Evidence for a common evolutionary rate in metazoan transcriptional networks

Anne-Ruxandra Carvunis, Tina Wang, Dylan Skola, Alice Yu, Jonathan Chen, Jason F Kreisberg, Trey Ideker

DOI: http://dx.doi.org/10.7554/eLife.11615

Cite as: eLife 2015;10.7554/eLife.11615

Received: 15 September 2015
Accepted: 17 December 2015
Published: 18 December 2015

Title: Evidence for a common evolutionary rate in metazoan transcriptional networks

Authors: Anne-Ruxandra Carvunis*, Tina Wang*, Dylan Skola*, Alice Yu, Jonathan Chen, Jason F. Kreisberg, Trey Ideker

Affiliations:

Department of Medicine, University of California, San Diego, La Jolla, 92093-0688, USA

* These authors contributed equally.

Correspondence to: tideker@ucsd.edu

Abstract:

Genome sequences diverge more rapidly in mammals than in other animal lineages such as birds or insects. However, the effect of this rapid divergence on transcriptional evolution remains unclear. Recent reports have indicated a faster divergence of transcription factor binding in mammals than in insects, but others found the reverse for mRNA expression. Here, we show that these conflicting interpretations resulted from differing methodologies. We performed an integrated analysis of transcriptional network evolution by examining mRNA expression, transcription factor binding and *cis*-regulatory motifs across >25 animal species including mammals, birds and insects. Strikingly, we found that transcriptional networks evolve at a common rate across the three animal lineages. Furthermore, differences in rates of genome divergence were greatly reduced when restricting comparisons to chromatin-accessible sequences. The evolution of transcription is thus decoupled from the global rate of genome sequence evolution, suggesting that a small fraction of the genome regulates transcription.

Introduction:

A long-standing question in biology is what fraction of the genome regulates transcription (1-4). Recent studies of chromatin structure have implicated half of the human genome in regulatory interactions (1). Comparative genomic studies, however, have shown that less than 10% of the human genome is evolutionarily conserved (5), suggesting that many of the experimentally-detected interactions are not functional (2). Recent studies have measured the association between sequence changes and changes in transcript levels, epigenetic modifications or binding of transcription factors regulating specific gene sets (gene-specific transcription factors, GSTF) (6-15). These experiments demonstrated that genomic sequences can influence transcription even in the absence of evolutionary conservation. For instance, some repetitive elements previously thought to be "junk" DNA have been shown to effectively regulate gene expression (16). The rapid evolution of repetitive and other rapidly-evolving sequences could cause pervasive rewiring of transcriptional networks through creation and destruction of regulatory motifs (11). Such rapid transcriptional evolution would set mammals apart from other metazoans like birds or insects, whose genomes contain far fewer repetitive elements (17) and tend to be more constrained (5, 18).

A few studies have attempted to assess whether transcriptional networks evolve more rapidly in mammals than in insects from the fruit fly genus *Drosophila*. These studies have reached conflicting conclusions. When examining the evolution of GSTF binding, chromatin immuno-precipitation (ChIP) studies in mammalian livers have generally described faster divergence rates than similar studies in fly embryos (11, 19). However, divergence rates were estimated with different analytical methods in the

different ChIP studies (**Supplementary File 1**) (11, 20). Another study found that gene expression levels may diverge at a slower rate in mammals than in flies, by comparing genome-wide correlations of mRNA abundances estimated by RNA sequencing (RNA-seq) for mammals but by a mixture of technologies for flies including microarrays (21). Although the inconsistencies between these conclusions may indicate that the evolution of transcriptional networks is fundamentally different in mammals and insects, they may also reflect a sensitivity of evolutionary rate estimations to technical methodology.

Here, we jointly examined the evolution of gene expression levels and the underlying genome-wide changes in GSTF binding and *cis*-regulatory sequences using consistent methodologies both within and across various animal lineages.

Results:

We assembled a comparative genomics platform encompassing >40 publicly available datasets spanning >25 organisms representative of the *Mammalia* (mammals), *Aves* (birds) and *Insecta* (insects) phylogenetic classes (**Figure 1 – figure supplement 1**). We designed a statistical framework to objectively compare the rates of divergence of these various datasets across lineages. In brief, an exponential model describing evolutionary divergence under a common, lineage-naïve rate was evaluated against a lineage-aware model, accounting for both statistical significance and effect size (**Figure 1; Materials and methods**). We assessed the power of this statistical framework using simulations and found that it could detect differences in divergence rates with high sensitivity (**Materials and methods; Figure 1 – figure supplement 2)**. As a baseline, we first performed a comparative analysis of the evolution of genome sequences. We randomly sampled

genomic segments from designated reference genomes: *Mus musculus domesticus* (C57BL/6) for mammals, *Gallus gallus* for birds and *Drosophila melanogaster* for insects. The rates at which genomic segments that retained homologs with the other species within each lineage accumulate nucleotide substitutions were then estimated and compared using our statistical framework. Segments retaining homologs displayed high sequence conservation across all three lineages, although our framework detected a slightly but significantly faster divergence in insects than in mammals or birds ($P < 0.05$; **Figure 2 – figure supplement 1**). Next, we compared the rates at which randomly sampled genomic segments lost homology with the other species within each lineage. We observed a much larger difference in evolutionary rates across lineages using this measure ($P < 0.05$; **Figure 2**; **Figure 1 – figure supplement 2**). For instance, after 100 million years (Myrs) of evolution, only ~30% of mammalian segments retained homology whereas >60% of bird and insect segments did. These findings recapitulated previous observations according to which genome sequences are less constrained in mammals than in insects (5) or birds (18)

We then studied the evolution of gene expression levels, using exclusively RNA-seq datasets. In mammals and birds these datasets were generated from adult livers; in insects, they were from whole bodies of adult female fruit flies (**Materials and methods; Figure 3 – source data 1**). After determining expression levels for each gene in each species using a common data processing pipeline, we correlated the expression levels of genes in the reference species with the expression levels of their one-to-one orthologs in all other species within the same lineage (**Materials and methods**). We found that correlations of gene expression levels decreased over time at similar rates that were

97 statistically indistinguishable: a lineage-naïve model describing the evolution of gene
98 expression levels under a common rate fitted the data as well as a lineage-aware model
99 (**Figure 3**). This result was robust to changes in correlation metrics or inclusion/exclusion
100 of poorly expressed genes (**Figure 3 – figure supplement 1**).

101 Several lines of evidence suggest that gene expression levels can remain relatively
102 stable even as the genomic locations bound by GSTFs change rapidly over time (12, 14,
103 22). Therefore, we next examined the evolution of GSTF binding patterns. We
104 considered all GSTFs that were profiled using ChIP followed by massively parallel
105 sequencing (ChIP-seq) in at least three related species, where separate ChIPs were
106 performed per species. GSTFs meeting these requirements were Twist and Giant in fruit
107 fly embryos, and CEBPA, FOXA1 and HNF4A in mammalian livers (**Materials and**
108 **methods**; **Figure 4 – source data 1; Supplementary File 1**). We aimed to measure
109 cross-species similarity in GSTF occupancy with a unified analytical method across all of
110 these datasets. Despite the widespread use of ChIP-seq, there is no consensus on the
111 appropriate analytical method (23). ChIP-seq analysis pipelines typically discretize
112 continuous occupancy profiles into a set of occupied segments ("peaks"), but this step
113 requires choosing a signal processing algorithm (a peak caller) and associated parameters
114 (**Figure 4a**). Further comparison of occupied segments across species requires additional
115 analytical choices (**Figure 4a**), some of which can strongly influence downstream
116 findings (20).

117 To explore the impact of these choices, we processed all ChIP-seq data using
118 systematic combinations of parameters representative of, and expanding from, previous
119 studies (**Supplementary File 1**) (24). In total, we executed 108 analytical pipelines to

compare divergence rates across 6 pairs of GSTFs (2 in insects each compared with 3 in mammals), the occupancy profiles of which were examined in 3 – 7 species per lineage (**Materials and methods**). The values of the estimated rates varied greatly from one combination of parameters to the next (**Figure 4b, c**). However, in the majority of cases (56 – 78% over the 6 comparisons), GSTF binding patterns diverged at statistically indistinguishable rates in mammals and insects (**Figure 4d; Figure 4 – source data 2**). Although the computed divergence rates were sensitive to technical methodology (**Figure 4 – figure supplement 1**), for a given method the results were generally similar across lineages for all of the five GSTFs investigated.

To substantiate these findings, we devised a method to compare genome-wide occupancy profiles at single-nucleotide resolution without discretization. We correlated occupancy profiles between pairs of species across all nucleotides where genomes aligned, after accounting for the differences in sequencing depth, read length and fragment size across datasets (**Materials and methods**). Again, we found indistinguishable divergence rates, regardless of which GSTF or lineage was examined (**Figure 4e**). After 100 Myrs of evolution, the correlation of GSTF occupancy profiles was 0.10 in mammals and 0.13 in insects. As a control, we also applied this method to CTCF, a pleiotropic DNA-binding protein that acts as chromatin insulator and looping factor (25). In mammals, patterns of DNA occupancy have been shown to be more conserved for CTCF than for GSTFs using unified analytical methods (26). In contrast, CTCF DNA occupancy was shown to diverge rapidly in insects, perhaps due to the existence of other insulator proteins (11, 27). Our analysis successfully recapitulated this

142 difference (**Figure 4f**), demonstrating that the common evolutionary rate observed among
143 GSTFs (**Figure 4e**) was not an artifact of our method for profile correlation.

