maeso, Thomas L Dunwell, and Jordi Paps. New genes from old: asymmetric divergence of gene duplicates and the evolution of development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1713):20150480, 2017.

[17] Rosa Fernández, Marina Marcet-Houben, Fabrice Legeai, Gautier Richard, Stéphanie Robin, Valentin Wucher, Cinta Pegueroles, Toni Gabaldón, and Denis Tagu. Selection following gene duplication shapes recent genome evolution in the pea aphid Acyrthosiphon pisum. *Molecular Biology and Evolution*, 37(9):2601–2615, 2020.

[18] Hana Jaafari, Carlos Bueno, Nicholas P Schafer, Jonathan Martin, Faruck Morcos, and Peter G Wolynes. The physical and evolutionary energy landscapes of devolved protein sequences corresponding to pseudogenes. *Proceedings of the National Academy of Sciences U.S.A.*, 121(21):e2322428121, 2024.

[19] UnJin Lee, Deanna Arsala, Shengqian Xia, Cong Li, Mujahid Ali, Nicolas Svetec, Christopher B Langer, Débora R Sobreira, Ittai Eres, Dylan Sosa, et al. The three-dimensional genome drives the evolution of asymmetric gene duplicates via enhancer capture-divergence. *Science Advances*, 10(51):eadn6625, 2024.

[20] Colin N Dewey. Positional orthology: putting genomic evolutionary relationships into context. *Briefings in Bioinformatics*, 12(5):401–412, 2011.

[21] Mitchell R Vollger, Philip C Dishuck, William T Harvey, William S DeWitt, Xavi Guitart, Michael E Goldberg, Allison N Rozanski, Julian Lucas, Mobin Asri, et al. Increased mutation and gene conversion within human segmental duplications. *Nature*, 617(7960):325–334, 2023.

[22] Lluís Armengol, Miguel Angel Pujana, Joseph Cheung, Stephen W Scherer, and Xavier Estivill. Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Human Molecular Genetics*, 12(17):2201–2208, 2003.

[23] Jeffrey A Bailey, Robert Baertsch, W James Kent, David Haussler, and Evan E Eichler. Hotspots of mammalian chromosomal evolution. *Genome Biology*, 5:1–7, 2004.

[24] Saiful Islam, Alber Aqil, Faraz Hach, Ibrahim Numanagic, Naoki Masuda, and Omer Gokcumen. Evolvability through segmental duplication follows species-specific dynamics. *bioRxiv*, 2024.

[25] Zhaoshi Jiang, Haixu Tang, Mario Ventura, Maria Francesca Cardone, Tomas Marques-Bonet, Xinwei She, Pavel A Pevzner, and Evan E Eichler. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nature Genetics*, 39(11):1361–1368, 2007.

[26] Tomas Marques-Bonet, Santhosh Girirajan, and Evan E Eichler. The origins and impact of primate segmental duplications. *Trends in Genetics*, 25(10):443–454, 2009.

[27] Tomas Marques-Bonet, Jeffrey M Kidd, Mario Ventura, Tina A Graves, Ze Cheng, LaDeana W Hillier, Zhaoshi Jiang, Carl Baker, Ray Malfavon-Borja, Lucinda A Fulton, et al. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature*, 457(7231):877–881, 2009.

[28] Peter H Sudmant, John Huddleston, Claudia R Catacchio, Maika Malig, LaDeana W Hillier, Carl Baker, Kiana Mohajeri, Ivanela Kondova, Ronald E Bontrop, Stephan Persengiev, et al. Evolution and diversity of copy number variation in the great ape lineage. *Genome Research*, 23(9):1373–1382, 2013.

[29] Xiaotian Feng, Jicai Jiang, Abinash Padhi, Chao Ning, Jinluan Fu, Aiguo Wang, Raphael Mrode, and Jian-Feng Liu. Characterization of genome-wide segmental duplications reveals a common genomic feature of association with immunity among domestic animals. *BMC Genomics*, 18:1–11, 2017.

[30] Mitchell R Vollger, Xavi Guitart, Philip C Dishuck, Ludovica Mercuri, William T Harvey, Ariel Gershman, Mark Diekhans, Arvis Sulovari, Katherine M Munson, Alexandra P Lewis, et al. Segmental duplications and their variation in a complete human genome. *Science*, 376(6588):eabj6965, 2022.

[31] Yuuta Moriyama and Kazuko Koshiba-Takeuchi. Significance of whole-genome duplications on the emergence of evolutionary novelties. *Briefings in Functional Genomics*, 17(5):329–338, 2018.

[32] Linda Z Holland and Daniel Ocampo Daza. A new look at an old question: when did the second whole genome duplication occur in vertebrate evolution? *Genome Biology*, 19:1–4, 2018.

[33] Daqi Yu, Yandong Ren, Masahiro Uesaka, Alan JS Beavan, Matthieu Muffato, Jieyu Shen, Yongxin Li, Iori Sato, Wenting Wan, James W Clark, et al. Hagfish genome elucidates vertebrate whole-genome duplication events and their evolutionary consequences. *Nature Ecology & Evolution*, 8(3):519–535, 2024.

