Amalgamated cross-species transcriptomes indicate organ-specific preadaptation for functional shifts in gene expression

Kenji Fukushima[1,*] and David D. Pollock[1,*]

[1]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045, USA.

* Correspondence to: kfuku52@gmail.com (K.F.) and david.pollock@ucdenver.edu (D.D.P.)

One Sentence Summary: Integration of 1,903 RNA-seq data reveals modular constraints on gene exchange among vertebrate organs.

Abstract:

The origins of complex multicellular physiology lie in the evolution of gene expression. Genes express differentially among organs as organisms evolve, but it is not well understood whether expression in certain organs increases the probability that they will be repurposed for expression in other organs. To examine this question, we amalgamated 1,903 RNA-seq datasets from 182 research projects, including 6 organs in 21 vertebrate species. We used various automated quality controls to eliminate project-specific biases, and defined expression regime evolution using tree-based Ornstein–Uhlenbeck models, allowing us to reconstruct evolutionary pathways of gene-family-wise expression on a genome-wide scale. Fluxes in organ-specific gene expression were non-random, suggesting that some ancestral expression patterns strongly preadapted genes for expression in certain organs, while others did not. For example, brain, ovary, and testis tend to change expression from one to another, while kidney and liver form a separate loop of gene expression evolution, illustrating a strong modularity of gene exchange among vertebrate organs. More subtle but significant differences depended on the type of gene duplication. Notably, RNA-based gene duplications tended to generate asymmetric exchange fluxes between organs. This detailed view of the evolutionary dynamics of gene expression supports a major role for preadaptive pathways.

Main Text:

Vertebrate organs serve as functional units of physiological activities, and their identities and functions are determined by the diverse expression patterns of thousands of genes. Because of this, the evolution of gene expression patterns plays a central role in organismal evolution. The degree of organ expression specificity affects how fast amino acids substitute (Zhang and Li, 2004), how rapidly they change expression levels (Liao and Zhang, 2006), and patterns of histone modifications (She et al., 2009). Lineage-defining organs such as the hominoid brain also depend on gene expression shifts (Chen et al., 2013; Kaessmann, 2010; Zhang and Long, 2014; Zhang et al., 2011). However, evolutionary dynamics and transitions among expression patterns remain poorly understood.

The evolution of new organismal traits necessarily depends on the pre-existing state of an organism. The idea that certain pre-existing adapted states are more conducive to evolution of specific new traits than other pre-existing states is known as preadaptation, also referred to as exaptation (Gould and Vrba, 1982). Evidence for preadaptation has been found in phenotypic traits, as it was traditionally described (Budd, 2007), and recently in molecular traits such as protein sequences during *de novo* gene birth (Wilson et al., 2017) or during functional innovations (Starr et al., 2017). However, as opposed to protein sequence evolution, which generally involves highly epistatic interactions and context-dependent changes (Podgornaia and Laub, 2015; Starr et al., 2017), the modular nature of expression regulation (de-Leon and Davidson, 2007) makes it unclear how much preexisting expression patterns influence evolutionary outcomes. Here, we ask whether preadaptation plays an important role in the dynamics of gene expression evolution using large amounts of transcriptome data from many studies.

To allow evolutionary gene expression analysis for a broad set of genes, we developed an extended phylogenetic approach tailored to gene family trees with duplications and losses, and applied it to 21 tetrapod genomes (Fig. S1). Only 1,377 genes per species (28,917 in total) in this dataset were ‘conserved orthologs’, whereas there were 15,282 gene trees in total, comprised of 15,787 genes per species on average (331,530 genes in total), including 21,242 human genes. This approach avoids problems with pairwise analyses (Assis and Bachtrog, 2015; Chen and Zhang, 2012; Guschanski et al., 2017; Kryuchkova-Mostacci et al., 2016; Lan and Pritchard, 2016; Rogozin et al.,

65 2014; Warnefors and Kaessmann, 2013) that ignore phylogenetic tree structure, and
66 allows inference of expression at ancestral nodes in the tree. Because we use gene
67 phylogenies rather than a single species phylogeny (Brawand et al., 2011), we can
68 analyze gene families with lineage-specific gene duplications and losses, and we did not
69 need to bias our study towards conserved genes and families with slow gene turnovers.
70 Such biases would tend to limit the number of organ-specific genes, which are enriched
71 in lineage-specific and young duplications (Guschanski et al., 2017).

72 To attain a good evolutionary resolution, we amalgamated 1,903 RNA-seq
73 experiments from 182 research projects and generated a dataset covering 6 organs from
74 21 vertebrate species (Tables S1–S2; Fig. S1). In comparison, other recent comparative
75 transcriptomic analyses of vertebrates (Barbosa-Morais et al., 2012; Breschi et al., 2016;
76 Carelli et al., 2016; Chen et al., 2017; Cortez et al., 2014; Guschanski et al., 2017;
77 Julien et al., 2012; Necsulea et al., 2014; Warnefors and Kaessmann, 2013) often used
78 the same dataset containing 131 RNA-seq experiments from 6 organs and 10 species
79 (Brawand et al., 2011), with some additional data in different studies. To allow rapid
80 integrated analysis of datasets, we employed automated multi-aspect quality controls,
81 including metadata curation (Table S3), sequence read filtering (Fig. S2), and
82 anomalous data removal by checking correlations between and within data categories
83 (Fig. 1A–C; Fig. S3). Finally, we applied surrogate variable analysis (SVA) (Leek and
84 Storey, 2007) to detect and correct hidden biases likely originating from heterogeneous
85 sampling conditions and sequencing procedures among experiments (Fig. 1D; Fig. S4;
86 for individual analysis of all 21 species, see Figs. S5–S7 and Supplementary Data). This
87 correction greatly improved the correlation of expression levels within organs from the
88 same species, even when data was derived from different research projects, while
89 comparisons among different organs yielded consistently low correlations (Fig. 1B).
90 Furthermore, a dimensionality reduction analysis produced excellent organ-wise
91 segregation in the multispecies comparison (Fig. 1E; Fig. S8). The curated expression
92 data thus appear sufficient for use in cross-species expression pattern analyses.

93 We next modeled expression evolution along 15,282 maximum-likelihood gene
94 family phylogenies (Table S4). We used Ornstein–Uhlenbeck models to allow for
95 possible adaptive shifts in optimal expression levels (Bastide et al., 2018; Hansen, 1997;
96 Khabbazian et al., 2016). Brownian motion models embody purely neutral expression
97 evolution (Bedford and Hartl, 2009), whereas Ornstein–Uhlenbeck models are designed

98 to detect purifying selection and adaptive evolution along with neutral fluctuation
99 (Brawand et al., 2011; Chen et al., 2017; Rohlfs et al., 2014). This analysis identifies
100 points where statistically significant expression 'regime shifts' (Hansen, 1997) occur
101 along each gene tree (Fig. 2A). To evaluate the effect of branching events preceding
102 expression shifts, we categorized gene tree nodes by whether they were associated with
103 DNA-based or RNA-based duplication, or not ('speciation') (Fig. 2B). Because
104 regulatory environments and expression patterns are more preserved among orthologous
105 genes in comparison with paralogous genes (Altenhoff et al., 2012; Castillo-Davis et al.,
106 2004; Chen and Zhang, 2012; Rogozin et al., 2014), we consider changes following
107 speciation nodes to be the 'baseline' mode of expression evolution for comparison with
108 duplication events.

109 An orthogroup of phosphoglycerol kinases (PGKs), for example, contains all three
110 categories of branching events followed by expression shifts (Fig. 2C). This protein
111 catalyzes the first ATP-generating step in the glycolytic pathway and is required for
112 most cell types including sperms (Danshina et al., 2010; Liu et al., 2015). PGK1, the
113 original copy on the X chromosome, is known to have duplicated independently in
114 eutherians and marsupials to produce the autosomal retrocopy PGK2 that compensates
115 the protein activity during X-inactivation (Boer et al., 1987; McCarrey and Thomas,
116 1987; Potrzebowski et al., 2008). Our automated analysis correctly recovered the
117 independent retrotranspositions and subsequent gains of testis-specific expression in
118 eutherian and marsupial lineages. This example illustrates that our genome-wide
119 analysis can readily recover expression evolutionary trajectories that are compatible
120 with focused gene family analyses (see Supplementary Data for individual gene trees).

121 To summarize gene expression evolution, we mapped shift event totals onto the
122 species tree (Fig. 3A–B). During vertebrate diversification, expression regimes were
123 usually conserved within gene groups formed by speciation (26,158 shifts/549,012
124 branches, 4.8% per branch, at a rate of 0.52 shifts/kiloyear). In contrast, expression
125 shifts occurred at three times this rate after DNA-based duplication (7,179 shifts/72,162
126 branches, 9.9% per branch, at a rate of 1.49 shifts/kiloyear). RNA-based duplications
127 occurred four times faster than DNA duplications (1,176 shifts/3,345 branches 35.2%
128 per branch, at a rate of 6.04 shifts/kiloyear). This may be explained by the idea that
129 functional divergence alters expression patterns and enhances long-term retention of
130 duplicated copies (Assis and Bachtrog, 2015; Huerta-Cepas et al., 2011). Genes that

131 shift genome location tend to have higher expression shift frequencies than those that do
132 not (Fig. S9), in line with previous observations from the human genome (Lan and
133 Pritchard, 2016). Expression levels varied most in ovary and testes, which had
134 significantly higher average stationary variances than the other four organs (Fig. 3C).
135 This is consistent with the observation of strong sex-related divergence of gene
136 expression in testes (Brawand et al., 2011), but shows strong sex-related divergence in
137 the reproductive organs of both sexes.

138 Expression shifts may coincide with changes in protein sequence evolution, and an
139 analysis of nonsynonymous/synonymous substitution rate ratios revealed a significant
140 association in all three branching categories, with a larger effect size in duplication than
141 in speciation (Fig. S10). An expression shift is associated with a 24% increase in
142 median nonsynonymous/synonymous substitution rate ratio in DNA-based duplications
143 (0.218 to 0.270) and 764% increase in RNA-based duplications (0.0456 to 0.394), in
144 contrast to only 4% increase in speciation (0.0941 to 0.0977). These differences are
145 much smaller than the differences between duplication and speciation, and small
146 compared to the variance among orthogroups (4.38), suggesting that the effect is not
147 large enough to be reliably detected in individual gene family analyses.

148 To further characterize how expression properties change from ancestral to derived
149 regimes, we examined changes in expression specificity, measured by τ (Yanai et al.,
150 2005). Genes with highly specific expression have τ values close to one, whereas
151 uniformly expressed genes yield τ values close to zero. Compared with speciation,
152 expression patterns following RNA-based duplications are more specific, and
153 DNA-based duplications were in between (Fig. 3D). We next examined the maximum
154 expression level among the six organs, because maximum rather than average values are
155 expected to reflect the 'on state' transcriptional activity. Duplication-associated shifts
156 tended to have bigger changes in maximum expression level, with a pronounced
157 enrichment of down-regulation (Fig. 3E). Down-regulation related to duplication events
158 may be best explained by either recent nonfunctionalization (Balakirev and Ayala,
159 2003; Mighell et al., 2000) or specialized expression in unanalyzed organs. Finally, we
160 compared expression complementarity (Huerta-Cepas et al., 2011) to evaluate the
161 differentiation in expression patterns between sister lineages. Duplication events gave
162 rise to greater complementarity in expression more often than speciation events (Fig.
163 3F), further supporting the role of gene duplication in functional specialization. Changes

164 in expression properties were more drastic in RNA-based duplications, followed by
165 DNA-based duplications and speciation events, probably reflecting the different modes
166 of inheritance of regulatory environments. RNA-based duplication accompanies the
167 elimination of regulatory elements outside transcribed regions, whereas DNA-based
168 duplication can retain certain regulatory regions depending on the range of duplicated
169 genomic segments (Babushok et al., 2007).