144 The similarity of divergence rates observed across lineages for gene expression
145 levels (**Figure 3**) and GSTF binding patterns (**Figure 4**) was unexpected given the rapid
146 evolution of genomic sequences in mammals relative to insects (5) or birds (18) (**Figure**
147 **2**). We therefore further examined these trends at the level of *cis*-regulatory sequences.
148 First, we considered the DNA sequence motifs thought to be specifically recognized by
149 the mammalian and insect GSTFs included in the previous ChIP-seq analysis (**Figure 4**).
150 We identified locations with significant matches to these motifs throughout the genomes
151 of the reference species and estimated how frequently these loci retained the same motifs
152 relative to background expectations (**Materials and methods**). We found similar,
153 indistinguishable retention rates in mammals and insects (**Figure 5a**). Next, we studied
154 the evolution of a broader set of motifs corresponding to GSTFs shared between *M.*
155 *musculus* and *D. melanogaster*. We found that these motifs were retained at similar rates
156 across lineages relative to background expectations in 8 out of 12 cases (one example
157 shown in **Figure 5b**; all other cases in **Figure 5 – figure supplement 1**).

158 Most active *cis*-regulatory sequences are located in genomic regions with
159 accessible chromatin (28). A recent study showed that chromatin-accessible sequences
160 were significantly more conserved between human and mouse than expected by chance
161 (29). We expanded this analysis to a wide range of species by using chromatin-accessible
162 sequences identified by DNAse I hypersensitivity in *M. musculus* livers, *D. melanogaster*
163 embryos and *G. gallus* MSB-1 cells (**Materials and methods**). We performed the
164 segment sampling procedure described previously (**Figure 2**), after excluding genes and

promoter regions since they typically are highly conserved (**Materials and methods**). Whereas inaccessible segments lost homology much faster in mammals than in insects and birds ($P < 0.05$; **Figure 5c**), accessible segments retained homologs at more similar rates in the three lineages (**Figure 5d; Figure 5 – figure supplement 2**). We still detected statistically significant differences across lineages ($P < 0.05$), but the effect sizes were considerably smaller than for inaccessible segments. For instance, ~60% of segments retained homology after 100 Myrs in birds and insects, independently of accessibility, whereas ~50% of chromatin-accessible segments and only ~20% of inaccessible segments did so in mammals.

Discussion:

To our knowledge, the analyses presented here represent the most comprehensive study conducted to date on the evolution of transcriptional networks across animal lineages. By applying unified analytical methods to data from different lineages, we were able to glean novel insights into the evolution of transcription in animals. We observed that gene expression levels, GSTF binding patterns, regulatory motifs and chromatin-accessible sequences each diverged at rates that were similar across mammals, birds and insects. These unexpected results reconcile previously conflicting findings (11, 21), highlighting the importance of unified study methodologies and providing evidence for a common evolutionary rate in metazoan transcriptional networks.

Most functional genomics studies have focused on humans and model organisms such as *D. melanogaster* or *M. musculus*, which are distantly related to each other. However, data on closely related species, like that we collected in this study, is needed to

188 investigate the dynamics of molecular network evolution. Unfortunately such data
189 remains scarce, leading to important limitations of our work. We only investigated three
190 lineages and six to twelve organisms per lineage with non-uniform coverage over
191 evolutionary time. In addition, we only examined a small number of tissues for each
192 lineage and a total of five GSTFs (none in birds). The generalizability of our observations
193 thus remains to be further evaluated as more data becomes available. Despite these
194 limitations, our finding that transcriptional networks evolve at a common rate per year
195 across animal lineages was strikingly robust across data layers.

196 The underlying mechanisms responsible for this concordance of evolutionary
197 rates are unclear. Mammals, birds and insects exhibit wide differences in the features that
198 are traditionally associated with evolutionary rates, such as generation times and breeding
199 sizes. Populations with small breeding sizes, such as mammals, are thought to be more
200 prone to genetic drift (30). This theory accounts for the abundance of repetitive elements
201 and the rapid evolution of genomic sequences in mammals relative to insects, which have
202 much larger breeding sizes. If the same theoretical principles also governed the evolution
203 of transcriptional networks, we would have expected that transcription would evolve
204 more rapidly in mammals than in insects. Instead, our results show that the evolution of
205 transcriptional networks, whether slow (*e.g.*, transcript levels) or fast (*e.g.*, GSTF
206 binding), is decoupled from the lineage-specific features that govern genome sequence
207 evolution.

208 One potential model could be that repetitive and rapidly-evolving sequences,
209 which make up the majority of the mammalian genome (5, 17), play a negligible role in
210 the global regulation of gene expression. Rather, chromatin-accessible regions may

represent the only portion of the mammalian genome that effectively regulates transcription. We observed that chromatin-accessible regions diverge much more slowly than other non-coding sequences in mammals, consistent with previous findings (29). These differences in divergence rates, however, were not found in birds and insects. As a result, chromatin-accessible regions in mammals are conserved at levels similar to those in birds and insects, in contrast to the genome as a whole. According to this model, the similar rates of evolution of chromatin-accessible sequences would constrain the dynamics of transcriptional evolution to be similar across lineages. The regulatory potential of repetitive and other rapidly-evolving elements could be rendered functionally inconsequential by silencing, or could be concentrated on controlling the expression of genetic elements that we did not investigate such as non-coding RNAs or species-specific genes (31).

An alternative model could be that the sequences that control transcriptional regulation in birds and insects evolve particularly rapidly within otherwise stable genomes. In these organisms, transcriptional networks would diverge under the action of natural selection, through specific single nucleotide substitutions resulting in rapid compensatory turnover (32). In mammals, transcriptional networks would diverge in a largely neutral fashion entrained for instance by transposons (31). Similar rates of transcriptional divergence would be achieved across lineages through very different evolutionary processes.

Importantly, none of the aforementioned models account for the differences in generation times between lineages. Evolutionary changes occurring based on chronological time and not generation time has also been observed for many protein-

234 coding sequences. Observations such as these led to the molecular clock theory (33). The
235 mechanisms through which environmental forces entrain these chronological
236 evolutionary clocks remain to be elucidated (33).

Materials and Methods:

Genome and Annotation Sources. We downloaded genome sequences for organisms belonging to three metazoan lineages: mammals, birds and insects. The mammalian and insect genome sequences were downloaded from the UCSC Genome Bioinformatics website (34): mm9 for *Mus musculus domesticus*, rn5 for *Rattus norvegicus* and hg19 for *Homo sapiens*; dm3 for *Drosophila melanogaster*, droSim1 for *Drosophila simulans*, droEre2 for *Drosophila erecta*, droYak2 for *Drosophila yakuba*, droAna3 for *Drosophila ananassae* and dp4 for *Drosophila pseudoobscura*. Genomes for mice strains and species not available from the UCSC Genome Bioinformatics site (*M. musculus domesticus* (AJ), *M. musculus castaneus* and *M. spretus*) were downloaded from (19). We downloaded bird genome sequences from Ensembl version 80 BioMart (35): galGal4 for *Gallus gallus*, Turkey_2.01 for *Meleagris gallopavo*, taeGut3.2.4 for *Taeniopygia guttata* and FicAlb_1.4 for *Ficedula albicollis*. Protein-coding gene names and symbols along with associated transcripts sequences were obtained from FlyBase (36) for insect species (dmel-r5.46, dsim-r1.4, dere-r1.3, dyak-r1.3, dana-r1.3 and dpse-r2.30), from Ensembl version 80 BioMart for bird species and from Ensembl version 59 BioMart for mammalian species (35). For *M. spretus* and *M. musculus castaneus*, we used the same transcript annotations as for *M. musculus*. Within the genomes of our designated reference organisms (*M. musculus domesticus*, *G. gallus* and *D. melanogaster*), we defined promoters as 2 kb upstream of transcription start site and delineated intergenic regions as regions that did not overlap annotated genes or promoters. Chromatin accessibility tracks used in **Figure 5c-d** and **Figure 5 – figure supplement 2** were downloaded from the UCSC bioinformatics website (34) for *M. musculus domesticus* and

D. melanogaster and obtained from (37) for *G. gallus*. We restricted our analyses to the sequences or annotations in, or homologous to, the well defined chromosome scaffolds of the reference organism. Specific reference chromosomes analyzed are as follows: *G. gallus* (1-28, Z, W), *D. melanogaster* (2L, 2R, 3L, 3R, 4, X) and *M. musculus* (1-19, X, Y).

Homology and Evolutionary Relationships. We obtained orthology relationships between protein-coding genes using Ensembl COMPARA (38), matching the Ensembl versions used for protein coding genes for each species described above. These relationships were used in **Figure 3**, **Figure 3 – figure supplement 1**, **Figure 5b and Figure 5 – figure supplement 1**. Homology between genomic segments was assigned using the LiftOver tool (34), for all analyses presented in **Figures 2, 4 and 5** and associated figure supplements, with the exception of the nucleotide-resolution analysis of GSTF occupancy profiles presented in **Figure 4e-f**. We used pre-computed chain files from UCSC matching the genome versions listed above when chains were readily available (34). When chain files were not available, we built chain files to map the UCSC M. *musculus* C57BL/6 mm9 to the genomes of *M. musculus domesticus* AJ, *Mus musculus castaneus* and *Mus spretus*, as well as to map the Ensembl 80 galGal4 to the genomes of *M. gallopavo*, *F. albicollis* and *T. guttata* (**Figure 1 – figure supplement 1**). These chains were constructed by following the steps recommended by UCSC (**Supplementary File 2**) (http://genomewiki.ucsc.edu/index.php/Whole_genome_alignment_howto).