[34] Kevin P Byrne and Kenneth H Wolfe. Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics*, 175(3):1341–1350, 2007.

[35] Liqing Zhang, Todd J Vision, and Brandon S Gaut. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in Arabidopsis thaliana. *Molecular Biology and Evolution*, 19(9):1464–1473, 2002.

[36] Jin Jun, Ion I Mandoiu, and Craig E Nelson. Identification of mammalian orthologs using local synteny. *BMC Genomics*, 10(1):1–13, 2009.

[37] Pavel S Novichkov, Igor Ratnere, Yuri I Wolf, Eugene V Koonin, and Inna Dubchak. ATGC: a database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes. *Nucleic Acids Research*, 37(suppl_1):D448–D454, 2009.

[38] David M Kristensen, Yuri I Wolf, Arcady R Mushegian, and Eugene V Koonin. Computational methods for gene orthology inference. *Briefings in Bioinformatics*, 12(5):379–391, 2011.

[39] Melvin Zhang and Hon Wai Leong. BBH-LS: an algorithm for computing positional homologs using sequence and gene context similarity. *BMC Systems Biology*, 6(1):1–11, 2012.

[40] Marcus Lechner, Maribel Hernandez-Rosales, Daniel Doerr, Nicolas Wieseke, Annelyse Thévenin, Jens Stoye, Roland K Hartmann, Sonja J Prohaska, and Peter F Stadler. Orthology detection combining clustering and synteny for very large datasets. *PLoS ONE*, 9(8):e105015, 2014.

[41] Raja H Ali, Sayyed A Muhammad, and Lars Arvestad. GenFamClust: an accurate, synteny-aware and reliable homology inference algorithm. *BMC Evolutionary Biology*, 16(1):1–19, 2016.

[42] Soheil Jahangiri-Tazehkand, Limsoon Wong, and Changiz Eslahchi. OrthoGNC: a software for accurate identification of orthologs based on gene neighborhood conservation. *Genomics, Proteomics & Bioinformatics*, 15(6):361–370, 2017.

[43] Kun Gao and Jonathan Miller. Primary orthologs from local sequence context. *BMC Bioinformatics*, 21(1):1–22, 2020.

[44] Bogdan M Kirilenko, Chetan Munegowda, Ekaterina Osipova, David Jebb, Virag Sharma, Moritz Blumer, Ariadna E Morales, Alexis-Walid Ahmed, Dimitrios-Georgios Kontopoulos, Leon Hilgers, et al. Integrating gene annotation with orthology inference at scale. *Science*, 380(6643):eabn3107, 2023.

[45] Olga Dudchenko, Sanjit S Batra, Arina D Omer, Sarah K Nyquist, Marie Hoeger, Neva C Durand, Muhammad S Shamim, Ido Machol, Eric S Lander, Aviva Presser Aiden, and Erez Lieberman Aiden. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*, 356(6333):92–95, 2017.

[46] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1):D766–D773, 2019.

[47] David M Emms and Steven Kelly. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1):1–14, 2015.

[48] David M Emms and Steven Kelly. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20:1–14, 2019.

[49] Brian J Haas, Arthur L Delcher, Jennifer R Wortman, and Steven L Salzberg. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics*, 20(18):3643–3646, 2004.

[50] Yupeng Wang, Haibao Tang, Jeremy D Debarry, Xu Tan, Jingping Li, Xiyin Wang, Tae-ho Lee, Huizhe Jin, Barry Marler, Hui Guo, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, 40(7):e49–e49, 2012.

[51] John T Lovell, Avinash Sreedasyam, M Eric Schranz, Melissa Wilson, Joseph W Carlson, Alex Harkess, David Emms, David M Goodstein, and Jeremy Schmutz. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *eLife*, 11:e78526, 2022.

[52] DM Emms and Steven Kelly. STAG: species tree inference from all genes. *bioRxiv*, 2018.

[53] David M Emms and Steven Kelly. STRIDE: species tree root inference from gene duplication events. *Molecular Biology and Evolution*, 34(12):3267–3278, 2017.

[54] Mira V Han and Matthew W Hahn. Identifying parent-daughter relationships among duplicated genes. In *Pacific Symposium on Biocomputing 14*, pages 114–125, 2009.

[55] Yi-Chieh Wu, Matthew D Rasmussen, Mukul S Bansal, and Manolis Kellis. Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. *Genome Research*, 24(3):475–486, 2014.

[56] Andreas R Gschwind, Kristy S Mualim, Alireza Karbalayghareh, Maya U Sheth, Kushal K Dey, Evelyn Jagoda, Ramil N Nurtdinov, Wang Xi, Anthony S Tan, Hank Jones, et al. An encyclopedia of enhancer-gene regulatory interactions in the human genome. *bioRxiv*, 2023.

[57] Rosa Karlić, Ho-Ryun Chung, Julia Lasserre, Kristian Vlahoviček, and Martin Vingron. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences U.S.A.*, 107(7):2926–2931, 2010.