170 To reveal the evolutionary dynamics of gene expression patterns, we quantified the
171 gain and loss of genes characterized by the organ in which they are most highly
172 expressed (primary-expressed). Across vertebrates, switching from one
173 primary-expressed organ to another was detected in 14,403, 4,364, and 783 regime
174 shifts after speciation, DNA-based duplication, and RNA-based duplication,
175 respectively (Fig. 4A; Fig. S11). Of note, the gain/loss ratios are heterogeneous among
176 organs, suggesting that vertebrate organs serve as both sources and sinks in expression
177 evolution, but that their relative contributions are organ-specific.

178 The preadaptation hypothesis predicts that expression shifts depend on ancestral
179 expression conditions prior to the shifts. To test this, we examined whether shifts in
180 primary-expressed organs are significantly different from random expectation
181 (controlled for the total number of shifts from and to each organ). There are clear
182 patterns of evolutionary transitions, many of which are largely shared between
183 speciation and duplication events (Fig. 4B). For example, brain, ovary, and testis
184 showed strong connections, indicating a solid exchange module. Kidney and liver also
185 donate genes to one another, forming a separate module from brain–ovary–testis.
186 Although flows are similar among the three branching categories, there were
187 significantly different trends, one of the strongest being more asymmetric interactions
188 between heart and kidney in duplication events compared with speciation events (χ^2 =
189 22.938, P = 1.05×10^{-5}; Fig. 4A; Table S5). In general, speciation-associated and
190 DNA-duplication-associated shifts were moderately symmetric in the flow between
191 pairs of organs, while RNA-duplication-associated shifts were more asymmetric (Fig.
192 S12). The lower symmetry in RNA duplications contributed to flux in expression
193 patterns over evolutionary timescales.

194 To eliminate the effect of subtle expression changes, we also analyzed shifts with
195 drastic expression changes by introducing a cutoff of expression specificity (τ>0.67) to
196 define organ-specific genes; with this cutoff, the second-highest expression level is no

197 greater than two-thirds of the highest expression in both ancestral and descendent states.
198 Although some previously significant trends were not recovered due to small sample
199 sizes, especially in RNA-based duplication, the result was largely consistent with the
200 broader analysis (Fig. S13; Table S5), demonstrating the robustness of our conclusion.

201 Following RNA-based duplication, genes often become expressed in testis (Fig.
202 S11). This result is somewhat concordant with the "out of the testis" hypothesis, which
203 explains the accelerated gains of testis expression based on the permissive chromatin
204 state, abundant transcriptional machinery, relatively simple promoters required for the
205 expression in spermatogenic cells and following gains of new expression patterns
206 (Kaessmann, 2010; Kleene, 2005).

207 The above results suggest that the landscape of expression evolution is strongly
208 shaped by mechanisms of gene birth. Patterns of organ switching depend on ancestral
209 states, suggesting prevalent preadaptation in the evolution of vertebrate gene expression.
210 Future study should address the mechanisms responsible for patterns of preadaptation.
211 We hypothesize that one possibility is a 'regulatory' mechanism, whereby frequent
212 gene-exchanging organs use similar sets of regulatory elements; altered expression
213 between such organs could occur with relatively few mutations in regulatory sequences.
214 Another hypothetical 'coding' mechanism is that frequently interacting organs may
215 have similar environmental requirements for expressed proteins, such that few amino
216 acid substitutions tend to be required to adjust biochemical properties. Protein
217 reusability may be determined by cellular environments such as pH and temperature or
218 by functional categories of proteins. The regulatory and coding hypotheses are not
219 mutually exclusive, and both factors may contribute to generate preadaptive patterns of
220 gene expression.

221 In this study, we established a method to standardize RNA-seq data from disparate
222 research projects and developed a pathway for data amalgamation. Using gene family
223 phylogenies, our analysis revealed non-randomness and directionality of expression
224 evolution. This suggests prevalent preadaptation in gene expression, that adaptation to
225 expression in certain organs is more conducive to future expression in other organs, and
226 shows how gene duplication helped to reshape the dynamics of expression evolution
227 that contributed to the vertebrate diversification.

228 **Materials and Methods:**

229 **Species selection.** A total of 105 species included in the Ensembl release 91 (Yates
230 et al., 2016) were searched for data availability in the NCBI SRA database (Leinonen et
231 al., 2011) (final search on May 1, 2018) and 22 species were found to have RNA-seq
232 data for 6 organs: brain, heart, kidney, liver, ovary, and testis. *Lepisosteus oculatus* was
233 excluded due to an insufficient quality of available expression data and therefore
234 remaining 21 species were selected for further analysis.

235 **Species tree.** The dated species tree for the 21 species was retrieved from
236 TimeTree (Hedges et al., 2006) (downloaded on March 15, 2018; Supplementary Data).
237 Some species were unavailable in the database and therefore they were temporarily
238 replaced by closely related species to obtain the dated species tree.

239 **Transcriptome metadata curation.** We developed an automated python program
240 for SRA metadata curation (Table S3; Supplementary Data). RNA-seq data were
241 selected from the NCBI SRA database by keyword searches limited to the 21 species,
242 the 6 organs, and Illumina sequencing platforms. Orthographical variants of annotations
243 were standardized with keyword libraries created by manually checking the original
244 annotations. Prenatal or unhealthy samples and small-scale sequencing samples (<5
245 million reads) were excluded. Data for non-messenger RNA sequencings were also
246 removed. In treatment and control RNA-seq pairs, only control experiments were
247 included.

248 **Transcriptome quantification.** Fastq files were extracted from downloaded SRA
249 files using parallel-fastq-dump 0.6.2 (https://github.com/rvalieris/parallel-fastq-dump)
250 with the minimum read length of 25 bp and the quality filter (-E option) (Leinonen et al.,
251 2011). The fastq sequences were then subjected to a quality filtering by fastp 0.12.3
252 (Chen et al., 2018). The filtered reads were mapped to genomic features annotated as
253 non-messenger RNAs in the Ensembl GTF files using bowtie2 2.3.4 (Langmead and
254 Salzberg, 2012) and resultant unmapped reads were used for expression level
255 quantification using kallisto 0.43.1 with the sequence-based bias correction (Bray et al.,
256 2016). Samples were removed if 20% or smaller percentages of reads were mappable
257 (Fig. S2). Log_2-transformed TPM (transcripts per million) values were subjected to
258 downstream analyses.

259 **Iterative transcriptome curation.** Anomalous RNA-seq samples were iteratively
260 removed by employing an expression bias correction and a correlation analysis. In each

261 iteration, unwanted biases in expression level were removed by surrogate variable
262 analysis (SVA) (Leek and Storey, 2007). SVA analysis was applied at the beginning of
263 each iteration so that it is not influenced by anomalous samples removed in previous
264 iterations. Subsequently, Pearson's correlation coefficients were calculated for every
265 RNA-seq data against mean expression level in each organ generated by averaging all
266 other data excluding those from the same BioProject (Fig. S3). We assume that the
267 sample's correlation coefficient against the same organ is higher than any of the values
268 against the other organs, and we removed samples that violated this assumption. These
269 steps were repeated until no violations were left and SVA-corrected expression levels
270 were finally reported. The curation steps were skipped if multiple samples were
271 unavailable in the species and hence SVA analysis was inapplicable. The final dataset
272 was comprised of 1,903 RNA-seq experiments from 182 BioProjects that cover 6
273 organs from 21 vertebrate species (Tables S1–S2).

274 **Orthogroup classification.** Orthogroups, which contain all genes descended from
275 one gene in the common ancestor, were inferred from coding sequences of the 21
276 species using OrthoFinder 2.1.2 (Emms and Kelly, 2015) guided by the species tree. In
277 total, 17,896 orthogroups were generated with the largest orthogroup comprised of
278 7,893 genes.

279 **Multiple sequence alignment.** Multiple fasta files containing coding sequences
280 were generated for each orthogroup. Stops and ambiguous codons were masked as gaps.
281 In-frame multiple codon sequence alignments were generated using MAFFT 7.394 with
282 the 'auto' option (Katoh and Standley, 2013) and tranalign in EMBOSS 6.5.7.0 (Rice et
283 al., 2000). Anomalous genes were excluded by MaxAlign (Gouveia-Oliveira et al.,
284 2007) which decreased the maximum orthogroup size from 7,893 to 6,553 genes.
285 Spurious codons were removed in-frame using pgtrimal in Phylogears2-2.0.2016.09.06
286 (https://www.fifthdimension.jp/products/phylogears/) with the 'gappyout' option
287 (Capella-Gutierrez et al., 2009). After complete gap deletion, 2,614 orthogroups showed
288 no parsimony-informative sites monitored by AMAS (Borowiec, 2016). Because
289 phylogenetic relationships cannot be inferred from such orthogroups, we discarded them
290 and retained 15,282 orthogroups for further analysis (Table S4).

291 **Gene tree reconstruction.** Maximum-likelihood trees were reconstructed using
292 IQ-TREE 1.6.5 (Nguyen et al., 2015) with the best-fit nucleotide substitution models
293 selected by ModelFinder with the Bayesian Information Criterion (Kalyaanamoorthy et

294 al., 2017). Larger orthogroups and longer genes tended to fit more complex substitution
295 matrices and larger numbers of categories of rate heterogeneity (Fig. S14A–B; Table
296 S4). Ultrafast bootstrapping with 1,000 replicates was performed to evaluate the
297 credibility of tree topology (Minh et al., 2013) with a further optimization of each
298 bootstrapping tree (-bnni option) (Hoang et al., 2018).

299 **Reconciliation-assisted gene tree rooting.** Candidate rooting positions were
300 inferred with different methods. Using the dated species tree, all rooting branches with
301 the minimum duplication-loss score were identified using the rooting mode of
302 NOTUNG 2.9 with the default parameters (duplication score = 1.5, loss score = 1.0)
303 (Chen et al., 2000). The midpoint of the longest path (Farris, 1972) and the position
304 with the minimal ancestor deviations (MAD) (Tria et al., 2017) were also considered as
305 candidates. The final rooting position was reported based on overlaps among those
306 rooting positions (Figs. S14C–S15; Table S4).