For the nucleotide-resolution analysis of GSTF occupancy profiles, we assigned homology relationships using the chain files, or, in the case of mice strains, using genome mapping tables from (19). We filtered the chain files to obtain one-to-one unambiguous mappings by retaining only highest scoring alignment for each position. These filtered mappings were then used to transfer data to from any organism onto the corresponding reference genome. Regions in the reference species genome lacking one-to-one unambiguous mappings were excluded from analysis.

To define evolutionary distances separating species in Myrs, we chose published estimates generated as homogenously as possible within each lineage using a combination of sequence alignments and fossil records. All distances between insect species were taken from (39); all distances between bird species were taken from (40); distances between mammalian species were taken from (19) and TimeTree (41).

Data Sources. For RNA-seq analyses (**Figure 3**; **Figure 3 – figure supplement 1**), sequencing data for the reference species corresponding to two experiments performed independently by different research groups, and, when possible, representing different genotypes, were downloaded from public repositories. For *M. musculus domesticus*, we used data from (42, 43), for *G. gallus* we used data from (44) and (45), for *D. melanogaster* we used data from (1, 46). Other species included were *M. musculus castaneus* (42), *M. spretus* (12), *R. norvegicus* (47), *H. sapiens* (1, 48), *G. gorilla (44)*, *D. simulans* (46), *D. yakuba* (46), *D. ananassae* (46), *D. pseudoobscura* (46), *M. gallopavo* (49), *A. platyrhynchos* (50) and *F. albicollis* (51). Specific accession numbers are listed in **Figure 3 – source data 1**.

For ChIP-seq analyses (**Figure 4**), we downloaded data for FOXA1 in *M. musculus domesticus* (C57BL/6) (19), *M. musculus domesticus* (AJ) (19), *M. musculus castaneus* (19), *M. spretus* (19) and *R. norvegicus* (19); HNF4A and CEBPA in *M. musculus domesticus* (C57BL/6) (19), *M. musculus domesticus* (AJ) (19), *M. musculus castaneus* (19), *M. spretus* (19), *R. norvegicus* (19), *H. sapiens (52)* and *C. familiaris (52)*; Twist in *D. melanogaster (53)*, *D. simulans (53)*, *D. erecta (53)*, *D. yakuba (53)*, *D. ananassae (53)* and *D. pseudoobscura (53)*; Giant *in D. melanogaster* (22, 54), *D. yakuba* (54) and *D. pseudoobscura (22)*. We also gathered data for CTCF in *M. musculus domesticus* (C57BL/6) (26), *R. norvegicus* (26), *H. sapiens* (26), *C. familiaris* (26), *D. melanogaster (27)*, *D. simulans (27)*, *D. yakuba (27)* and *D. pseudoobscura (27)*. Accession numbers corresponding to the specific experimental replicates and control samples are listed in **Figure 4 – source data 1**.

For motif analyses (**Figure 5a-b**; **Figure 5 – figure supplement 1**), we gathered known position-weight matrixes from the JASPAR database (55) and the Fly Factor survey (56). We focused on the motifs corresponding to Twist and Giant in *D. melanogaster*, to CEBPA, HNF4A and FOXA1 in *M. musculus domesticus*, and on a set of 12 other motifs corresponding to GSTFs conserved across mammals and insects. This set was constructed by downloading all Core A vertebrata motifs from JASPAR (55), identifying those corresponding to conserved GSTFs with one-to-one orthologs between *M. musculus domesticus* and *D. melanogaster* using COMPARA (38), and filtering the list down to those 12 instances where a position-weight matrix was also described in Fly Factor (56) and were not already analyzed.

Comparing evolutionary rates. We developed a statistical framework to compare evolutionary rates between lineages, and implemented it in R (57). This framework takes as inputs: measures of pairwise cross-species similarity (*e.g.,* correlation of gene expression or sequence conservation), pairwise cross-species evolutionary distances and lineage labels. Conceptually, the framework estimates both a statistical significance and an effect size to determine whether rates of evolutionary divergence are indistinguishable or different between lineages (**Figure 1**).

In practice, we model evolutionary divergence by an exponential decay in log-linear space. First, the `nls` in R function is applied to the log-transformed cross-species similarity data as a function of evolutionary distances to derive the following linear models:

- a lineage-naïve model that estimates a shared intercept and slope for all the data without specifying the lineage labels
- a lineage-aware model that estimates a shared intercept for all the data and lineage-specific slopes based on lineage labels
- lineage-specific models that estimate intercept and slope individually for each lineage

Second, an R function written in-house to handle `nls` model structures estimates the significance level of an ANOVA with a likelihood ratio test comparing the lineage-naïve and the lineage-aware model. Third, we define the effect size as the predicted absolute difference in similarity between lineage pairs after 100 Myrs of divergence as estimated from the lineage-specific models. We consider that the framework detected a difference

between evolutionary divergence rates when the significance level is <0.05 and the effect size is >5%.

We chose to use an exponential decay function because it is the simplest evolutionary model that fit all our input measures of cross-species similarity reasonably well. We chose to model the exponential decay in log-linear space because we noted that a simple exponential decay in linear space failed to capture the conservation observed between distant species (mouse versus human at 91 Myrs and dog at 97.4 Myrs) when analyzing the evolutionary dynamics of GSTF binding (**Figure 4**) and motif retention (**Figure 5**). We hypothesize that these data layers likely follow a more complex decay model, but we did not want to explore this with our current data set to avoid over-fitting.

The power of this statistical framework was assessed by simulating data for two lineages with measure of cross-species similarity decaying exponentially at different rates over time (**Figure 1 – figure supplement 2**). We fixed one lineage to decay at set rates: -0.007, -0.005 and -0.003. We fixed the second lineage to be faster by a range of given differences. Over 1,000 simulations, we sampled two values from a normal distribution centered on the expected values from the set exponential decay rates corresponding to the evolutionary distances shown in **Figure 4b,** with standard deviations set at 0.5% or 5%. Our framework detected an absolute rate difference of 0.001 in 39.3% of simulations and an absolute rate difference of 0.003 in 88.9% of simulations when the standard deviation was high (5%). When the standard deviation was low (0.5%), our framework detected an absolute rate difference of 0.001 in 25.7% of simulations and an absolute rate difference of 0.003 in 100% of simulations.

Gene expression evolutionary rates (related to Figure 3). Analysis of gene expression evolutionary rates was performed in four steps. First, we preprocessed the raw RNA sequencing data downloaded for public data sources. Second, we quantified the abundance of all annotated transcripts corresponding to protein-coding genes. Third, we estimated cross-species similarity by correlating transcript abundances at the genome-scale. Finally, we used these cross-species similarity estimates as input to our statistical framework to evaluate a common model against a lineage-aware model.

RNA sequencing data was first preprocessed using FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/) and Trimmomatic (58). In order to quantify transcript abundances, we then used the program Sailfish (59) to 1) build transcriptome indices for each species using the transcriptome sequences described above, using the parameters "-p 8 -k 20"; 2) quantify transcript abundance using the transcriptome indices with the parameters "-p 8 -l 'T=PE:O=><:S=U' " for samples with paired-end reads and "-p 8 –l 'T=SE:S=U' " for samples with single-end reads. The bias-corrected transcripts per million (TPM) abundances estimated by Sailfish were then summed over the transcripts corresponding to the same gene locus.

To estimate cross-species similarities in gene expression levels, for each lineage, we used R (57) to build a matrix containing the gene expression values for all the protein-coding genes of the reference organism and their one-to-one orthologs across other organisms within each lineage. We discarded instances where the abundance of a particular gene locus was less than or equal to 5 TPM. We then calculated the Spearman's rank correlation for the expression of all genes between the reference and all other organisms within each lineage and plotted these correlations as against the

evolutionary distance separating each organism pair (**Figure 3**). We also repeated the calculations using Kendall's rank correlation coefficient and Pearson's product-moment correlation on $\log_2$-transformed expression values (**Figure 3 – figure supplement 1a-b**). Finally, we calculated Spearman's correlations among all genes including those with less than 5 TPM (**Figure 3 – figure supplement 1c**). All these scenarios were evaluated using our statistical framework. None indicated that a lineage-aware model described the data better than a common model.

GSTF Occupancy – Segment-resolution (related to Figure 4a-d). The first step of all our occupancy analyses was to align the ChIP-seq reads to the corresponding genomes in order to obtain occupancy profiles (**Figure 4a**). For each accession (**Figure 4 – source data 1**), the sequencing reads were aligned to reference genomes using Bowtie2 version 2.2.4 (60) with the parameters "-very-sensitive -N 1." Reads containing the 'XS:' field (multi-mappers) were removed. Reads having the same start site were presumed to be PCR duplicates and removed using the "rmdup" command of SAMtools version 1.1 (61). The filtered reads were then converted to tagAlign format. The tagAlign files corresponding to CEBPA, HNF4A, FOXA1, Twist and Giant were then processed using 108 different segment-resolution methods and one nucleotide-resolution method; the tagAlign files corresponding to CTCF were only processed using the nucleotide-resolution method. The nucleotide-resolution method is described below and relates to **Figure 4e-f**.