[58] Mark B Gerstein, Joel Rozowsky, Koon-Kiu Yan, Daifeng Wang, Chao Cheng, James B Brown, Carrie A Davis, LaDeana Hillier, Cristina Sisu, Jingyi Jessica Li, et al. Comparative analysis of the transcriptome across distant species. *Nature*, 512(7515):445–448, 2014.

[59] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[60] Ritambhara Singh, Jack Lanchantin, Gabriel Robins, and Yanjun Qi. Deepchrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, 32(17):i639–i648, 2016.

[61] Carson Holt and Mark Yandell. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, 12(1):1–14, 2011.

[62] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60, 2015.

[63] Stijn Van Dongen. Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications*, 30(1):121–141, 2008.

[64] Riccardo Dondi, Manuel Lafond, and Celine Scornavacca. Reconciling multiple genes trees via segmental duplications and losses. *Algorithms for Molecular Biology*, 14:1–19, 2019.

[65] Mattéo Delabre, Nadia El-Mabrouk, Katharina T Huber, Manuel Lafond, Vincent Moulton, Emmanuel Noutahi, and Miguel Sautie Castellanos. Evolution through segmental duplications and losses: a super-reconciliation approach. *Algorithms for Molecular Biology*, 15:1–15, 2020.

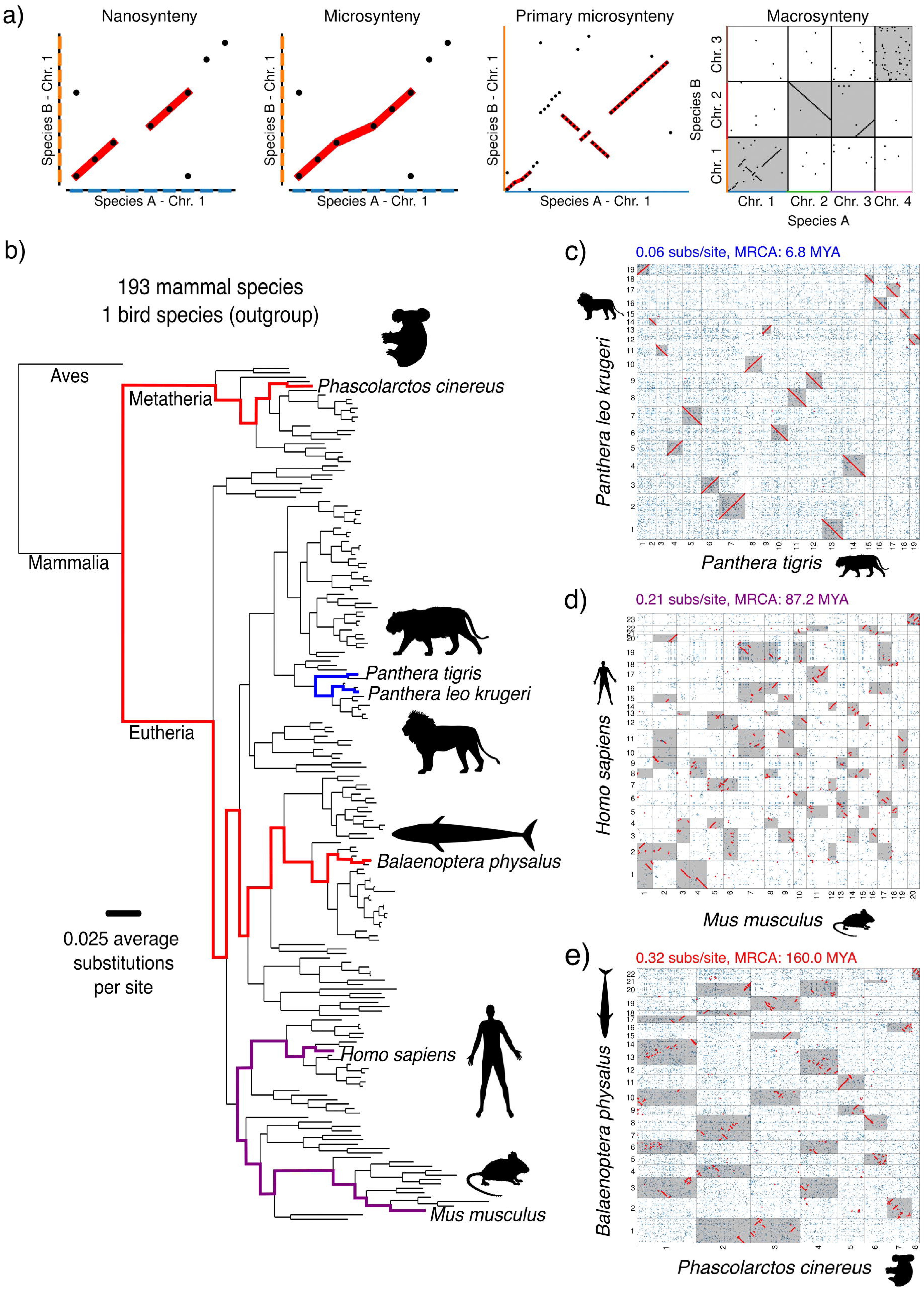
[66] Sudhir Kumar, Michael Suleski, Jack M Craig, Adrienne E Kasprowicz, Maxwell Sanderford, Michael Li, Glen Stecher, and S Blair Hedges. TimeTree 5: an expanded resource for species divergence times. *Molecular Biology and Evolution*, 39(8):msac174, 2022.