307 **Reconciliation-assisted divergence time estimation.** To prepare dated gene trees,
308 we first matched species tree nodes with corresponding gene tree nodes using the
309 reconciliation mode of NOTUNG 2.9 (Chen et al., 2000) and created time constraints of
310 speciation nodes (Fig. S15). Duplication nodes were constrained with the upper and
311 lower age limits derived from corresponding speciation nodes. If the root node is a
312 duplication node and is not covered by the range of the species tree, the upper age limit
313 was set to 1,105 million years ago, which corresponds to the split of animals and fungi
314 (Hedges et al., 2006). Divergence time was then estimated by a penalized likelihood
315 method (Sanderson, 2002) implemented in an R package 'ape' ('chronos' function)
316 (Popescu et al., 2012) with time constraints on speciation, duplication, and root nodes.
317 On 4,472 out of 15,282 gene trees, reasonable initial parameters were not found after
318 1,000 trials with different initial values and thus the above constraints were partly
319 relaxed (Fig. S14D; Table S4).

320 **Modeling and shift detection of expression evolution.** Using the dated gene trees
321 and SVA-corrected expression matrices, expression shifts were detected as shifts of
322 optimal trait values in Ornstein–Uhlenbeck (OU) models determined by a Lasso-based
323 model selection with AICc in an R package '*l*1ou' (Khabbazian et al., 2016) or by an
324 expectation–maximization algorithm with a penalized likelihood in an R package
325 'PhylogeneticEM' (Bastide et al., 2018). To handle gene trees recalcitrant to this

326 analysis (especially those with a large number of genes), we skimmed gene trees by
327 collapsing clades with small changes in expression level (Fig. S16). Specifically, we
328 first calculated all-vs-all Pearson's correlation coefficients of gene expression level
329 among all genes that belong the clade. The clade was replaced with a single tip if the
330 minimum correlation coefficient was greater than 0.5. Phylogenetic means of the clade
331 were calculated by assuming Brownian motion and were used as expression level at the
332 new tip. Although the use of PhylogeneticEM and the tree skimming reduced the
333 number of detected regime shifts, the positions of detected shifts were largely
334 overlapped with those inferred with *l*1ou without tree skimming (Fig. S17). In addition,
335 this strategy allowed us to analyze the largest gene tree with 6,553 genes. Therefore, we
336 used PhylogeneticEM and/or the tree skimming if the gene tree cannot be analyzed by
337 *l*1ou with non-skimmed topology (Table S4; 8/15,282 trees). The upper limit of regime
338 shifts was set as $\max\left[\min(N/2, 100), \sqrt[2]{N}\right]$. In most gene trees (15,278/15,282), the
339 number of detected regime shifts was smaller than the upper limit (Fig. S18).

340 **Analysis of expression patterns.** We characterized expression patterns of extant
341 and ancestral genes by calculating different metrics from fitted values (μ) in the OU
342 models. Organ specificity was measured by τ (Yanai et al., 2005), which outperformed
343 other methods in a benchmark for tissue specificity (Kryuchkova-Mostacci and
344 Robinson-Rechavi, 2016). Expression complementarity between sister lineages was
345 measured by TEC (Huerta-Cepas et al., 2011). Because μ was estimated from
346 log-transformed expression levels, these expression metrics were calculated with
347 unlog-transformed μ values. Primary-expressed organs of gene expression patterns were
348 defined as the organ in which the highest expression levels were observed among the
349 six organs we analyzed.

350 **Estimation of protein evolution rate.** Parameters for codon substitution matrix,
351 shape parameter of discrete gamma distribution for rate heterogeneity (α), equilibrium
352 transition/transversion rate ratio (κ), equilibrium nonsynonymous/synonymous
353 substitution rate ratio (ω) were estimated using IQ-TREE 1.6.5 (Nguyen et al., 2015) by
354 fitting GY+F3X4+G4 models to each gene tree. Equilibrium base composition (θ) was
355 estimated from empirical codon state frequencies which are calculated from the
356 alignment by counting. To obtain θ at the root node, we calculated θ at subroot nodes by

357 taking advantage of IQ-TREE's empirical Bayesian method for ancestral sequence
358 reconstruction. Considering subroot branch length, a weighted average of the subroot θ
359 values were calculated as the θ value at the root node. Using all those parameters,
360 branch-wise nonsynonymous/synonymous substitution rate ratios were estimated by
361 stochastic substitution mapping (*mapdNdS*) (Guéguen and Duret, 2018) using bio++
362 library (Guéguen et al., 2013). To examine the robustness of the *mapdNdS*-based ω
363 estimation, we compared the results with those obtained by maximum-likelihood ω
364 estimation by fitting MG94W9 models in HyPhy 2.3.11 (Pond et al., 2005). The two
365 methods yielded a high correlation of ω values (Spearman's $\rho = 0.71$, $N = 622{,}906$
366 branches) and consistent results on the effect of branching events and expression shifts
367 (Fig. S10), suggesting methodological robustness. We reported *mapdNdS*-based results
368 in the main text.

369 **Analysis of gene structure and location.** The number of introns and chromosomal
370 location were obtained for each gene from the Ensembl gene models (GFF3 files). The
371 intron numbers were subsequently converted to binary values that represent intronless
372 and intron-containing states. Chromosomal locations were categorized into autosome, X
373 chromosome, and Y chromosome. Genes from non-therian species were treated as
374 missing data because mechanisms of their sex determination are not homologous to the
375 mammalian XY system (Veyrunes et al., 2008). Genes from *Chinchilla lanigera* were
376 also treated as missing data because their sequenced genomes are not anchored to
377 chromosomes in the Ensembl release 91. The posterior probabilities of ancestral
378 character states were inferred by the stochastic character mapping of discrete traits
379 (Bollback, 2006) implemented in an R package 'phytools' (Revell, 2012). Because
380 functional retrotranspositions (Jun et al., 2009; Marques et al., 2005), intron gains (Roy
381 and Gilbert, 2005), and inter-chromosomal duplications (Pace et al., 2009) are rare
382 events relative to the timescale of the vertebrate evolution on the per-gene basis, we set
383 the transition rate parameters to a sufficiently small value (1×10^{-4} per gene per million
384 years) that reduces the model into the maximum parsimony paradigm.

385 **Analysis of branching events.** Speciation and duplication nodes were classified by
386 a species-overlap method (Huerta-Cepas et al., 2007) and were mapped to the species
387 tree on the basis of species coverages. Duplication confidence scores (Vilella et al.,
388 2009) were calculated for the duplication nodes and scores greater than 0 were
389 considered duplication events. A transition from intron-containing to intronless states

with a posterior probability greater than 0.5 was classified as a retrotransposition event. Although our classification cannot detect RNA-based duplication from originally intronless genes, we expect such situations would be rare because most vertebrate genes contain at least one intron (e.g., 20,160/21,242 human genes in our dataset). Inter-chromosomal translocation was detected by considering chromosomal locations with the highest posterior probability as the ancestral states. Because of the difficulty in determining rooting positions of deep phylogenies, gene tree nodes older than the root node of the species tree were removed from the analysis.

Gene tree annotations and visualization. Phylogenetic trees were visualized using a python package 'ETE 3' (Huerta-Cepas et al., 2016) and an R package 'ggtree' (Yu et al., 2017). A part of animal silhouettes in Fig. 3A were obtained from PhyloPic (http://phylopic.org).

Code availability. Scripts and parameter values used in this study are available as Supplementary Data (Dryad ID available on publication).

Acknowledgements:

We acknowledge the following sources for funding: MEXT/JSPS KAKENHI 18J00178 (K.F.) and NIH R01 GM083127 (D.D.P.). Computations were partially performed on the NIG supercomputer. The authors declare no competing financial interests.

Author Contributions:

K.F. and D.D.P. jointly designed the study and wrote the paper. K.F. designed and wrote all programs and performed data analysis.

414 **References:**

415 Altenhoff, A.M., Studer, R.A., Robinson-Rechavi, M., Dessimoz, C., and Couto, F.
416 (2012). Resolving the ortholog conjecture: orthologs tend to be weakly, but
417 significantly, more similar in function than paralogs. PLoS Comput. Biol. *8*,
418 e1002514.
419 Assis, R., and Bachtrog, D. (2015). Rapid divergence and diversification of mammalian
420 duplicate gene functions. BMC Evol. Biol. *15*, 138.
421 Babushok, D. V., Ostertag, E.M., and Kazazian, H.H. (2007). Current topics in genome
422 evolution: Molecular mechanisms of new gene formation. Cell. Mol. Life Sci. *64*,
423 542–554.
424 Balakirev, E.S., and Ayala, F.J. (2003). Pseudogenes: Are they "junk" or functional
425 DNA? Annu. Rev. Genet. *37*, 123–151.
426 Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J.,
427 Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., et al. (2012). The evolutionary
428 landscape of alternative splicing in vertebrate species. Science *338*, 1587–1593.
429 Bastide, P., Ané, C., Robin, S., Mariadassou, M., and Alfaro, M. (2018). Inference of
430 adaptive shifts for multivariate correlated traits. Syst. Biol. *67*, 662–680.
431 Bedford, T., and Hartl, D.L. (2009). Optimization of gene expression by natural
432 selection. Proc. Natl. Acad. Sci. U. S. A. *106*, 1133–1138.
433 Boer, P.H., Adra, C.N., Lau, Y.F., and McBurney, M.W. (1987). The testis-specific
434 phosphoglycerate kinase gene pgk-2 is a recruited retroposon. Mol. Cell. Biol. *7*,
435 3107–3112.
436 Bollback, J. (2006). SIMMAP: Stochastic character mapping of discrete traits on
437 phylogenies. BMC Bioinformatics *7*, 88.
438 Borowiec, M.L. (2016). AMAS: a fast tool for alignment manipulation and computing
439 of summary statistics. PeerJ *4*, e1660.
440 Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier,
441 M., Liechti, A., Aximu-Petri, A., Kircher, M., et al. (2011). The evolution of gene
442 expression levels in mammalian organs. Nature *478*, 343–348.
443 Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic
444 RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.
445 Breschi, A., Djebali, S., Gillis, J., Pervouchine, D.D., Dobin, A., Davis, C.A., Gingeras,
446 T.R., and Guigó, R. (2016). Gene-specific patterns of expression variation across

447 organs and species. Genome Biol. *17*, 151.
448 Brunner, E., and Munzel, U. (2000). The nonparametric Behrens-Fisher problem:
449 asymptotic theory and a small-sample approximation. Biometrical J. *42*, 17–25.
450 Budd, G.E. (2007). On the origin and evolution of major morphological characters. Biol.
451 Rev. *81*, 609–628.
452 Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009). trimAl: a tool for
453 automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics
454 *25*, 1972–1973.
455 Carelli, F.N., Hayakawa, T., Go, Y., Imai, H., Warnefors, M., and Kaessmann, H.
456 (2016). The life history of retrocopies illuminates the evolution of new mammalian
457 genes. Genome Res. *26*, 301–314.
458 Castillo-Davis, C.I., Hartl, D.L., and Achaz, G. (2004). cis-Regulatory and protein
459 evolution in orthologous and duplicate genes. Genome Res. *14*, 1530–1536.
460 Chen, X., and Zhang, J. (2012). The ortholog conjecture is untestable by the current
461 gene ontology but is supported by RNA sequencing data. PLoS Comput. Biol. *8*,
462 e1002784.
463 Chen, J., Swofford, R., Johnson, J., Cummings, B.B., Rogel, N., Lindblad-Toh, K.,
464 Haerty, W., Palma, F. di, and Regev, A. (2017). A quantitative model for
465 characterizing the evolutionary history of mammalian gene expression. BioRxiv
466 229096.
467 Chen, K., Durand, D., and Farach-Colton, M. (2000). NOTUNG: a program for dating
468 gene duplications and optimizing gene family trees. J. Comput. Biol. *7*, 429–447.
469 Chen, S., Krinsky, B.H., and Long, M. (2013). New genes as drivers of phenotypic
470 evolution. Nat. Rev. Genet. *14*, 645–660.
471 Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ
472 preprocessor. BioRxiv 274100.
473 Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P.D.,
474 Grützner, F., and Kaessmann, H. (2014). Origins and functional evolution of Y
475 chromosomes across mammals. Nature *508*, 488–493.
476 Danshina, P. V., Geyer, C.B., Dai, Q., Goulding, E.H., Willis, W.D., Kitto, G.B.,
477 McCarrey, J.R., Eddy, E.M., and O'Brien, D.A. (2010). Phosphoglycerate kinase 2
478 (PGK2) is essential for sperm function and male fertility in mice. Biol. Reprod. *82*,
479 136–145.