The aim of our segment-resolution analyses was to examine how robust the evolution of GSTF binding patterns was across 108 different analysis pipelines (**Figure**

4a-d). We implemented all these pipelines, which follow the same general framework and differ only in the choice of 5 parameters, described and underlined below.

First, the occupancy profiles in the tagAlign files were discretized into candidate occupied segments using a peak caller algorithm that aims at identifying segments where the ChIP sample is enriched in reads relative to the control sample. We implemented two peak callers: MACS version 2 (M) (62) and SPP (S) (63).

The occupied segments were then selected from the candidate set using a quality filter: stringent (S), lenient (L) or asymmetric (A). When using MACS2 (62) as a peak caller, lenient segments were called using a p-value cutoff of 10^{-5} (default) and merged across replicates when available using the merge function in BEDTools (64). Stringent segments were called using a p-value cutoff of 10^{-22} and intersected across replicates when replicates were available. The intersection procedure, inspired from (19), used BEDTools (64) to implement the following two steps: 1) merge the two replicates 2) select the merged segments corresponding to at least one segment in each original replicate. When using SPP (63) as a peak caller, lenient segments were called using a q-value of 10^{-2} (default), and merged across replicates when available (64). Stringent segments were called by selecting all candidate segments assigned to the lowest possible q-value in the sample, then intersected across replicates when available using the same intersection procedure. The asymmetric quality filter, inspired by (20, 53), indicates that segments were called stringently in the reference species and leniently in the other organism.

The coordinates of the occupied segments called in the reference organism were projected onto the other organism's genome using the LiftOver tool from the UCSC

442 genome browser (34) and specifying a sequence similarity filter through the minMatch
443 parameter. We used 3 different minMatch thresholds: stringent (S: 0.95 default), lenient
444 (L: 0.5), and none (N: 0.001).

445 After cross-species coordinate projection, a reference subset was chosen to define
446 the set of reference-occupied segments that would be further analyzed. Three choices
447 were implemented: all reference-occupied segments independently of whether they map
448 to any other species (A); for each pair of species, only reference-occupied segments with
449 a homolog in the second species (P); only reference-occupied segments that had
450 homologs across all the other species considered within the lineage (S).

451 The projected coordinates of the reference subset were then overlapped with the
452 coordinates of the occupied segments in the other species using the intersect function in
453 BEDTools (64). The overlap requirement was either lenient (L; default parameter of 1
454 bp) or stringent (S; required a reciprocal overlap of half of the segments length: " -f 0.5 -
455 r").

456 We systematically executed all combinations of the aforementioned 2 peak
457 callers, 3 quality filters, 3 sequence similarity filters, 3 reference subsets, and 2 overlap
458 requirements, yielding a total of 108 pipelines. The output of each pipeline was the
459 fraction of reference subset segments that overlapped segments occupied in the others
460 species (*i.e.* segments retaining occupancy between the two species). This output was
461 used as a cross-species similarity measure for GSTF binding patterns. We analyzed these
462 similarity measures for 6 pairs of GSTFs (Twist and Giant were each compared to
463 FOXA1, CEBPA and HNF4A) using our statistical framework. Two GSTFs were
464 considered to diverge differently from each other over time when 1) the significance of

the test was less than 0.05 and 2) the effect size was greater than 5%. In summary we found that the choice of parameters greatly influenced what the evolutionary dynamics of a given GSTF looked like (**Figure 4b-c**) but that in general the rate of divergence of mammal and insects GSTFs were statistically indistinguishable (**Figure 4d**). The results of these tests for all GSTF pairs considered across 108 pipelines are reported in **Figure 4 – source data 2** and summarized as pie-charts in **Figure 4**. Observations about general trends of parameters and evolutionary divergence are further elaborated in **Figure 4 – figure supplement 1**.

As a control we also conducted an analysis between FOXA1 and CEBPA since FOXA1 lacks data past 20 Myrs of evolutionary divergence, whereas for all others GSTFs we have broader resolution across in the 100Myrs range. We applied the same statistical framework to the within-lineage comparison between FOXA1 and CEBPA and detected that FOXA1 evolves faster than CEBPA in 74/108 instances. We believe that most of these detected differences are artifacts because the conservation of binding patterns for FOXA1 and CEBPA is in fact highly correlated throughout all combinations of parameters when restricting analyses to data points up to 20 Myrs (Pearson's R = 0.96). We suspect that this type of artifact also affects the results of comparing FOXA1 with Twist or Giant (**Figure 4d**).

GSTF Occupancy – Nucleotide-resolution (related to Figure 4e-f). In order to compare occupancy profiles directly without discretizing them into occupied segments and unoccupied segments, we correlated sets of imputed fragment density vectors across species. The inputs to this method were the tagAlign files described above. To generate

these vectors we first estimated the mean fragment size using a method adapted from (63), whereby the mean fragment size is computed as the number of base pairs of offset between the positive and negative strands that maximizes the Pearson's correlation coefficient of their mapped read density. We used a modified approach that considered only the density of 5' read start sites on each strand, rather than the density of the entire read. The first peak of the cross-correlation values was identified by approximating the first derivative by the finite difference method, smoothing the derivative values with a Gaussian kernel of bandwidth 10, and identifying the first downward zero-crossing of the curve. This position was used as the estimated mean fragment size *L*. We created imputed fragments by extending each read start site by *L* base pairs in the 3' direction. We then calculated a fragment density vector for each chromosome as the number of such imputed fragments that overlap each genomic position. When multiple replicates were available, replicates were merged by adding the fragment density vectors.

In order to minimize bias introduced by the presence of unmappable regions, we implemented a masking scheme that adaptively normalizes each dataset depending on the read length and estimated fragment size of each sequencing run. First, all possible error-free reads of a given length were generated synthetically and aligned back to the genome using Bowtie2 2.2.4 with the following parameters: "-r -N 0 -D 0 -R 0 --dpad 0 --score-min `C,0,-1`". Any multi-mapping reads with the 'XS:' flag were removed and the 5' and 3'-most positions of the remaining read alignments recorded. The imputed fragment densities computed from the ChIP data were then normalized by dividing the density at each position by the fraction of positions within *L* base pairs upstream that were covered by the start site (5' for positive-strand density and 3' for negative-strand density) of a

uniquely-mapped genomic read. Positions with 0 uniquely-mappable read start sites within *L* base pairs upstream regions were excluded from further analysis.

In order to compare between species, we transferred data from query organisms to the reference genome using the one-to-one filtered chain files described previously, and calculated the Pearson's correlation between the concatenated chromosome vectors of reference and reference-mapped query data. The evolution of the correlation was modeled and compared using the statistical framework described above.

Genome sequence evolutionary rates (related to Figure 2 and Figure 5c-d). We calculated the percentage of randomly sampled segments retaining homology. Within the genomes of the reference species, we delineated the boundaries of the regions from which to sample: whole genome (**Figure 2; Figure 2 – figure supplement 1**), intergenic regions in accessible chromatin and intergenic regions in inaccessible chromatin (**Figure 5; Figure 5 – figure supplement 2**). We used the BEDTools shuffle (64) to randomize the locations of 5,000 segments of 75 bp length within the delineated boundaries using the option "-noOverlapping." The resulting 5,000 shuffled segments were then mapped across species using the LiftOver tool with minMatch parameter 0.001 (34). We then calculated the percentage of segments that were successfully mapped (*i.e.*, retained homology), excluding segments that mapped to a region longer than 1,000 bp. The entire simulation was repeated 20 times, starting each time with different sets of 5,000 segments. The percentages of segments retaining homology were recorded for each of the 20 simulations, and averaged for each pair of species. These averages were plotted and used as inputs for our statistical framework. Varying the minMatch parameter of the

LiftOver tool to 0.5 and segment length to 150 bp allowed us to verify that the observed trends were robust to sequence similarity thresholds and length sampled (**Figure 2 – figure supplement 2; Figure 5 – figure supplement 2)**.

Nucleotide substitution rate within retained genomic segments (related to Figure 2 – figure supplement 1). The nucleotide sequences of the genomic segments from **Figure 2** that retained enough homology to undergo a pairwise alignment were extracted using the getfasta function of BEDTools (64). These sequences were then pairwise aligned using EMBOSS suite's implementation of Smith-Waterman local alignment (65). Default values for gap open penalty (10), gap extend penalty (0.5) and scoring matrix (EDNAFULL) were used to dynamically choose the best local alignment between reference and query sequences. For each cross-species comparison, we calculated the average percent identity of the ungapped alignments of all the segments across 20 randomizations. This procedure yielded values similar to those described previously for the mouse / human (66) and *D. melanogaster* / *D. pseudoobscura* comparisons (67). The average percent identity of ungapped alignments were used as inputs for our statistical framework, revealing that a model that incorporates lineage labels significantly improved fit to the data relative to a common model ($P < 0.05$; **Figure 2 – figure supplement 1**).