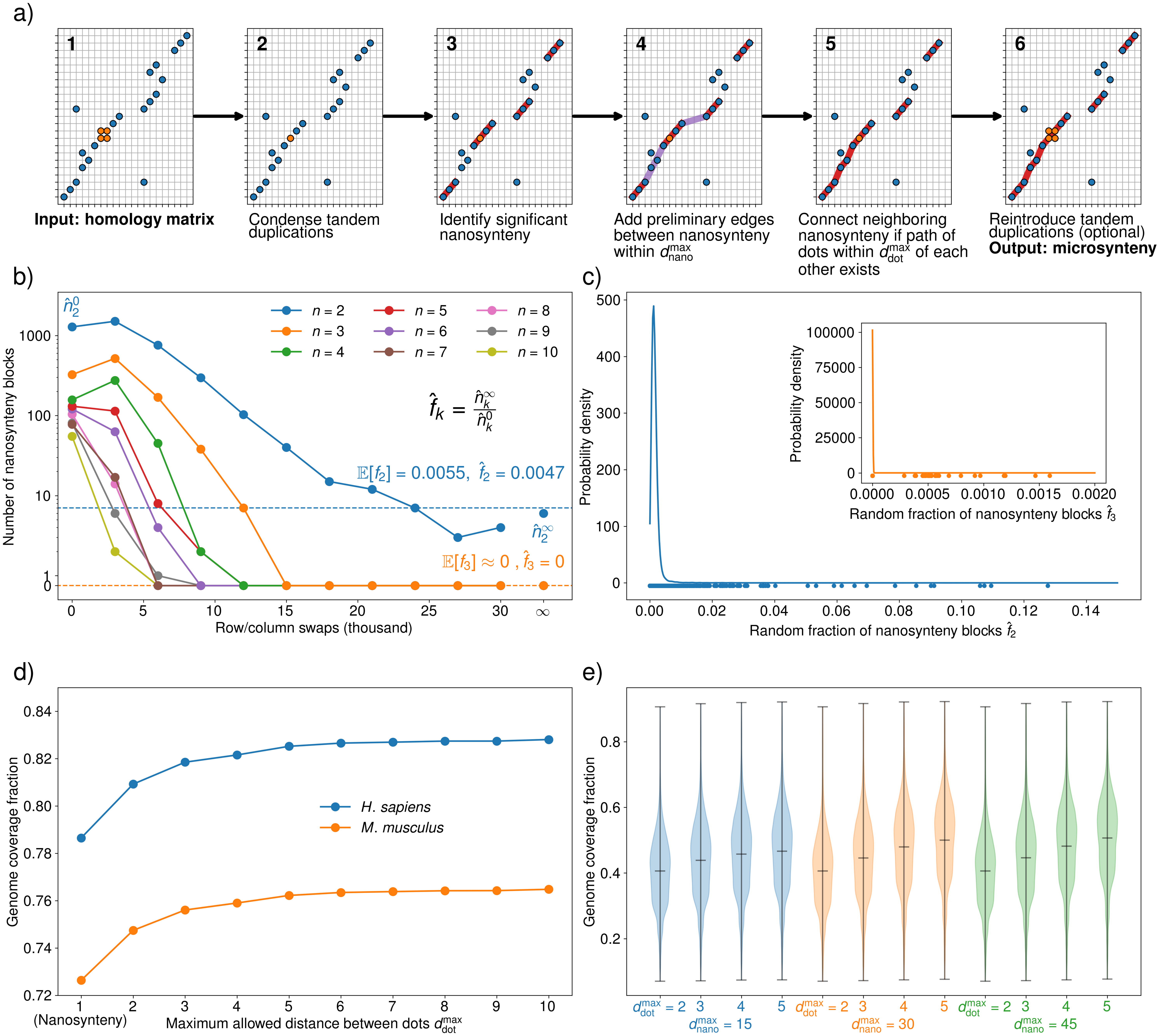
[67] ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57, 2012.