480 de-Leon, S.B.-T., and Davidson, E.H. (2007). Gene regulation: gene control network in
481 development. Annu. Rev. Biophys. Biomol. Struct. *36*, 191–212.
482 Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole
483 genome comparisons dramatically improves orthogroup inference accuracy.
484 Genome Biol. *16*, 157.
485 Farris, J.S. (1972). Estimating phylogenetic trees from distance matrices. Am. Nat. *106*,
486 645–668.
487 Gould, S.J., and Vrba, E.S. (1982). Exaptation—a missing term in the science of form.
488 Paleobiology *8*, 4–15.
489 Gouveia-Oliveira, R., Sackett, P.W., and Pedersen, A.G. (2007). MaxAlign:
490 maximizing usable data in an alignment. BMC Bioinformatics *8*, 312.
491 Guéguen, L., and Duret, L. (2018). Unbiased estimate of synonymous and
492 nonsynonymous substitution rates with nonstationary base composition. Mol. Biol.
493 Evol. *35*, 734–742.
494 Guéguen, L., Gaillard, S., Boussau, B., Gouy, M., Groussin, M., Rochette, N.C., Bigot,
495 T., Fournier, D., Pouyet, F., Cahais, V., et al. (2013). Bio++: efficient extensible
496 libraries and tools for computational molecular evolution. Mol. Biol. Evol. *30*,
497 1745–1750.
498 Guschanski, K., Warnefors, M., and Kaessmann, H. (2017). The evolution of duplicate
499 gene expression in mammalian organs. Genome Res. *27*, 1461–1474.
500 Hansen, T.F. (1997). Stabilizing selection and the comparative analysis of adaptation.
501 Evolution *51*, 1341–1351.
502 Hedges, S.B., Dudley, J., and Kumar, S. (2006). TimeTree: a public knowledge-base of
503 divergence times among organisms. Bioinformatics *22*, 2971–2972.
504 Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018).
505 UFBoot2: Improving the ultrafast bootstrap approximation. Mol. Biol. Evol. *35*,
506 518–522.
507 Huerta-Cepas, J., Dopazo, H., Dopazo, J., Gabaldón, T., McPherson, J., Marra, M.,
508 Hillier, L., Waterston, R., Chinwalla, A., Wallis, J., et al. (2007). The human
509 phylome. Genome Biol. *8*, 934–941.
510 Huerta-Cepas, J., Dopazo, J., Huynen, M.A., and Gabaldón, T. (2011). Evidence for
511 short-time divergence and long-time conservation of tissue-specific expression after
512 gene duplication. Brief. Bioinform. *12*, 442–448.

513 Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, analysis, and
514 visualization of phylogenomic data. Mol. Biol. Evol. *33*, 1635–1638.
515 Julien, P., Brawand, D., Soumillon, M., Necsulea, A., Liechti, A., Schütz, F., Daish, T.,
516 Grützner, F., and Kaessmann, H. (2012). Mechanisms and evolutionary patterns of
517 mammalian and avian dosage compensation. PLoS Biol. *10*, e1001328.
518 Jun, J., Ryvkin, P., Hemphill, E., Mandoiu, I., and Nelson, C. (2009). The birth of new
519 genes by RNA- and DNA-mediated duplication during mammalian evolution. J.
520 Comput. Biol. *16*, 1429–1444.
521 Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes.
522 Genome Res. *20*, 1313–1326.
523 Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S.
524 (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat.
525 Methods *14*, 587–589.
526 Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software
527 version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–
528 780.
529 Khabbazian, M., Kriebel, R., Rohe, K., and Ané, C. (2016). Fast and accurate detection
530 of evolutionary shifts in Ornstein-Uhlenbeck models. Methods Ecol. Evol. *7*, 811–
531 824.
532 Kleene, K.C. (2005). Sexual selection, genetic conflict, selfish genes, and the atypical
533 patterns of gene expression in spermatogenic cells. Dev. Biol. *277*, 16–26.
534 Kryuchkova-Mostacci, N., and Robinson-Rechavi, M. (2016). A benchmark of gene
535 expression tissue-specificity metrics. Brief. Bioinform. *44*, bbw008.
536 Kryuchkova-Mostacci, N., Robinson-Rechavi, M., Shmoish, M., Chalifa-Caspi, V.,
537 Shklar, M., and Ophir, R. (2016). Tissue-specificity of gene expression diverges
538 slowly between orthologs, and rapidly between paralogs. PLOS Comput. Biol. *12*,
539 e1005274.
540 Lan, X., and Pritchard, J.K. (2016). Coregulation of tandem duplicate genes slows
541 evolution of subfunctionalization in mammals. Science *352*, 1009–1013.
542 Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2.
543 Nat. Methods *9*, 357–359.
544 Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies
545 by surrogate variable analysis. PLoS Genet. *3*, e161.

546 Leinonen, R., Sugawara, H., and Shumway, M. (2011). The Sequence Read Archive.
547 Nucleic Acids Res. *39*, D19–D21.
548 Liao, B.-Y., and Zhang, J. (2006). Low rates of expression profile divergence in highly
549 expressed genes and tissue-specific genes during mammalian evolution. Mol. Biol.
550 Evol. *23*, 1119–1128.
551 Liu, X.-X., Zhang, H., Shen, X.-F., Liu, F.-J., Liu, J., and Wang, W.-J. (2015).
552 Characteristics of testis-specific phosphoglycerate kinase 2 and its association with
553 human sperm quality. Hum. Reprod. *126*, dev301.
554 Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A., and Kaessmann, H.
555 (2005). Emergence of young human genes after a burst of retroposition in primates.
556 PLoS Biol. *3*, e357.
557 McCarrey, J.R., and Thomas, K. (1987). Human testis-specific PGK gene lacks introns
558 and possesses characteristics of a processed gene. Nature *326*, 501–505.
559 Mighell, A.J., Smith, N.R., Robinson, P.A., and Markham, A.F. (2000). Vertebrate
560 pseudogenes. FEBS Lett. *468*, 109–114.
561 Minh, B.Q., Nguyen, M.A.T., and von Haeseler, A. (2013). Ultrafast approximation for
562 phylogenetic bootstrap. Mol. Biol. Evol. *30*, 1188–1195.
563 Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker,
564 J.C., Grützner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires
565 and expression patterns in tetrapods. Nature *505*, 635–640.
566 Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: A
567 fast and effective stochastic algorithm for estimating maximum-likelihood
568 phylogenies. Mol. Biol. Evol. *32*, 268–274.
569 Pace, J.K., Sen, S.K., Batzer, M.A., and Feschotte, C. (2009). Repair-mediated
570 duplication by capture of proximal chromosomal DNA has shaped vertebrate
571 genome evolution. PLoS Genet. *5*, e1000469.
572 Podgornaia, A.I., and Laub, M.T. (2015). Pervasive degeneracy and epistasis in a
573 protein-protein interface. Science *347*, 673–677.
574 Pond, S.L.K., Frost, S.D.W., and Muse, S. V. (2005). HyPhy: hypothesis testing using
575 phylogenies. Bioinformatics *21*, 676–679.
576 Popescu, A.-A., Huber, K.T., and Paradis, E. (2012). ape 3.0: New tools for
577 distance-based phylogenetics and evolutionary analysis in R. Bioinformatics *28*,
578 1536–1537.

579 Potrzebowski, L., Vinckenbosch, N., Marques, A.C., Chalmel, F., Jégou, B., and
580 Kaessmann, H. (2008). Chromosomal gene movements reflect the recent origin and
581 biology of therian sex chromosomes. PLoS Biol. *6*, e80.
582 Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and
583 other things). Methods Ecol. Evol. *3*, 217–223.
584 Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular
585 Biology Open Software Suite. Trends Genet. *16*, 276–277.
586 Rogozin, I.B., Managadze, D., Shabalina, S.A., and Koonin, E. V (2014). Gene family
587 level comparative analysis of gene expression in mammals validates the ortholog
588 conjecture. Genome Biol. Evol. *6*, 754–762.
589 Rohlfs, R. V., Harrigan, P., and Nielsen, R. (2014). Modeling gene expression evolution
590 with an extended Ornstein–Uhlenbeck process accounting for within-species
591 variation. Mol. Biol. Evol. *31*, 201–211.
592 Roy, S.W., and Gilbert, W. (2005). Rates of intron loss and gain: implications for early
593 eukaryotic evolution. Proc. Natl. Acad. Sci. U. S. A. *102*, 5773–5778.
594 Sanderson, M.J. (2002). Estimating absolute rates of molecular evolution and
595 divergence times: a penalized likelihood approach. Mol. Biol. Evol. *19*, 101–109.
596 She, X., Rohl, C.A., Castle, J.C., Kulkarni, A. V, Johnson, J.M., and Chen, R. (2009).
597 Definition, conservation and epigenetics of housekeeping and tissue-enriched genes.
598 BMC Genomics *10*, 269.
599 Soubrier, J., Steel, M., Lee, M.S.Y., Der Sarkissian, C., Guindon, S., Ho, S.Y.W., and
600 Cooper, A. (2012). The influence of rate heterogeneity among sites on the time
601 dependence of molecular rates. Mol. Biol. Evol. *29*, 3345–3358.
602 Starr, T.N., Picton, L.K., and Thornton, J.W. (2017). Alternative evolutionary histories
603 in the sequence space of an ancient protein. Nature *549*, 409–413.
604 Tria, F.D.K., Landan, G., and Dagan, T. (2017). Phylogenetic rooting using minimal
605 ancestor deviation. Nat. Ecol. Evol. *1*, 0193.
606 Veyrunes, F., Waters, P.D., Miethke, P., Rens, W., McMillan, D., Alsop, A.E., Grutzner,
607 F., Deakin, J.E., Whittington, C.M., Schatzkamer, K., et al. (2008). Bird-like sex
608 chromosomes of platypus imply recent origin of mammal sex chromosomes.
609 Genome Res. *18*, 965–973.
610 Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009).
611 EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in