Motif evolutionary rates (related to Figure 5a-b). Using the FIMO tool (68) in the MEME suite (69), the genomes of *D. melanogaster* and *M. musculus domesticus* were scanned for matches to experimentally-determined position-weight matrixes corresponding to the GSTFs of interest. Motif matches were called significant according

to the default threshold of FIMO, $P < 10^{-4}$. The genomic coordinates of significant motif matches were mapped to the other species within the same lineage using LiftOver (minMatch 0.001). The corresponding coordinates (*Mapped*) were then extended by 50 bp, and the resulting segments were scanned for motif occurrence (*Mappedwithmotif*). In order to estimate background expectation, we randomly shuffled the locations of the *Mapped* segments and scanned these shuffled segments for motifs (*ShuffledMappedwithmotif*). The percentage of motifs retained relative to background was calculated as:

$$F = \frac{Mappedwithmotif - ShuffledMappedwithmotif}{Mapped} * 100$$

The percentages F were then used as measures of cross-species similarity to estimate whether a lineage-aware model would describe the evolution of DNA binding motifs better than a common model (**Figure 5 – figure supplement 1**).

Acknowledgments:

This work was supported by the National Institutes of Health: R01 GM084279 and P50 GM085764 awarded to T.I., T32 GM008666 awarded to T.W. and the Pathway to Independence K99 GM108865 awarded to A-R.C. The authors are grateful to Drs. Pollard K.S., Wittkopp P.J., Burke M., Charloteaux B. and Coolon J.D. for review of the manuscript prior to submission and to the editors and reviewers for their help in improving the manuscript after submission.

Competing Interests Statement:

Authors declare no competing financial interests.

Figure Titles and Legends:

Figure 1. Statistical framework to evaluate differences in evolutionary rates of change. Throughout this study we frequently evaluated whether the rate of evolutionary divergence of a given layer of transcriptional regulation differs between lineages. Our approach is equivalent to asking: if the lineage labels were hidden, would one be able to tell that the data points correspond to several lineages or would they seem equally likely to belong to a common distribution? **a, b,** Depict an example of statistically indistinguishable evolutionary rates. Without lineage labels (**a**), the similarity data are modeled by an exponential decay as well as with lineage labels (**b**). Adding lineage labels does not significantly improve the fit. **c, d,** Depict an example of statistically different evolutionary rates. Adding lineage labels (**d**) significantly improves the fit of an exponential decay model over unlabeled data (**c**).

Figure 2. Genomic sequences evolve more rapidly in mammals than in birds and insects. The evolutionary retention of 5,000 randomly sampled 75 bp segments was averaged over 20 trials. Organisms compared to reference species are as follows: *M. musculus domesticus* (AJ), *M. musculus castaneus*, *M. spretus*, rat, guinea pig, rabbit, human, chimpanzee and dog for *Mammalia*; turkey, zebrafinch and flycatcher for *Aves*; *D. simulans*, *D. erecta*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. virilis*, *D. willistoni* and *D. Grimshawi* for *Insecta*. Colored dashed lines: lineage-specific exponential fits, here and in all following displays. The trends were robust to variations in segment length and sequence similarity filters (**Figure 2 – figure supplement 2**).

Figure 3. Gene expression levels diverge at a common rate in mammals, birds and insects. Gene expression levels were derived independently from two RNA-seq experiments for each reference species and then correlated against each other and against gene expression levels derived from individual experiments in other species within the same lineage. Black dashed line: lineage-naïve exponential fit of all the data, without differentiating the lineages, here and in all following displays. Organisms compared to reference species are as follows: *M. musculus castaneus*, *M. spretus*, rat, human and gorilla for *Mammalia*; turkey, duck and flycatcher for *Aves*; *D. simulans*, *D. yakuba*, *D. ananassae* and *D. pseudoobscura* for *Insecta*.

Figure 4. GSTF occupancy diverges at a common rate in mammals and insects. a, Estimating shared GSTF occupancy across species requires multiple parameter choices. This diagram summarizes the main steps involved in comparing GSTF-occupied segments across species, showing a representative sample of choices at each step (steps represented by purple shapes, specific choices by the first letter bolded). The detailed methods and specific choices illustrated here and implemented in panels **b – d** are described in **Materials and methods**. **b, c,** An example of different analytical choices leading to different results despite starting from the same underlying data. Organisms compared to reference species are as follows: *M. musculus domesticus* (AJ), *M. musculus castaneus*, *M. spretus*, rat, human and dog for *Mammalia*; *D. simulans*, *D. erecta*, *D. yakuba*, *D. ananassae* and *D. pseudoobscura* for *Insecta*. **d**, Most combinations of choices yield indistinguishable evolutionary rates of GSTF binding patterns across lineages. The comparison of Twist and CEBPA is enlarged to show the color labels corresponding to the statistical interpretation regarding relative evolutionary rates. **e,** A

genome-wide comparison of GSTF occupancy profiles at single-nucleotide resolution shows indistinguishable evolutionary rates for CEBPA, HNF4A and FOXA1 in mammals and for Twist and Giant in insects. PCC: Pearson correlation coefficient. **f,** CTCF occupancy is highly conserved in mammals**.** Transparent points and lines are identical as panel **e**. Hexagons correspond to cross-species correlations of CTCF occupancy at single-nucleotide resolution.

Figure 5. Regulatory sequences diverge at similar rates across lineages. a, The motifs for CEBPA, HNF4A and FOXA1 in mammals and for Twist and Giant in insects are retained at a common rate. Organisms compared to reference species are the same as **Figure 4**. **b,** The motifs for GSTFs shared in mammals and insects are retained at common rates. One example is shown here for the motifs corresponding to PHO (FBgn0002521) in *D. melanogaster* and YY1 (ENSMUSG00000021264) in *M. musculus*, which are orthologous GSTFs. Eleven other cases of motif evolution for shared GSTFs conserved in mammals and insects are shown in **Figure 5 – figure supplement 1**. Organisms compared to reference species are as in **Figure 4**. **c, d,** Chromatin-accessible sequences are retained at similar rates in mammals, birds and insects. Analyses were performed as in **Figure 2**, limiting sampling to the inaccessible (**c**) and accessible (**d**) portions of the intergenic regions. Organisms compared to reference species are the same as **Figure 2**. The trends were robust to variations in segment length and sequence similarity filters (**Figure 5 – figure supplement 2**).

Figure Supplements:

Figure 1 – figure supplement 1: Comparative genomics platform for studying transcriptional network evolution across three metazoan lineages.

Figure 1 – figure supplement 2: Power of the statistical framework to evaluate differences in evolutionary rates.

Figure 2 – figure supplement 1: Genomic segments retaining homologs are highly conserved at the nucleotide level.

Figure 2 – figure supplement 2: Retention of genomic segments is robust to changes in sampled region size and sequence identity threshold.

Figure 3 – figure supplement 1: The common evolutionary rate of gene expression levels presented in Figure 3 is robust to changes in correlation metrics or expression threshold.

Figure 4 - figure supplement 1: Measured GSTF binding divergence rates are influenced by parameter choices.

Figure 5 - figure supplement 1: Conservation of *cis*-regulatory motifs for GSTFs conserved across insects and mammals.

Figure 5 - figure supplement 2: Retention of intergenic genomic segments in accessible- and inaccessible- chromatin is robust to changes in sampled region size and sequence identity threshold.

Supplementary Files:

Supplementary File 1: Published ChIP-seq studies comparing binding locations of GSTFs in closely related metazoans used different technical methodologies to estimate divergence rates.

Supplementary File 2: Parameters used to build chain files among vertebrate genomes.

669

Source Data Files:

671 Figure 3 - source data 1: Accession numbers used in RNA-seq analyses.

672 Figure 4 - source data 1: Accession numbers used in ChIP-seq analyses.

673 Figure 4 - source data 2: 648 segment-based ChIP analyses.

References:

1. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74. PubMed PMID: 22955616. Pubmed Central PMCID: 3439153.
2. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. 2013;5(3):578-90. PubMed PMID: 23431001. Pubmed Central PMCID: 3622293.
3. Niu DK, Jiang L. Can ENCODE tell us how much junk DNA we carry in our genome? Biochem Biophys Res Commun. 2013 Jan 25;430(4):1340-3. PubMed PMID: 23268340.
4. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A. 2014 Apr 29;111(17):6131-8. PubMed PMID: 24753594. Pubmed Central PMCID: 4035993.
5. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005 Aug;15(8):1034-50. PubMed PMID: 16024819. Pubmed Central PMCID: 1182216.
6. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. Nat Rev Genet. 2009 Mar;10(3):184-94. PubMed PMID: 19223927.
7. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, et al. Identification of genetic variants that affect histone modifications in human cells. Science. 2013 Nov 8;342(6159):747-9. PubMed PMID: 24136359. Pubmed Central PMCID: 3947669.
8. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, et al. Extensive variation in chromatin states across humans. Science. 2013 Nov 8;342(6159):750-2. PubMed PMID: 24136358. Pubmed Central PMCID: 4075767.
9. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, et al. Variation in transcription factor binding among humans. Science. 2010 Apr 9;328(5975):232-5. PubMed PMID: 20299548. Pubmed Central PMCID: 2938768.
10. Heinz S, Romanoski CE, Benner C, Allison KA, Kaikkonen MU, Orozco LD, et al. Effect of natural genetic variation on enhancer selection and function. Nature. 2013 Nov 28;503(7477):487-92. PubMed PMID: 24121437. Pubmed Central PMCID: 3994126.
11. Villar D, Flicek P, Odom DT. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. Nat Rev Genet. 2014 Apr;15(4):221-33. PubMed PMID: 24590227. Pubmed Central PMCID: 4175440.
12. Wong ES, Thybert D, Schmitt BM, Stefflova K, Odom DT, Flicek P. Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals. Genome Res. 2015 Feb;25(2):167-78. PubMed PMID: 25394363. Pubmed Central PMCID: 4315291.
13. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science. 2002 Apr 26;296(5568):752-5. PubMed PMID: 11923494.
14. Chan ET, Quon GT, Chua G, Babak T, Trochesset M, Zirngibl RA, et al. Conservation of core gene expression in vertebrate tissues. J Biol. 2009;8(3):33. PubMed PMID: 19371447. Pubmed Central PMCID: 2689434.
15. Shibata Y, Sheffield NC, Fedrigo O, Babbitt CC, Wortham M, Tewari AK, et al. Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. PLoS Genet. 2012 Jun;8(6):e1002789. PubMed PMID: 22761590. Pubmed Central PMCID: 3386175.
16. Rebollo R, Romanish MT, Mager DL. Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet. 2012;46:21-42. PubMed PMID: 22905872.
17. Taft RJ, Pheasant M, Mattick JS. The relationship between non-protein-coding DNA and eukaryotic complexity. BioEssays : news and reviews in molecular, cellular and developmental biology. 2007 Mar;29(3):288-99. PubMed PMID: 17295292.
18. Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. Science. 2014 Dec 12;346(6215):1311-20. PubMed PMID: 25504712. Pubmed Central PMCID: 4390078.