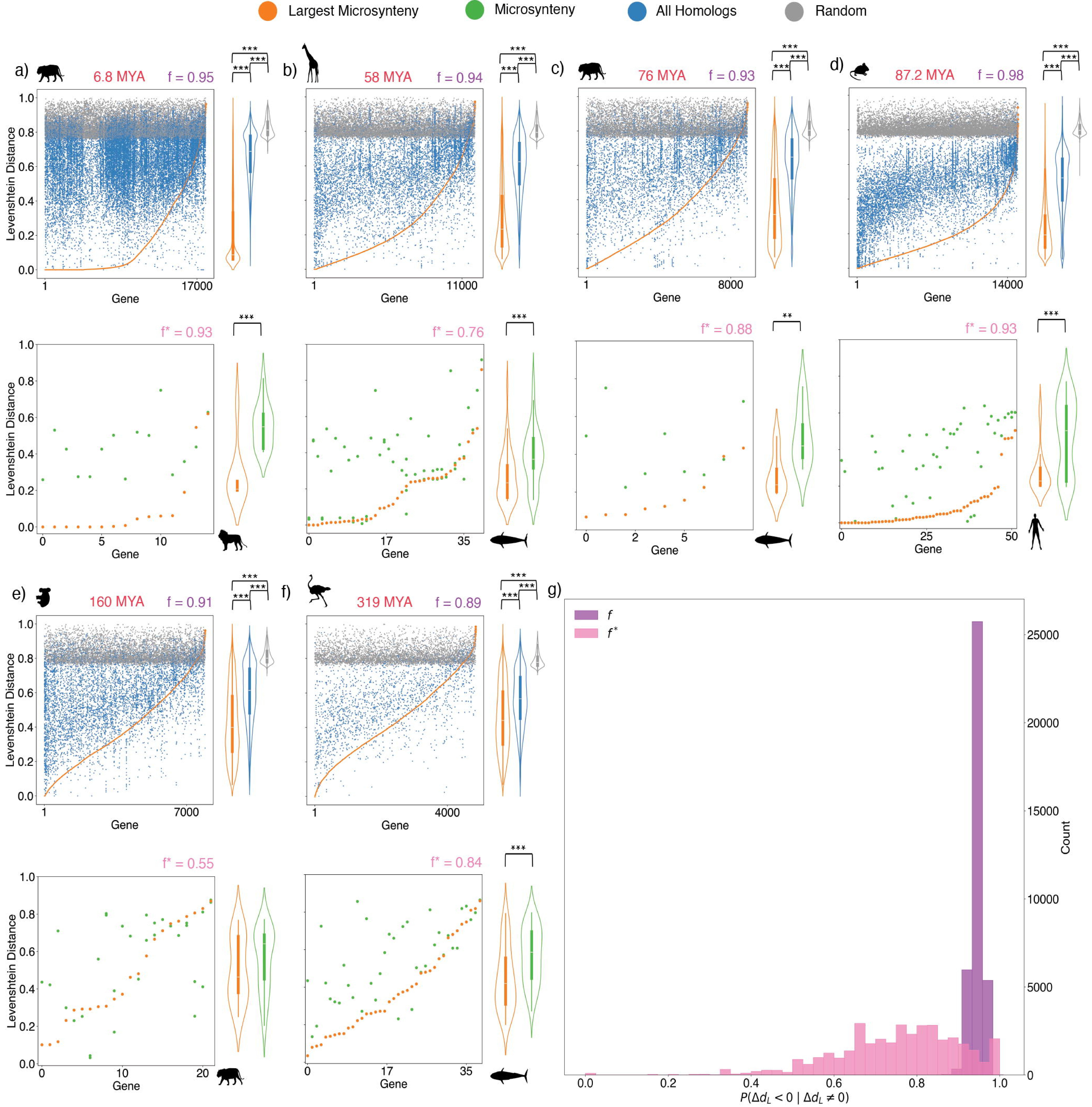
[68] Yunhai Luo, Benjamin C Hitz, Idan Gabdank, Jason A Hilton, Meenakshi S Kagda, Bonita Lam, Zachary Myers, Paul Sud, Jennifer Jou, Khine Lin, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Research*, 48(D1):D882–D889, 2020.