612 vertebrates. Genome Res. *19*, 327–335.
613 Warnefors, M., and Kaessmann, H. (2013). Evolution of the correlation between
614 expression divergence and protein divergence in mammals. Genome Biol. Evol. *5*,
615 1324–1335.
616 Wilson, B.A., Foy, S.G., Neme, R., and Masel, J. (2017). Young genes are highly
617 disordered as predicted by the preadaptation hypothesis of de novo gene birth. Nat.
618 Ecol. Evol. *1*, 0146.
619 Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R.,
620 Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., et al. (2005). Genome-wide
621 midrange transcription profiles reveal expression level relationships in human tissue
622 specification. Bioinformatics *21*, 650–659.
623 Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences
624 with variable rates over sites: Approximate methods. J. Mol. Evol. *39*, 306–314.
625 Yang, Z. (1995). A space-time process model for the evolution of DNA sequences.
626 Genetics *139*, 993–1005.
627 Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D.,
628 Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2016). Ensembl 2016.
629 Nucleic Acids Res. *44*, D710–D716.
630 Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.-Y. (2017). ggtree : an R package
631 for visualization and annotation of phylogenetic trees with their covariates and other
632 associated data. Methods Ecol. Evol. *8*, 28–36.
633 Zhang, L., and Li, W.-H. (2004). Mammalian housekeeping genes evolve more slowly
634 than tissue-specific genes. Mol. Biol. Evol. *21*, 236–239.
635 Zhang, Y.E., and Long, M. (2014). New genes contribute to genetic and phenotypic
636 novelties in human evolution. Curr. Opin. Genet. Dev. *29*, 90–96.
637 Zhang, Y.E., Landback, P., Vibranovski, M.D., Long, M., and Stark, A. (2011).
638 Accelerated recruitment of new brain development genes into the human genome.
639 PLoS Biol. *9*, e1001179.
640

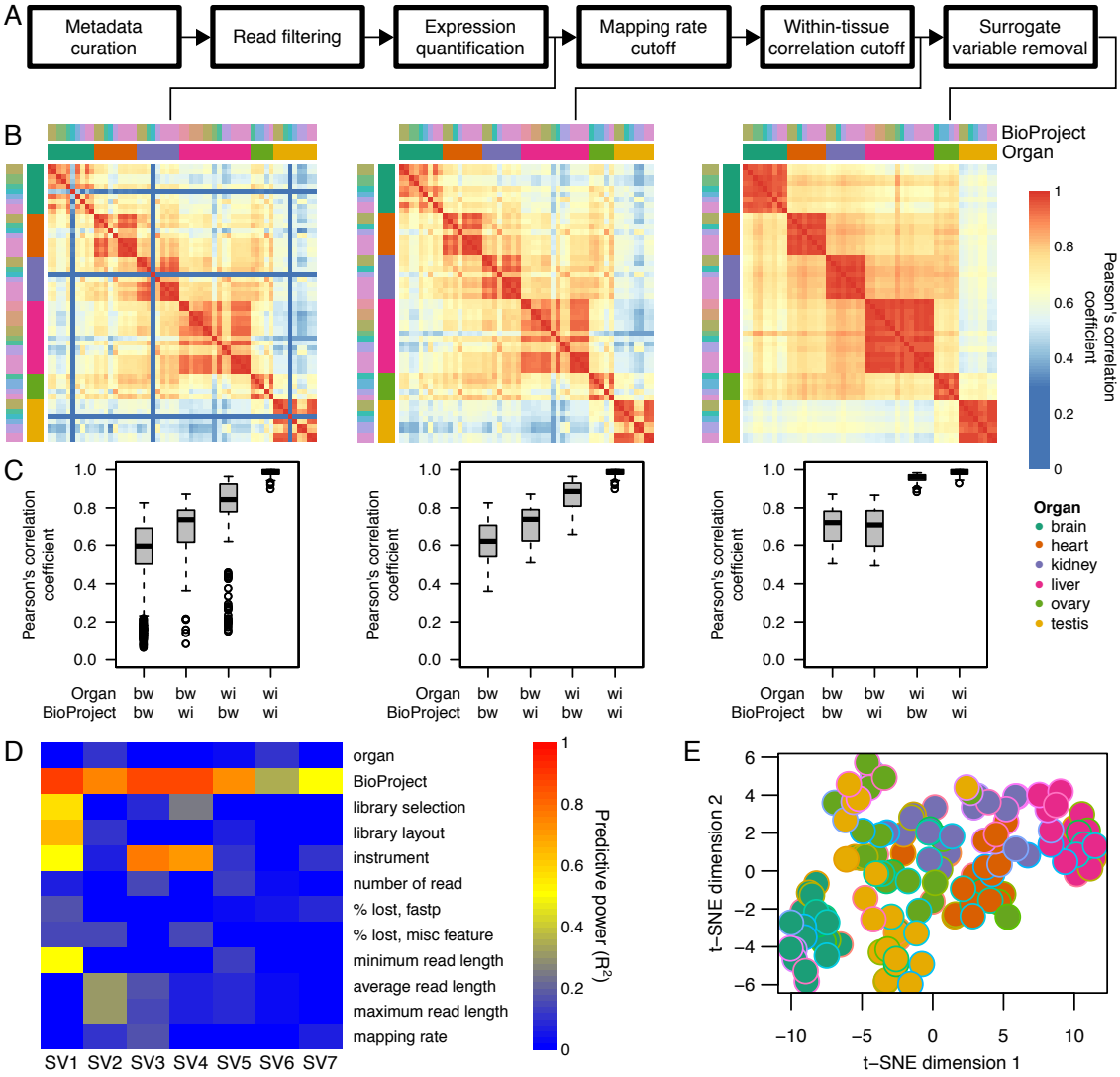


Figure 1. Transcriptome corrections to amalgamate heterogeneous RNA-seq experiments. (**A**) A simplified flow chart of the transcriptome curation. The full chart is available in Fig. S1. (**B–D**) Transcriptome curation within species. Data from *Monodelphis domestica* are shown as an example. The heatmaps show Pearson's correlation coefficients among RNA-seq samples (**B**). Pearson's correlation coefficients were calculated using expression levels of all genes. Note that project-specific correlations visible in the uncorrected data are absent in the corrected data. The box plots show distinct distributions of Pearson's correlation coefficients depending on whether a pair of samples are the same organ or whether they are from the same research project (**C**, bw, between categories; wi, within category). The correlation coefficients are largely improved in within-organ comparisons when surrogate variables are removed, while within-project biases are attenuated. In this species, 7 surrogate

654 variables were detected against 52 RNA-seq data from 8 projects (**D**). Analysis of those
655 variables by linear regression highlights the BioProject feature as the strongest source of
656 removed biases. For full description of predictors, see Fig. S7. (**E**) Nonlinear
657 dimensionality reduction using expression levels of 1,377 conserved orthologs from 21
658 species. Each point corresponds to the averaged transcriptome of one organ in one
659 species. The six organs are colored in the same way as in **B**. Species are identified by
660 border colors (see Fig. S8).

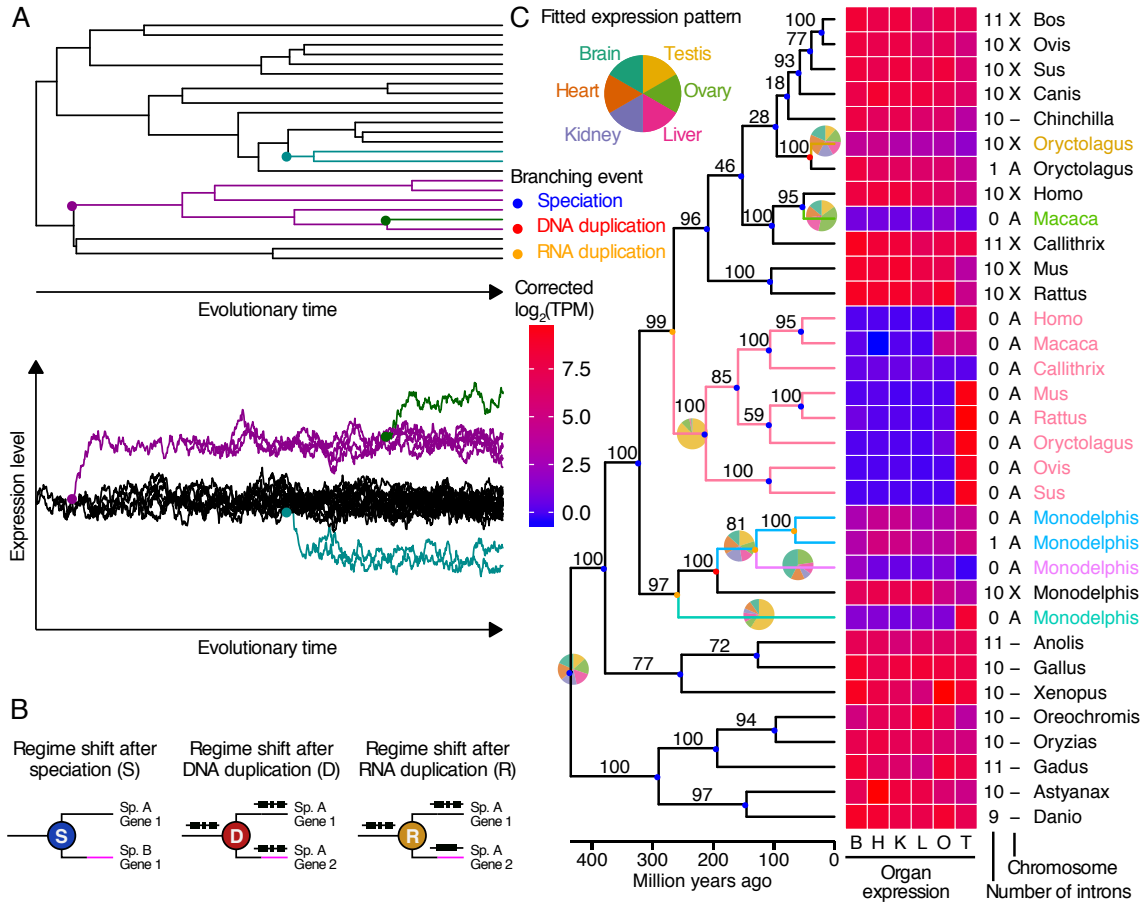


Figure 2. Expression evolution in a complex history of gene family evolution. (**A**) A phylogenetic simulation of univariate Ornstein–Uhlenbeck process. Regime shifts (indicated by colors) appear as a substantial change in optimal trait values. (**B**) Branching events on gene trees were categorized into speciation, DNA duplication and RNA duplication. (**C**) The gene tree of phosphoglycerol kinases (orthogroup ID: OG0002332) is shown as an example. This orthogroup shows ortholog-specific expression patterns as well as regime shifts after speciation or lineage-specific gene duplication. Branch colors indicate expression regimes. Node colors mean diversification events: blue, speciation; red, DNA duplication; orange, RNA duplication. Bootstrap supports are given above branches. The heatmap shows expression levels in, from the left, brain, heart, kidney, liver, ovary, and testis. To the right, the number of introns and located chromosomes (A, autosome; X, X chromosome; Y, Y chromosome) are indicated.