19. Stefflova K, Thybert D, Wilson MD, Streeter I, Aleksic J, Karagianni P, et al. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. Cell. 2013 Aug 1;154(3):530-40. PubMed PMID: 23911320. Pubmed Central PMCID: 3732390.
20. Bardet AF, He Q, Zeitlinger J, Stark A. A computational pipeline for comparative ChIP-seq analyses. Nat Protoc. 2012 Jan;7(1):45-61. PubMed PMID: 22179591.
21. Coolon JD, McManus CJ, Stevenson KR, Graveley BR, Wittkopp PJ. Tempo and mode of regulatory evolution in Drosophila. Genome Res. 2014 May;24(5):797-808. PubMed PMID: 24567308. Pubmed Central PMCID: 4009609.
22. Paris M, Kaplan T, Li XY, Villalta JE, Lott SE, Eisen MB. Extensive divergence of transcription factor binding in Drosophila embryos with highly conserved gene expression. PLoS Genet. 2013;9(9):e1003748. PubMed PMID: 24068946. Pubmed Central PMCID: 3772039.
23. Wilbanks EG, Facciotti MT. Evaluation of algorithm performance in ChIP-seq peak detection. PLoS One. 2010;5(7):e11471. PubMed PMID: 20628599. Pubmed Central PMCID: 2900203.
24. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012 Sep;22(9):1813-31. PubMed PMID: 22955991. Pubmed Central PMCID: 3431496.
25. Ohlsson R, Lobanenkov V, Klenova E. Does CTCF mediate between nuclear organization and gene expression? BioEssays : news and reviews in molecular, cellular and developmental biology. 2010 Jan;32(1):37-50. PubMed PMID: WOS:000273506800007.
26. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, et al. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell. 2012 Jan 20;148(1-2):335-48. PubMed PMID: 22244452. Pubmed Central PMCID: 3368268.
27. Ni X, Zhang YE, Negre N, Chen S, Long M, White KP. Adaptive evolution and the birth of CTCF binding sites in the Drosophila genome. PLoS Biol. 2012;10(11):e1001420. PubMed PMID: 23139640. Pubmed Central PMCID: 3491045.
28. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods. 2009 Apr;6(4):283-9. PubMed PMID: 19305407. Pubmed Central PMCID: 2668528.
29. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014 Nov 20;515(7527):355-64. PubMed PMID: 25409824. Pubmed Central PMCID: 4266106.
30. Ohta T. The Nearly Neutral Theory of Molecular Evolution. Annu Rev Ecol Syst. 1992;23:263-86. PubMed PMID: WOS:A1992JZ28100011.
31. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 2014 Dec;24(12):1963-76. PubMed PMID: 25319995. Pubmed Central PMCID: 4248313.
32. He BZ, Holloway AK, Maerkl SJ, Kreitman M. Does positive selection drive transcription factor binding site turnover? A test with Drosophila cis-regulatory modules. PLoS Genet. 2011 Apr;7(4):e1002053. PubMed PMID: 21572512. Pubmed Central PMCID: 3084208.
33. Kumar S. Molecular clocks: four decades of evolution. Nat Rev Genet. 2005 Aug;6(8):654-62. PubMed PMID: 16136655.
34. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, et al. The UCSC Genome Browser database: 2015 update. Nucleic Acids Res. 2015 Jan;43(Database issue):D670-81. PubMed PMID: 25428374. Pubmed Central PMCID: 4383971.
35. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res. 2015 Jan;43(Database issue):D662-9. PubMed PMID: 25352552. Pubmed Central PMCID: 4383879.
36. dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, et al. FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Res. 2015 Jan;43(Database issue):D690-7. PubMed PMID: 25398896. Pubmed Central PMCID: 4383921.
37. He Y, Carrillo JA, Luo J, Ding Y, Tian F, Davidson I, et al. Genome-wide mapping of DNase I hypersensitive sites and association analysis with gene expression in MSB1 cells. Frontiers in genetics. 2014;5:308. PubMed PMID: 25352859. Pubmed Central PMCID: 4195362.

38. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 2009 Feb;19(2):327-35. PubMed PMID: 19029536. Pubmed Central PMCID: 2652215.
39. Tamura K, Subramanian S, Kumar S. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. Mol Biol Evol. 2004 Jan;21(1):36-44. PubMed PMID: 12949132.
40. Lu L, Chen Y, Wang Z, Li X, Chen W, Tao Z, et al. The goose genome sequence leads to insights into the evolution of waterfowl and susceptibility to fatty liver. Genome Biol. 2015;16:89. PubMed PMID: 25943208. Pubmed Central PMCID: 4419397.
41. Hedges SB. The Timetree of Life: Oxford University Press; 2009. 551 p.
42. Goncalves A, Leigh-Brown S, Thybert D, Stefflova K, Turro E, Flicek P, et al. Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. Genome Res. 2012 Dec;22(12):2376-84. PubMed PMID: 22919075. Pubmed Central PMCID: 3514667.
43. Sugathan A, Waxman DJ. Genome-wide analysis of chromatin states reveals distinct mechanisms of sex-dependent gene regulation in male and female mouse liver. Mol Cell Biol. 2013 Sep;33(18):3594-610. PubMed PMID: 23836885. Pubmed Central PMCID: 3753870.
44. Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011 Oct 20;478(7369):343-8. PubMed PMID: 22012392.
45. Coble DJ, Fleming D, Persia ME, Ashwell CM, Rothschild MF, Schmidt CJ, et al. RNA-seq analysis of broiler liver transcriptome reveals novel responses to high ambient temperature. BMC Genomics. 2014;15:1084. PubMed PMID: 25494716. Pubmed Central PMCID: 4299486.
46. Chen ZX, Sturgill D, Qu J, Jiang H, Park S, Boley N, et al. Comparative validation of the D. melanogaster modENCODE transcriptome annotation. Genome Res. 2014 Jul;24(7):1209-23. PubMed PMID: 24985915. Pubmed Central PMCID: 4079975.
47. Gong B, Wang C, Su Z, Hong H, Thierry-Mieg J, Thierry-Mieg D, et al. Transcriptomic profiling of rat liver samples in a comprehensive study design by RNA-Seq. Scientific data. 2014;1:140021. PubMed PMID: 25977778. Pubmed Central PMCID: 4322565.
48. Lin S, Lin Y, Nery JR, Urich MA, Breschi A, Davis CA, et al. Comparison of the transcriptional landscapes between human and mouse tissues. Proc Natl Acad Sci U S A. 2014 Dec 2;111(48):17224-9. PubMed PMID: 25413365. Pubmed Central PMCID: 4260565.
49. Monson MS, Settlage RE, McMahon KW, Mendoza KM, Rawal S, El-Nezami HS, et al. Response of the hepatic transcriptome to aflatoxin B1 in domestic turkey (Meleagris gallopavo). PLoS One. 2014;9(6):e100930. PubMed PMID: 24979717. Pubmed Central PMCID: 4076218.
50. Huang Y, Li Y, Burt DW, Chen H, Zhang Y, Qian W, et al. The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. Nat Genet. 2013 Jul;45(7):776-83. PubMed PMID: 23749191. Pubmed Central PMCID: 4003391.
51. Uebbing S, Kunstner A, Makinen H, Ellegren H. Transcriptome sequencing reveals the character of incomplete dosage compensation across multiple tissues in flycatchers. Genome Biol Evol. 2013;5(8):1555-66. PubMed PMID: 23925789. Pubmed Central PMCID: 3762201.
52. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science. 2010 May 21;328(5981):1036-40. PubMed PMID: 20378774. Pubmed Central PMCID: 3008766.
53. He Q, Bardet AF, Patton B, Purvis J, Johnston J, Paulson A, et al. High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. Nat Genet. 2011 May;43(5):414-20. PubMed PMID: 21478888.
54. Bradley RK, Li XY, Trapnell C, Davidson S, Pachter L, Chu HC, et al. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. PLoS Biol. 2010 Mar;8(3):e1000343. PubMed PMID: 20351773. Pubmed Central PMCID: 2843597.
55. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 2014 Jan;42(Database issue):D142-7. PubMed PMID: 24194598. Pubmed Central PMCID: 3965086.
56. Zhu LJ, Christensen RG, Kazemian M, Hull CJ, Enuameh MS, Basciotta MD, et al. FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the