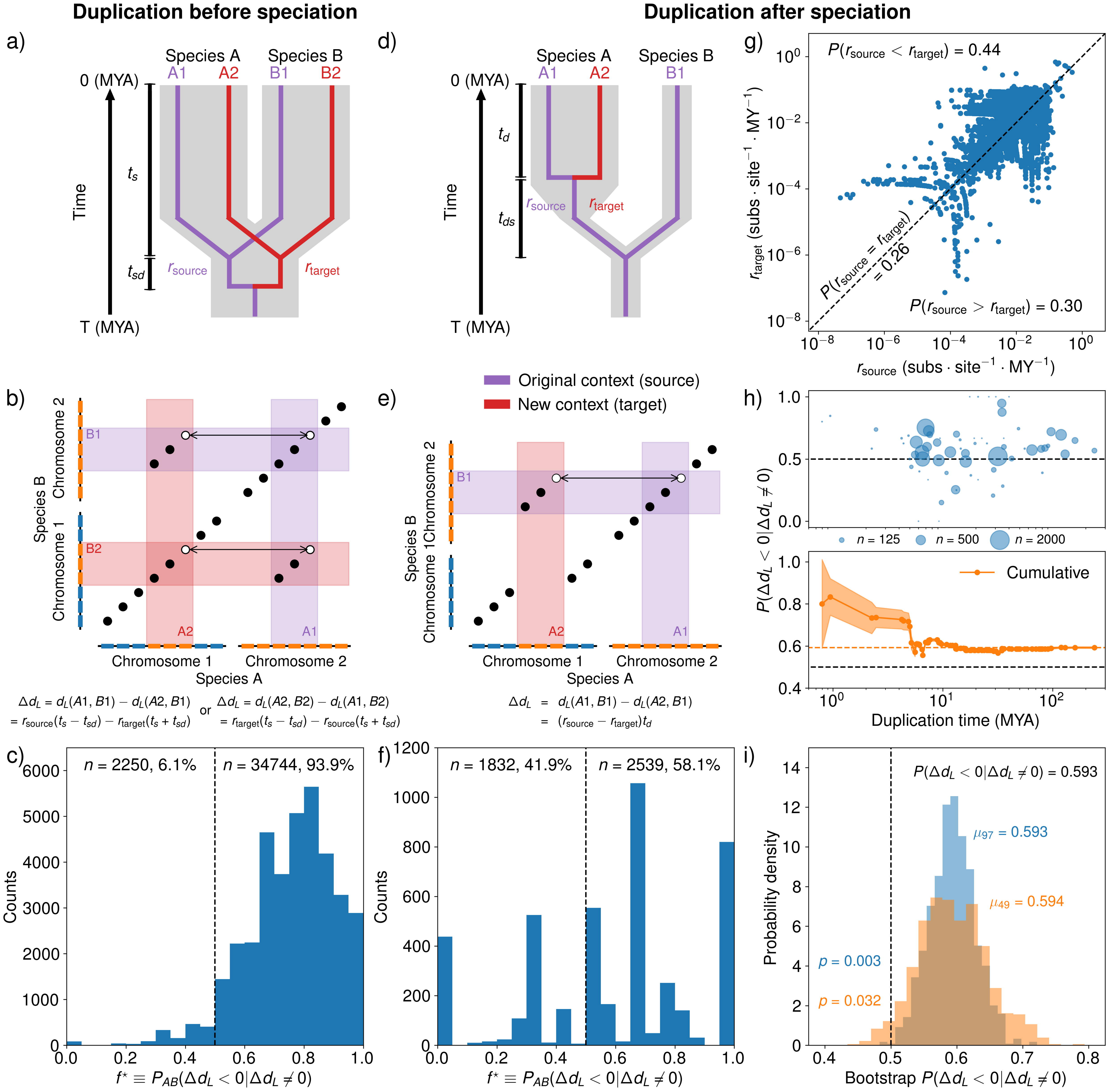
[69] Benjamin C Hitz, Lee Jin-Wook, Otto Jolanki, Meenakshi S Kagda, Keenan Graham, Paul Sud, Idan Gabdank, J Seth Strattan, Cricket A Sloan, Timothy Dreszer, et al. The ENCODE uniform analysis pipelines. *bioRxiv*, 2023.

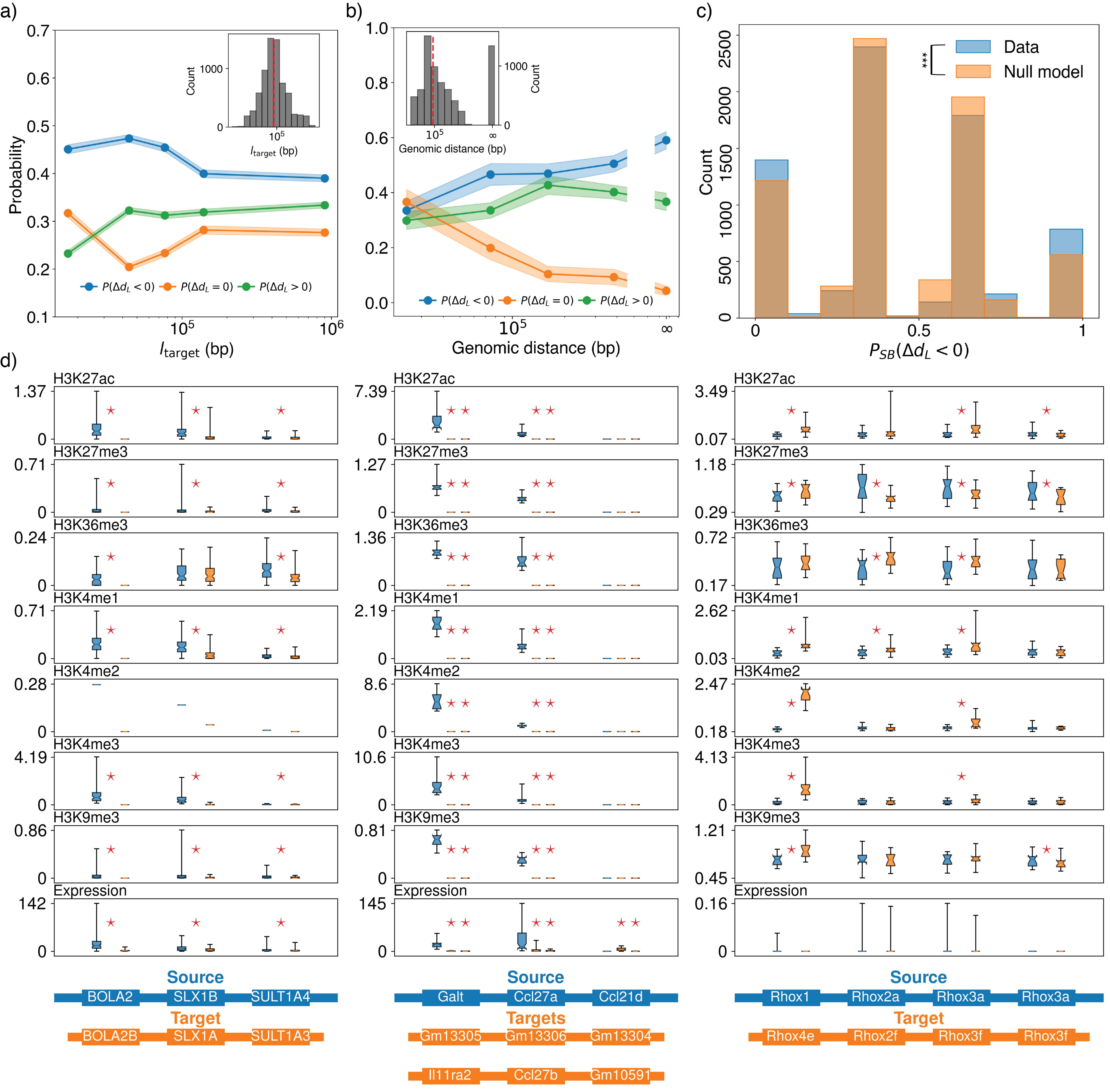
[70] Sean Upchurch, Emilio Palumbo, Jeremy Adams, David Bujold, Guillaume Bourque, Jared Nedzel, Keenan Graham, Meenakshi S Kagda, Pedro Assis, Benjamin Hitz, et al. RNAget: an API to securely retrieve RNA quantifications. *Bioinformatics*, 39(4):btad126, 2023.