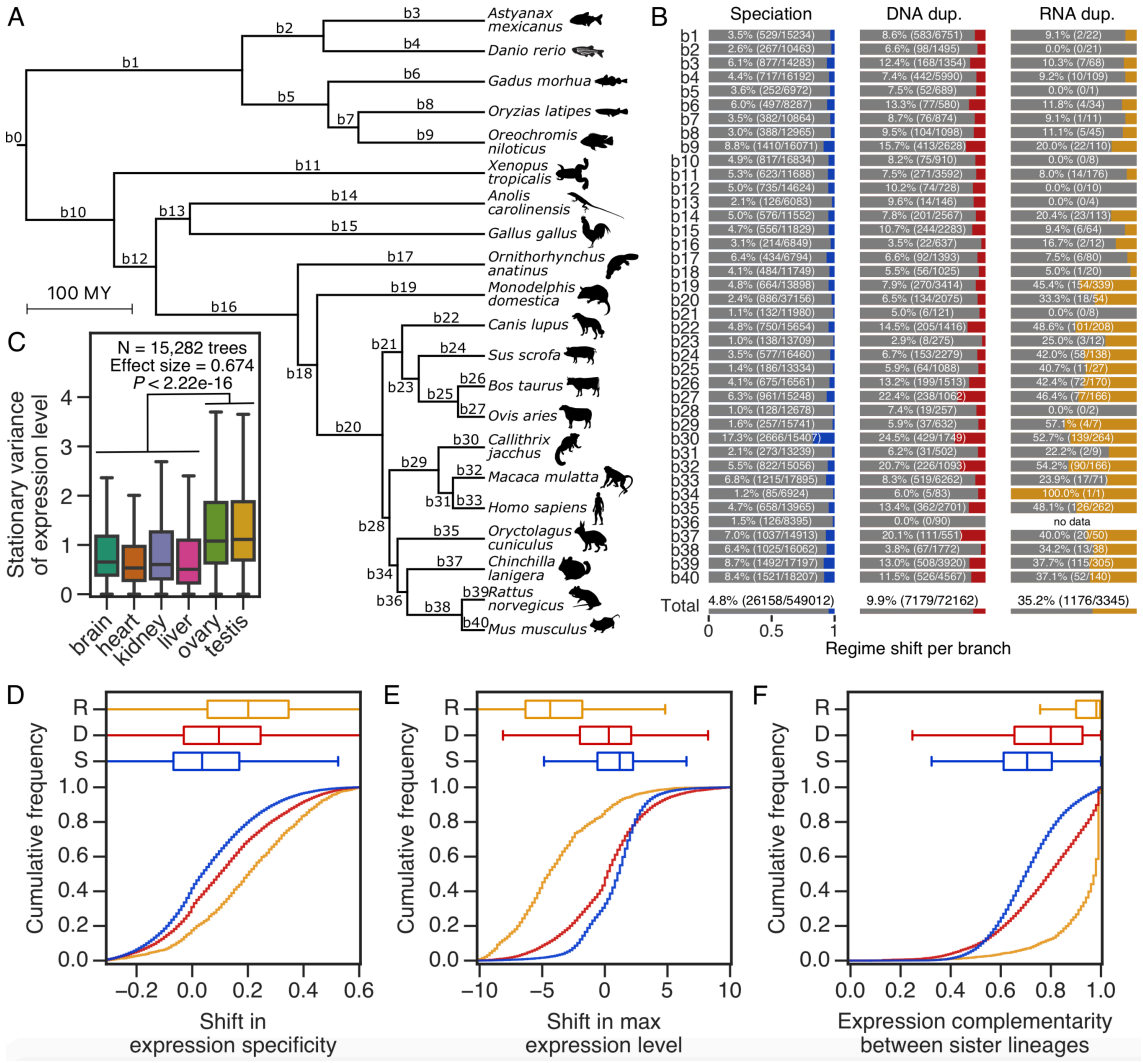


Figure 3. Characteristics of expression shifts in 15,282 gene trees. (**A**) The species tree showing analyzed genomes. (**B**) Mapping of 34,513 expression shifts in the species tree. The number and proportion of expression shifts in S, D, and R branches (see Fig. 2B) are shown. Corresponding branches in the species tree are indicated in **A**. Note that speciation only occurs at the left node of a branch in the species tree, whereas duplication can occur anywhere in the branch. (**C**) Organ-specific stationary variances of expression level evolution in vertebrates. (**D–F**) Cumulative frequency of change in expression specificity (**D**), change in maximum expression level (**E**), and expression complementarity between sister lineages (**F**) among detected expression shifts.

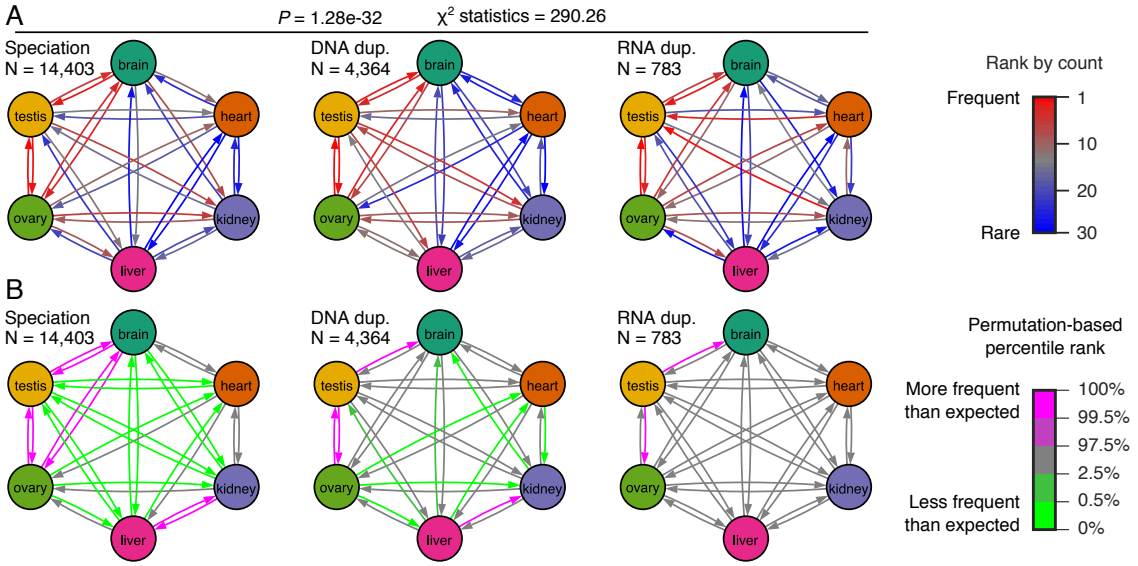


Fig. 4. Evolutionary dynamics of gene expression. (**A–B**) Transitions of primary-expressed organs (PEOs). Arrows represent transitions from ancestral PEOs to derived PEOs. Networks show the rank with the number of transitions (**A**) and its statistical significance based on 10,000 permutations (**B**). In **A**, the patterns of transitions among the three branching categories were statistically compared by a χ^2 test.

Supplementary Materials for

Amalgamated cross-species transcriptomes indicate organ-specific preadaptation for functional shifts in gene expression

Kenji Fukushima and David D. Pollock

List of Supplementary Materials:

Tables S1–S5 (separate file)

Figs. S1–S18

Supplementary Data (separate file)

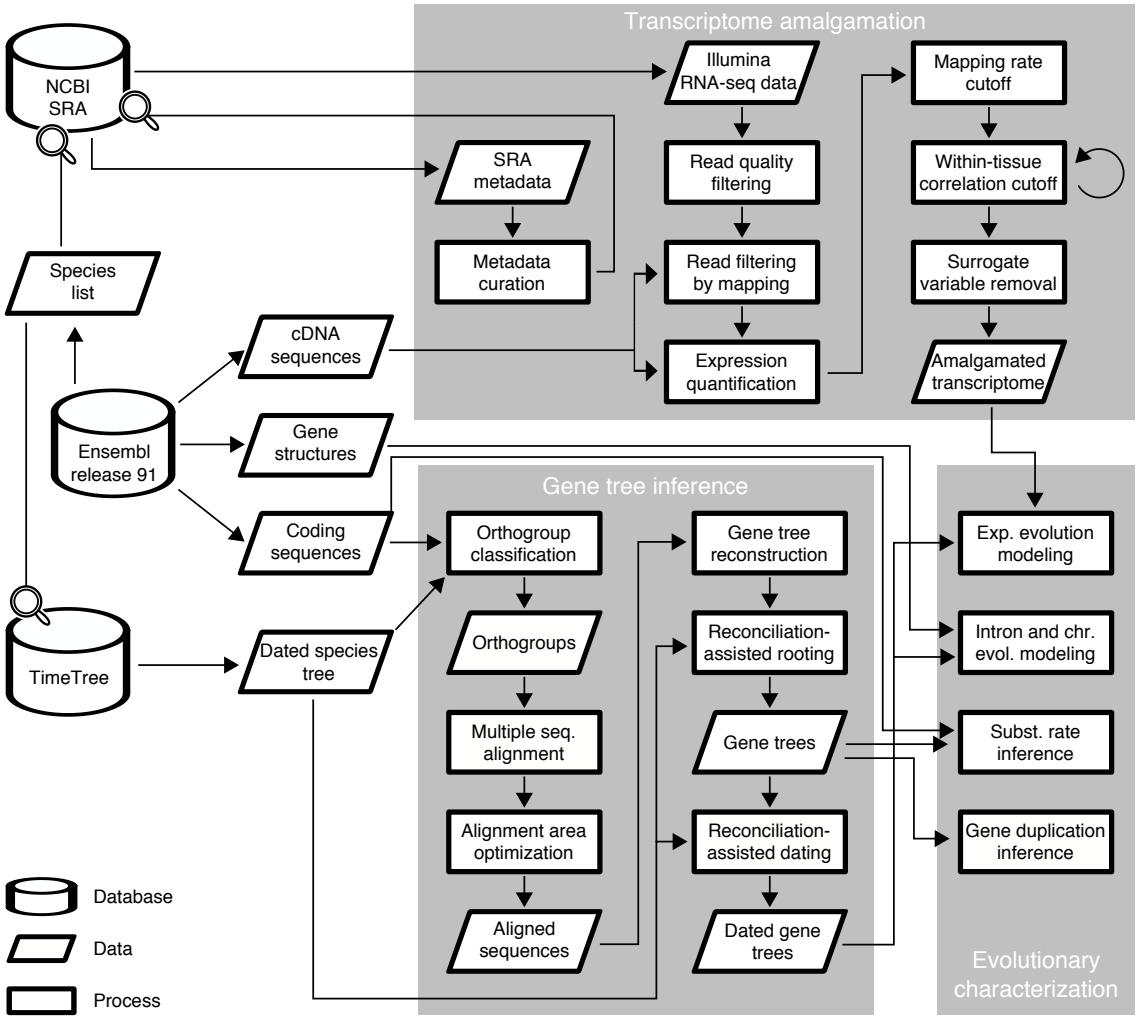


Fig. S1. A flow-chart of transcriptome amalgamation, gene tree inference, and evolutionary characterization in this study.