bacterial one-hybrid system. Nucleic Acids Res. 2011 Jan;39(Database issue):D111-7. PubMed PMID: 21097781. Pubmed Central PMCID: 3013762.
57. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2011.
58. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20. PubMed PMID: 24695404. Pubmed Central PMCID: 4103590.
59. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014 May;32(5):462-4. PubMed PMID: 24752080. Pubmed Central PMCID: 4077321.
60. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Apr;9(4):357-9. PubMed PMID: 22388286. Pubmed Central PMCID: 3322381.
61. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. PubMed PMID: 19505943. Pubmed Central PMCID: 2723002.
62. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137. PubMed PMID: 18798982. Pubmed Central PMCID: 2592715.
63. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol. 2008 Dec;26(12):1351-9. PubMed PMID: 19029915. Pubmed Central PMCID: 2597701.
64. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841-2. PubMed PMID: 20110278. Pubmed Central PMCID: 2832824.
65. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000 Jun;16(6):276-7. PubMed PMID: 10827456.
66. Mouse Genome Sequencing C, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002 Dec 5;420(6915):520-62. PubMed PMID: 12466850.
67. Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, et al. Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. Genome Res. 2005 Jan;15(1):1-18. PubMed PMID: 15632085. Pubmed Central PMCID: 540289.
68. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011 Apr 1;27(7):1017-8. PubMed PMID: 21330290. Pubmed Central PMCID: 3065696.
69. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W202-8. PubMed PMID: 19458158. Pubmed Central PMCID: 2703892.

Figure 1

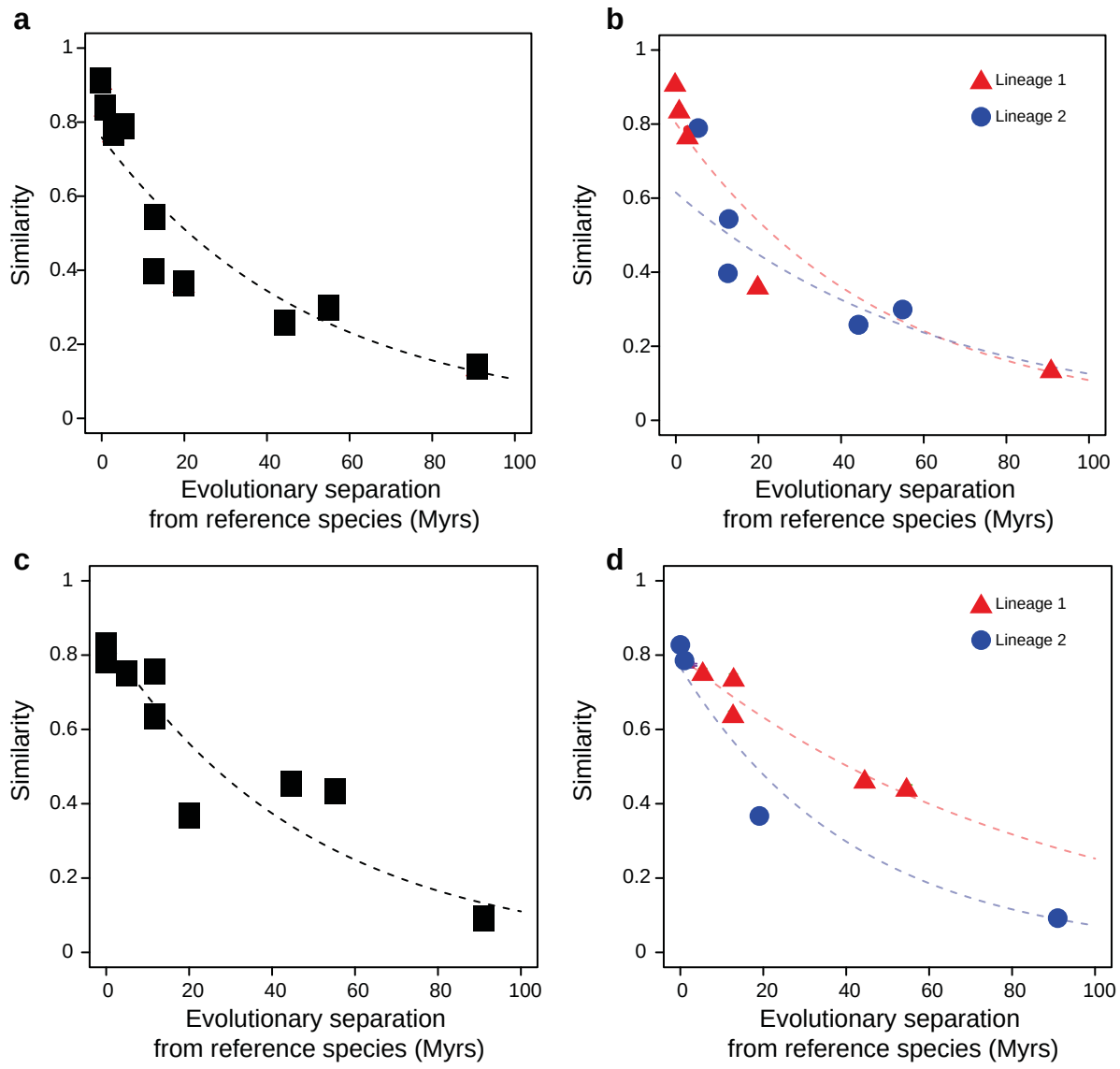

a
b
c
d
Similarity
Evolutionary separation
from reference species (Myrs)
Lineage 1
Lineage 2
0
0.2
0.4
0.6
0.8
1
0
20
40
60
80
100

Figure 2

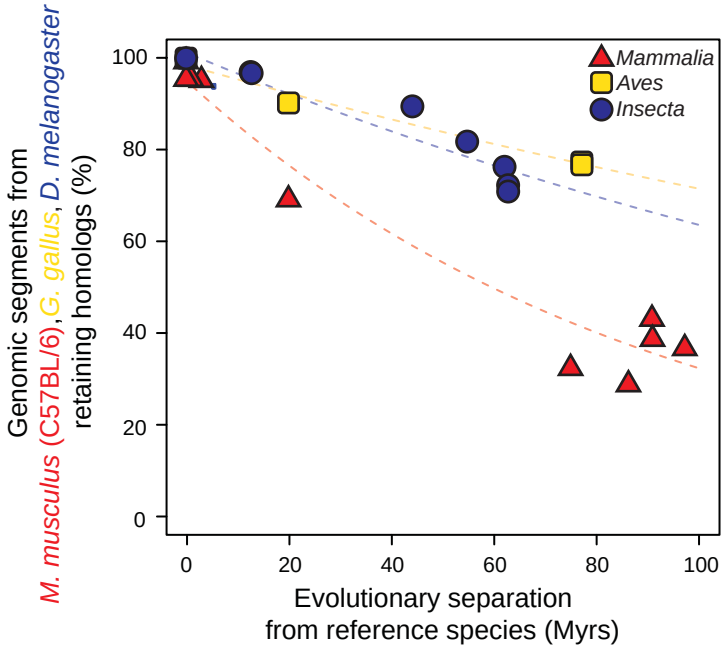

Genomic segments from
M. musculus (C57BL/6), G. gallus, D. melanogaster
retaining homologs (%)
100
80
60
40
20
0
0
20
40
60
80
100
Evolutionary separation
from reference species (Myrs)
Mammalia
Aves
Insecta

Figure 3

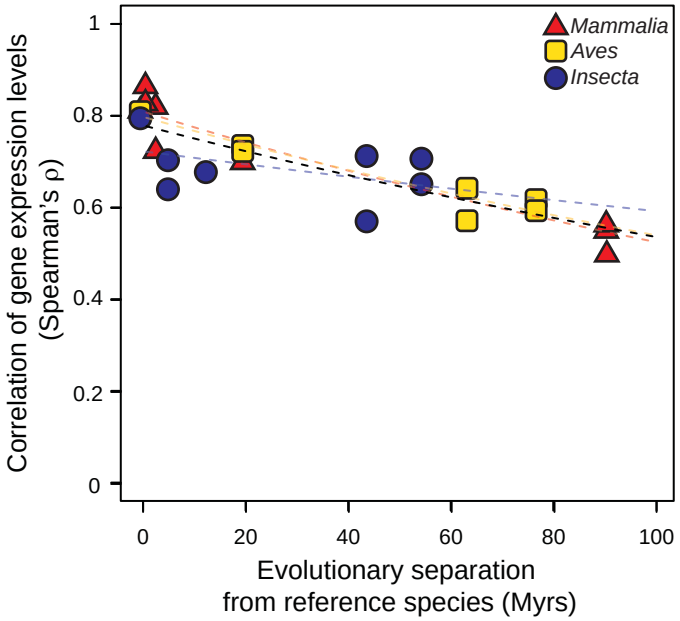

Mammalia
Aves
Insecta
Correlation of gene expression levels
(Spearman's ρ)
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
Evolutionary separation
from reference species (Myrs)