a)
Nanosynteny
Microsynteny
Primary microsynteny
Macrosynteny
Species B - Chr. 1
Species A - Chr. 1
Species B
Chr. 1
Chr. 2
Chr. 3
Chr. 4
Species A
b)
193 mammal species
1 bird species (outgroup)
Aves
Metatheria
Mammalia
Eutheria
Phascolarctos cinereus
Panthera tigris
Panthera leo krugeri
Balaenoptera physalus
Homo sapiens
Mus musculus
0.025 average substitutions per site
c)
0.06 subs/site, MRCA: 6.8 MYA
Panthera leo krugeri
Panthera tigris
d)
0.21 subs/site, MRCA: 87.2 MYA
Homo sapiens
Mus musculus
e)
0.32 subs/site, MRCA: 160.0 MYA
Balaenoptera physalus
Phascolarctos cinereus


a)
1
2
3
4
5
6
Input: homology matrix
Condense tandem duplications
Identify significant nanosynteny
Add preliminary edges between nanosynteny within $d_{\mathrm{nano}}^{\max}$
Connect neighboring nanosynteny if path of dots within $d_{\mathrm{dot}}^{\max}$ of each other exists
Reintroduce tandem duplications (optional)
Output: microsynteny
b)
$\hat{n}_2^0$
$n = 2$
$n = 3$
$n = 4$
$n = 5$
$n = 6$
$n = 7$
$n = 8$
$n = 9$
$n = 10$
$\hat{f}_k = \frac{\hat{n}_k^\infty}{\hat{n}_k^0}$
$\mathbb{E}[f_2] = 0.0055,\ \hat{f}_2 = 0.0047$
$\hat{n}_2^\infty$
$\mathbb{E}[f_3] \approx 0\ ,\ \hat{f}_3 = 0$
Number of nanosynteny blocks
1000
100
10
1
0
0
5
10
15
20
25
30
∞
Row/column swaps (thousand)
c)
Probability density
500
400
300
200
100
0
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Random fraction of nanosynteny blocks $\hat{f}_2$
Probability density
100000
75000
50000
25000
0
0.0000
0.0005
0.0010
0.0015
0.0020
Random fraction of nanosynteny blocks $\hat{f}_3$
d)
Genome coverage fraction
0.84
0.82
0.80
0.78
0.76
0.74
0.72
H. sapiens
M. musculus
1
(Nanosynteny)
2
3
4
5
6
7
8
9
10
Maximum allowed distance between dots $d_{\mathrm{dot}}^{\max}$
e)
Genome coverage fraction
0.8
0.6
0.4
0.2
$d_{\mathrm{dot}}^{\max} = 2$
3
4
5
$d_{\mathrm{nano}}^{\max} = 15$
$d_{\mathrm{dot}}^{\max} = 2$
3
4
5
$d_{\mathrm{nano}}^{\max} = 30$
$d_{\mathrm{dot}}^{\max} = 2$
3
4
5
$d_{\mathrm{nano}}^{\max} = 45$