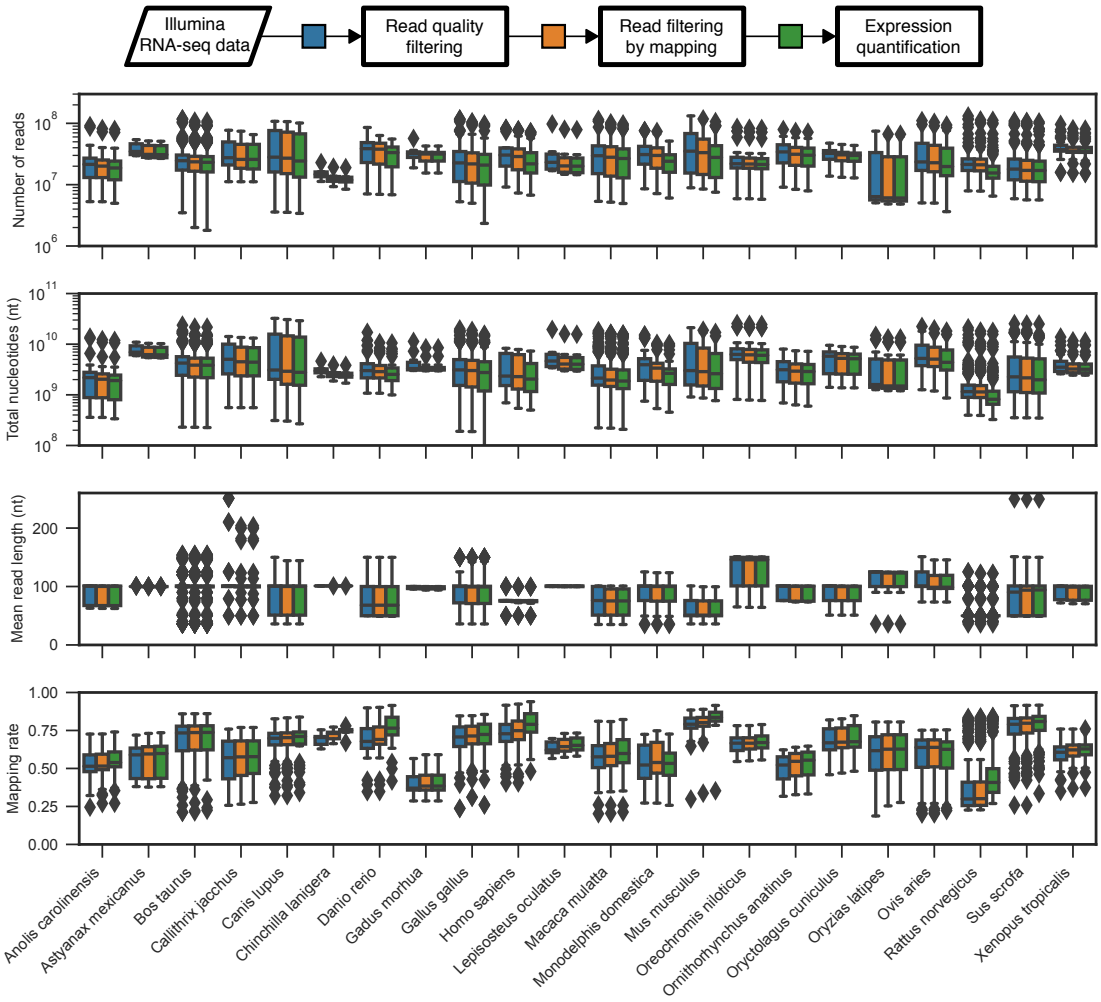
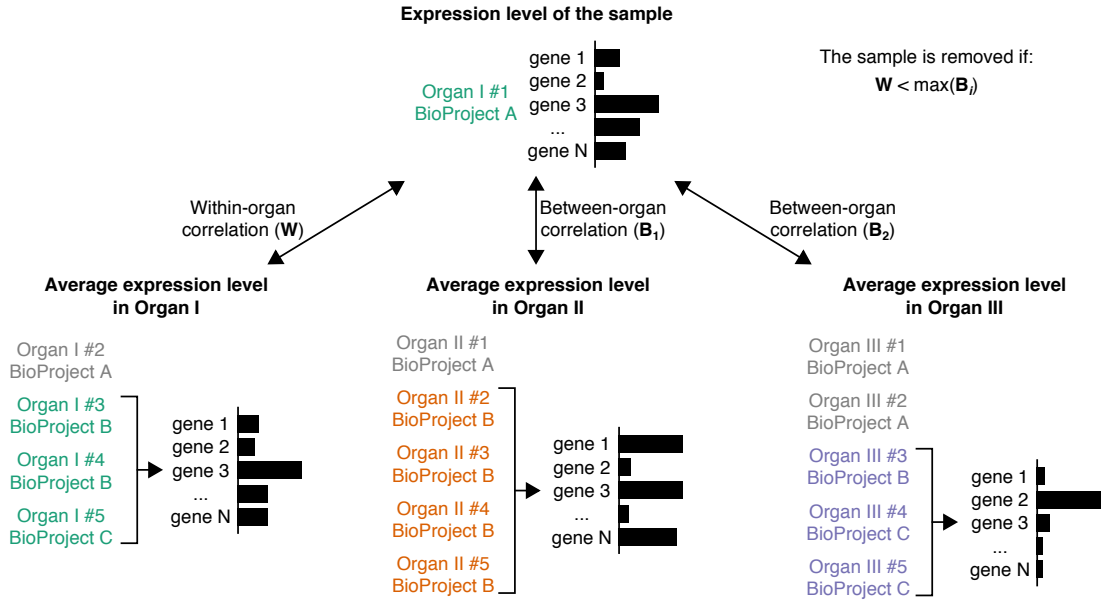


Fig. S2. Changes in number of reads, total nucleotide length, mean read length, and mapping rate by RNA-seq read filterings. Mapping rates were, in almost every case, improved by both read quality filtering and read filtering by mapping to miscellaneous genomic features.



714

715 **Fig. S3. A correlation analysis for the detection and removal of anomalous**
716 **RNA-seq samples.** Expression levels of all genes were compared between the sample
717 and the organ averages. A sample was removed if any between-organ comparisons
718 yielded a correlation coefficient higher than the within-organ comparison.

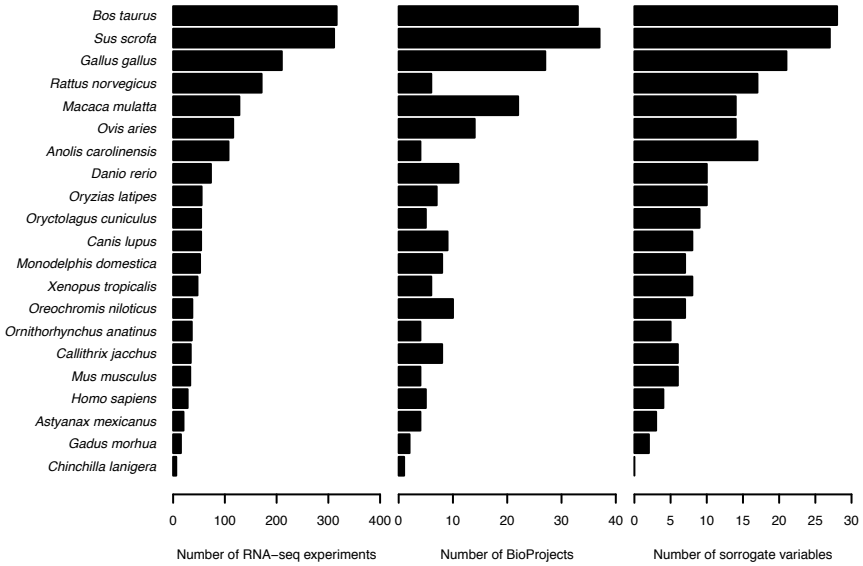


Fig. S4. The number of RNA-seq experiments, BioProjects, and detected surrogate variables in each species. The counts were derived from the final dataset. The numbers of surrogate variables are correlated with the numbers of RNA-seq experiments and BioProjects.

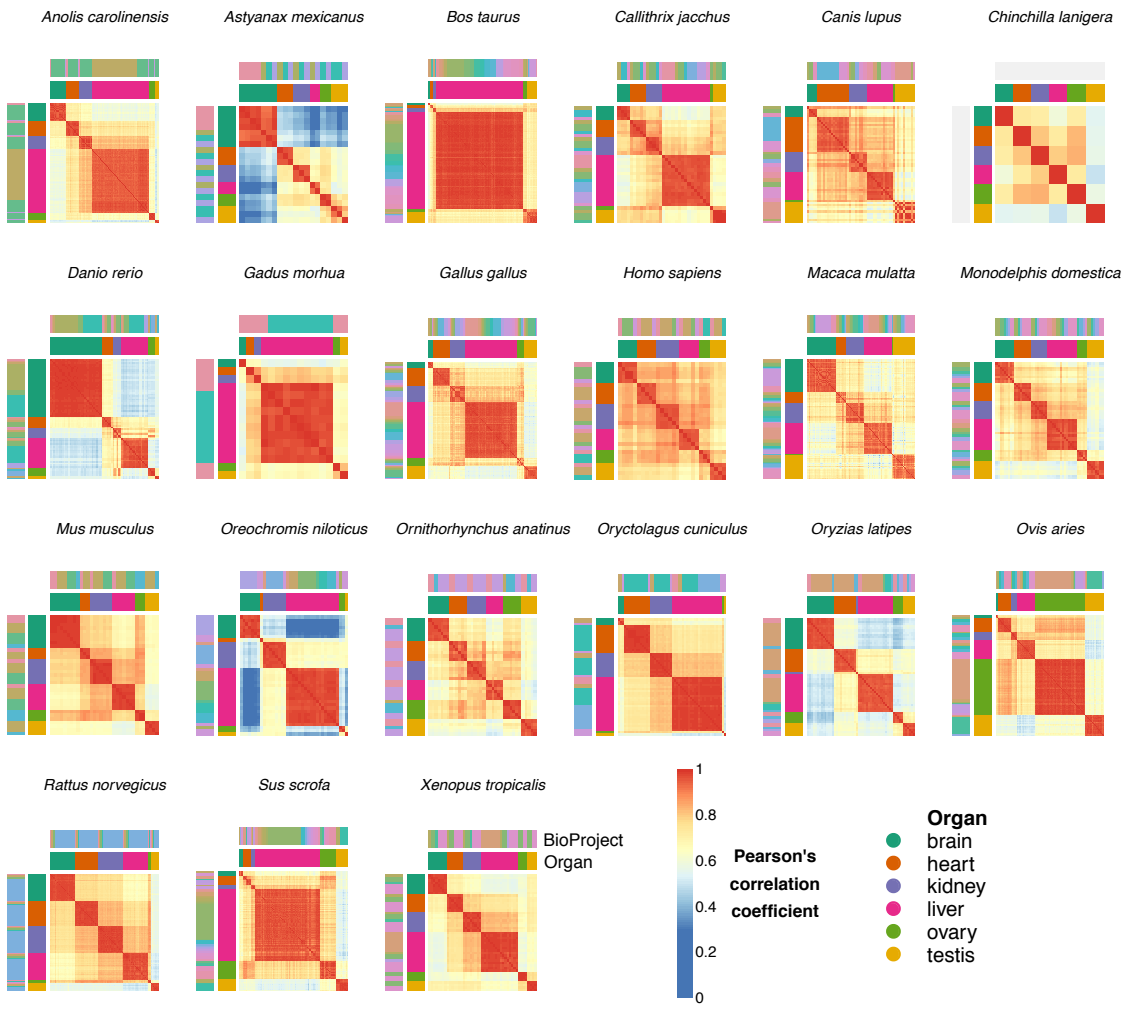


Fig. S5. Correlation heatmaps of corrected transcriptomes. See Supplementary Data for full descriptions including RNA-seq sample IDs and BioProject IDs.