Figure 4

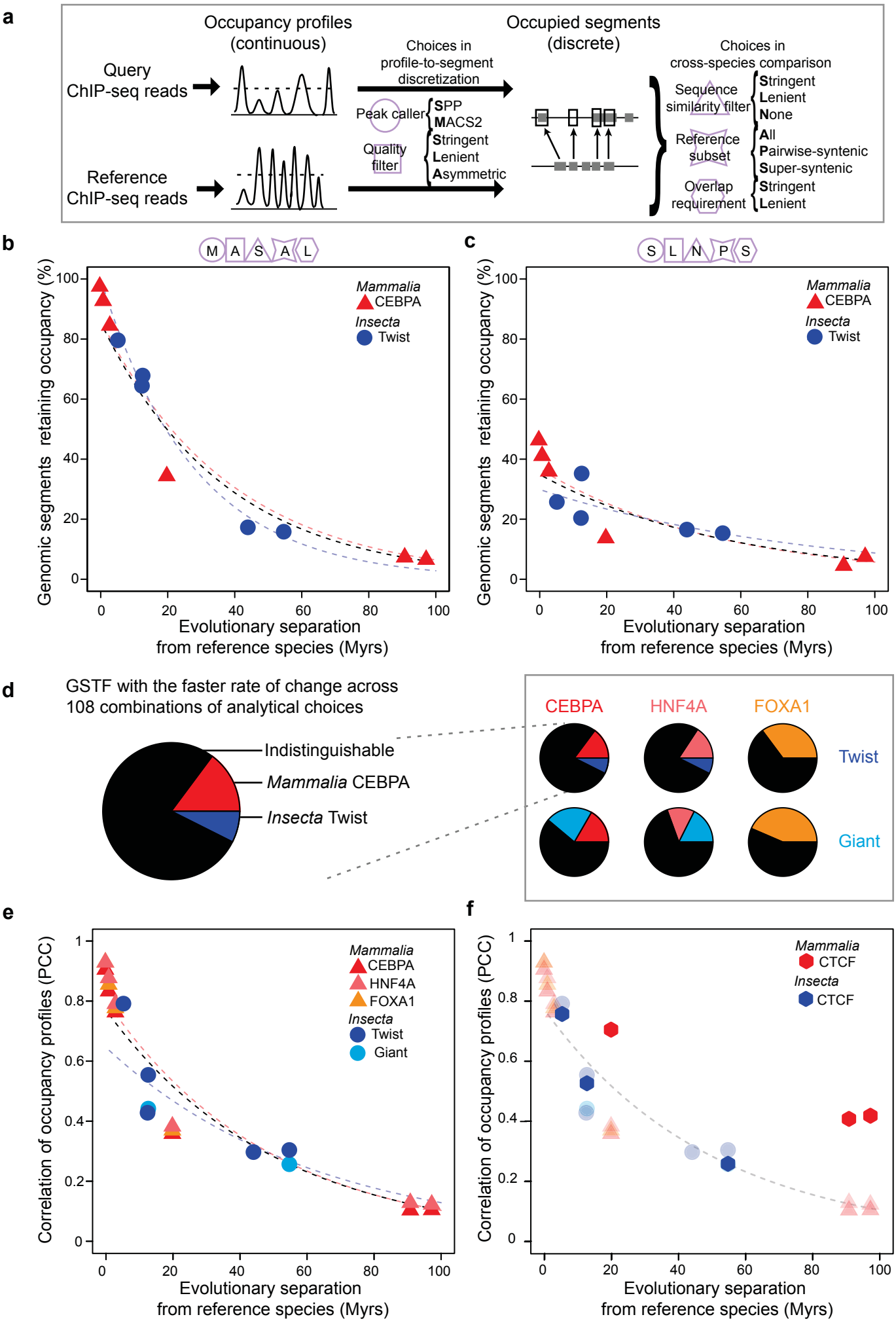
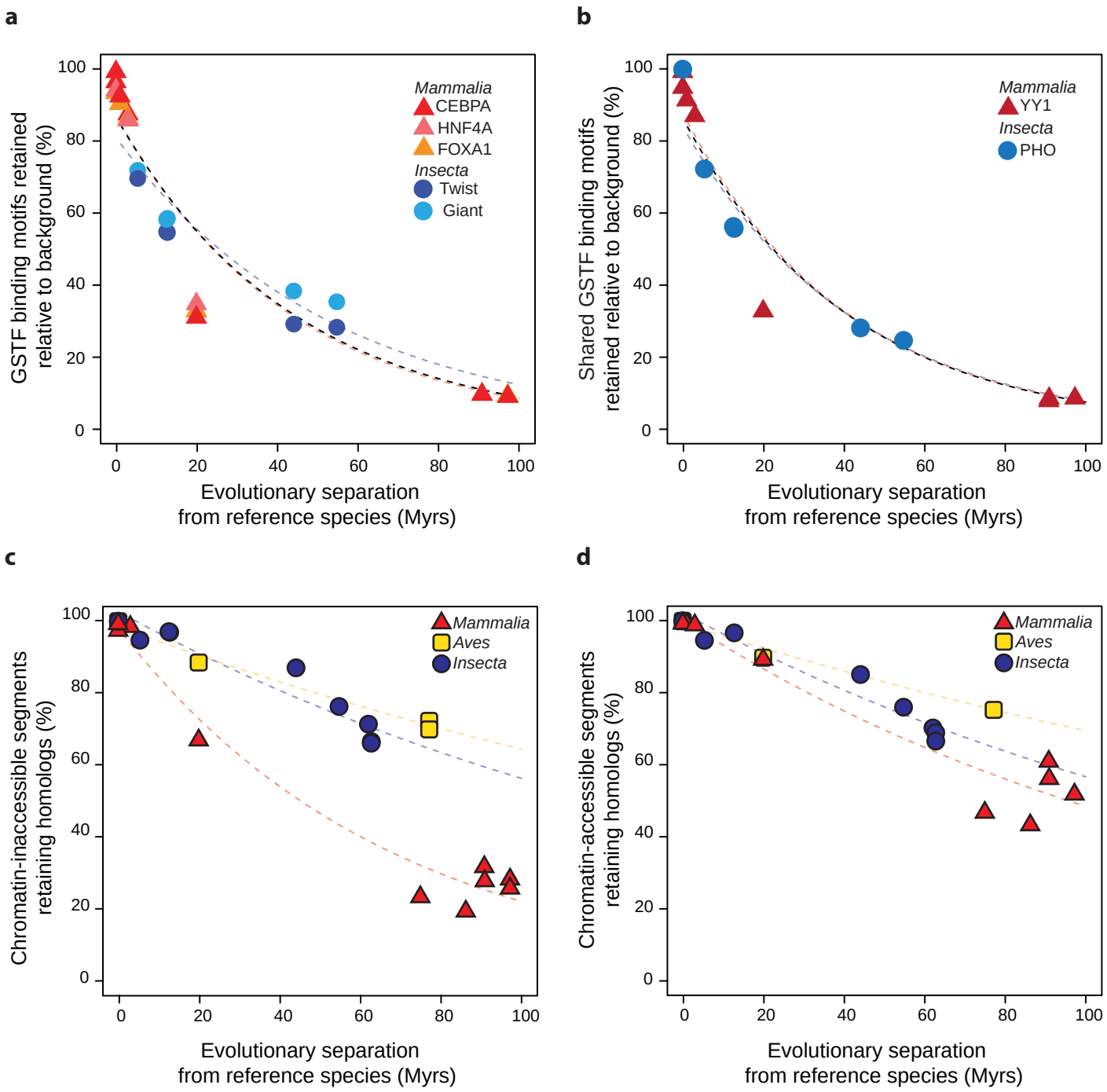

a
Occupancy profiles (continuous)
Occupied segments (discrete)
Query ChIP-seq reads
Reference ChIP-seq reads
Choices in profile-to-segment discretization
Peak caller
SPP
MACS2
Quality filter
Stringent
Lenient
Asymmetric
Choices in cross-species comparison
Sequence similarity filter
Stringent
Lenient
None
Reference subset
All
Pairwise-syntenic
Super-syntenic
Overlap requirement
Stringent
Lenient
b
MASAL
Genomic segments retaining occupancy (%)
Mammalia
CEBPA
Insecta
Twist
Evolutionary separation from reference species (Myrs)
c
SLNPS
Genomic segments retaining occupancy (%)
Mammalia
CEBPA
Insecta
Twist
Evolutionary separation from reference species (Myrs)
d
GSTF with the faster rate of change across 108 combinations of analytical choices
Indistinguishable
Mammalia CEBPA
Insecta Twist
CEBPA
HNF4A
FOXA1
Twist
Giant
e
Correlation of occupancy profiles (PCC)
Mammalia
CEBPA
HNF4A
FOXA1
Insecta
Twist
Giant
Evolutionary separation from reference species (Myrs)
f
Correlation of occupancy profiles (PCC)
Mammalia
CTCF
Insecta
CTCF
Evolutionary separation from reference species (Myrs)

Figure 5


a
GSTF binding motifs retained relative to background (%)
Evolutionary separation from reference species (Myrs)
Mammalia
CEBPA
HNF4A
FOXA1
Insecta
Twist
Giant
b
Shared GSTF binding motifs retained relative to background (%)
Evolutionary separation from reference species (Myrs)
Mammalia
YY1
Insecta
PHO
c
Chromatin-inaccessible segments retaining homologs (%)
Evolutionary separation from reference species (Myrs)
Mammalia
Aves
Insecta
d
Chromatin-accessible segments retaining homologs (%)
Evolutionary separation from reference species (Myrs)
Mammalia
Aves
Insecta