Largest Microsynteny
Microsynteny
All Homologs
Random
a)
6.8 MYA
f = 0.95
b)
58 MYA
f = 0.94
c)
76 MYA
f = 0.93
d)
87.2 MYA
f = 0.98
Levenshtein Distance
Gene
1
17000
11000
8000
14000
***
**
f* = 0.93
f* = 0.76
f* = 0.88
f* = 0.93
e)
160 MYA
f = 0.91
f)
319 MYA
f = 0.89
7000
4000
f* = 0.55
f* = 0.84
g)
f
f^*
Count
$P(\Delta d_L < 0 \mid \Delta d_L \neq 0)$
0.0
0.2
0.4
0.6
0.8
1.0
5000
10000
15000
20000
25000


Duplication before speciation
Duplication after speciation
a)
Species A
Species B
A1
A2
B1
B2
0 (MYA)
T (MYA)
Time
t_s
t_{sd}
r_{source}
r_{target}
d)
Species A
Species B
A1
A2
B1
0 (MYA)
T (MYA)
Time
t_d
t_{ds}
r_{source}
r_{target}
g)
$P(r_{\text{source}} < r_{\text{target}}) = 0.44$
$P(r_{\text{source}} = r_{\text{target}}) = 0.26$
$P(r_{\text{source}} > r_{\text{target}}) = 0.30$
r_{target} (subs · site^{-1} · MY^{-1})
r_{source} (subs · site^{-1} · MY^{-1})
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}
b)
Species B
Chromosome 2
Chromosome 1
B1
B2
A2
A1
Chromosome 1
Chromosome 2
Species A
$\Delta d_L = d_L(A1, B1) - d_L(A2, B1)$
$= r_{\text{source}}(t_s - t_{sd}) - r_{\text{target}}(t_s + t_{sd})$
or
$\Delta d_L = d_L(A2, B2) - d_L(A1, B2)$
$= r_{\text{target}}(t_s - t_{sd}) - r_{\text{source}}(t_s + t_{sd})$
e)
Original context (source)
New context (target)
Species B
Chromosome 1
Chromosome 2
B1
A2
A1
Chromosome 1
Chromosome 2
Species A
$\Delta d_L = d_L(A1, B1) - d_L(A2, B1)$
$= (r_{\text{source}} - r_{\text{target}})t_d$
h)
$P(\Delta d_L < 0|\Delta d_L \neq 0)$
1.0
0.5
0.0
n = 125
n = 500
n = 2000
1.0
0.8
0.6
0.4
Cumulative
10^{0}
10^{1}
10^{2}
Duplication time (MYA)
c)
n = 2250, 6.1%
n = 34744, 93.9%
Counts
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0
$f^\star \equiv P_{AB}(\Delta d_L < 0|\Delta d_L \neq 0)$
f)
n = 1832, 41.9%
n = 2539, 58.1%
Counts
1200
1000
800
600
400
200
0
0.0
0.2
0.4
0.6
0.8
1.0
$f^\star \equiv P_{AB}(\Delta d_L < 0|\Delta d_L \neq 0)$
i)
$P(\Delta d_L < 0|\Delta d_L \neq 0) = 0.593$
$\mu_{97} = 0.593$
$\mu_{49} = 0.594$
$p = 0.003$
$p = 0.032$
Probability density
14
12
10
8
6
4
2
0
0.4
0.5
0.6
0.7
0.8
Bootstrap $P(\Delta d_L < 0|\Delta d_L \neq 0)$


a)
Probability
0.7
0.6
0.5
0.4
0.3
0.2
0.1
10^5
10^6
l_{target} (bp)
Count
1000
0
10^5
l_{target} (bp)
$P(\Delta d_L < 0)$
$P(\Delta d_L = 0)$
$P(\Delta d_L > 0)$
b)
1.0
0.8
0.6
0.4
0.2
0.0
10^5
∞
Genomic distance (bp)
1000
Count
Genomic distance (bp)
c)
Count
2500
2000
1500
1000
500
0
0.5
1
$P_{SB}(\Delta d_L < 0)$
***
Data
Null model
d)
H3K27ac
H3K27me3
H3K36me3
H3K4me1
H3K4me2
H3K4me3
H3K9me3
Expression
1.37
0.71
0.24
0.71
0.28
4.19
0.86
142
7.39
1.27
1.36
2.19
8.6
10.6
0.81
145
3.49
0.07
1.18
0.29
0.72
0.17
2.62
0.03
2.47
0.18
4.13
1.21
0.45
0.16
Source
BOLA2
SLX1B
SULT1A4
Target
BOLA2B
SLX1A
SULT1A3
Source
Galt
Ccl27a
Ccl21d
Targets
Gm13305
Gm13306
Gm13304
Il11ra2
Ccl27b
Gm10591
Source
Rhox1
Rhox2a
Rhox3a
Rhox3a
Target
Rhox4e
Rhox2f
Rhox3f
Rhox3f