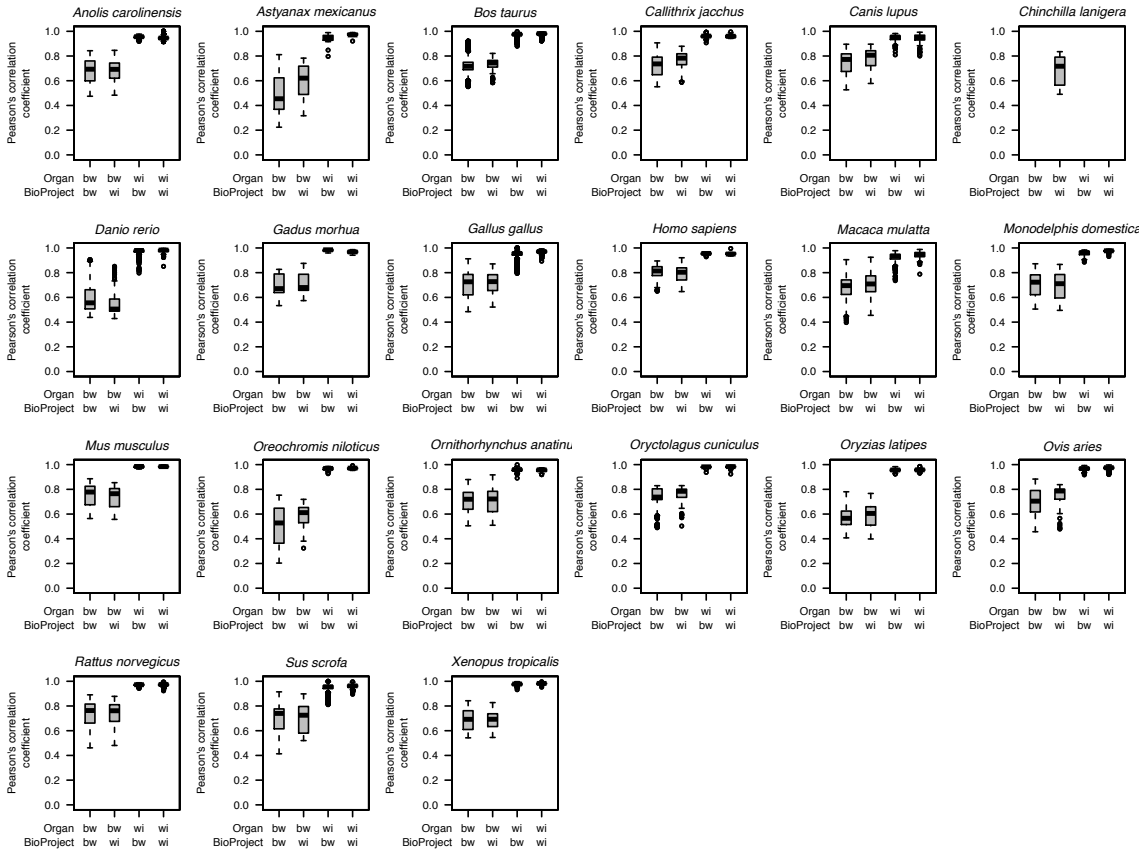


Fig. S6. Distinct distributions of Pearson's correlation coefficients depending on whether a pair of RNA-seq samples are the same organ or whether they are from the same research project. bw, between categories; wi, within category.

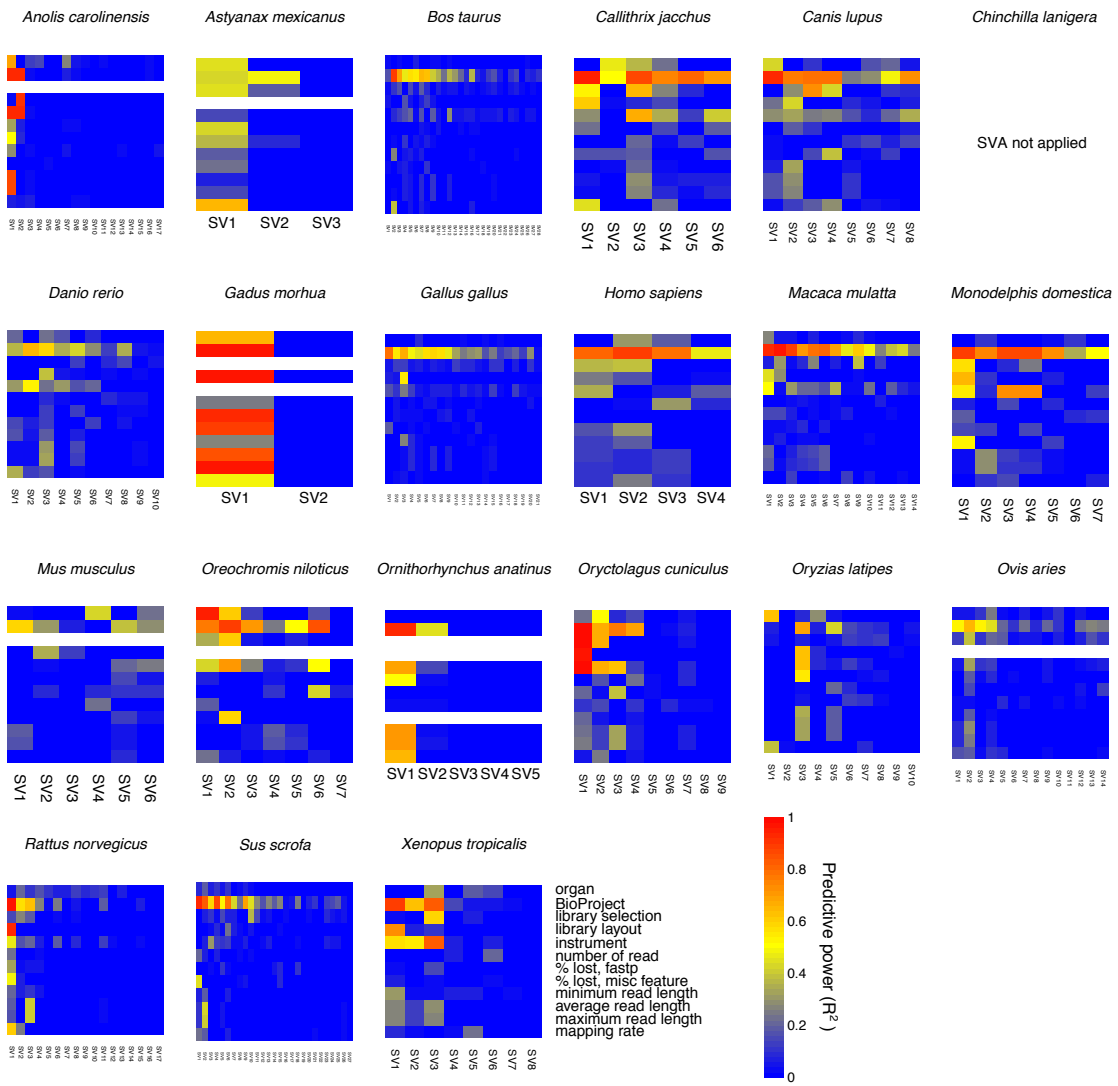


Fig. S7. Predictor analysis of detected surrogate variables. The predictive power was analyzed by linear regression using different properties of RNA-seq experiments: organ (brain, heart, kidney, liver, ovary, and testis), BioProject (e.g., PRJNA176589), library selection (e.g., cDNA and polyA), library layout (single and paired), instrument (e.g., Illumina HiSeq 2500 and NextSeq 550), number of read (e.g., 91,641,467 reads), % lost, fastp (percentage of reads that are removed by fastp; e.g., 5%), % lost, misc feature (percentage of reads that are mapped to non-nuclear-mRNA features and are removed from the analysis; e.g., 5%), minimum read length (e.g., 25 nt), average read length (e.g., 70 nt), maximum read length (e.g., 75 nt), and mapping rate (e.g., 80%). The predictors are summarized in Table S1.

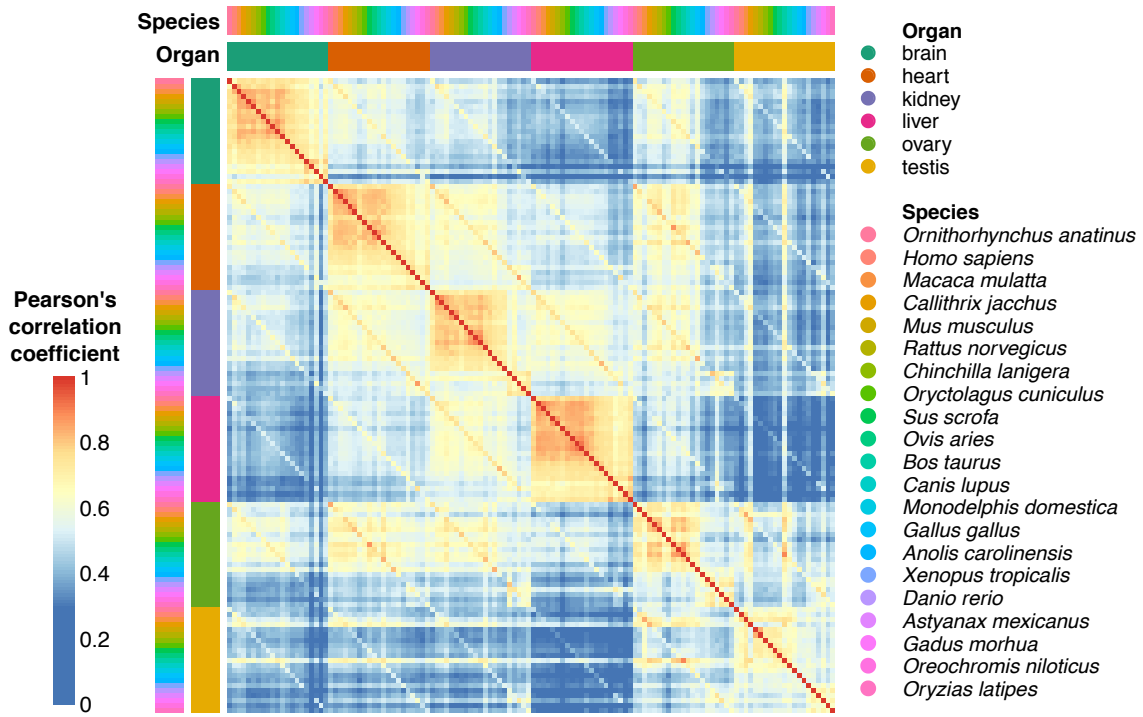


Fig. S8. Multispecies correlation analysis of averaged organ expression. Corrected expression levels of 1,377 conserved orthologs were used to calculate pairwise Pearson's correlation coefficients.

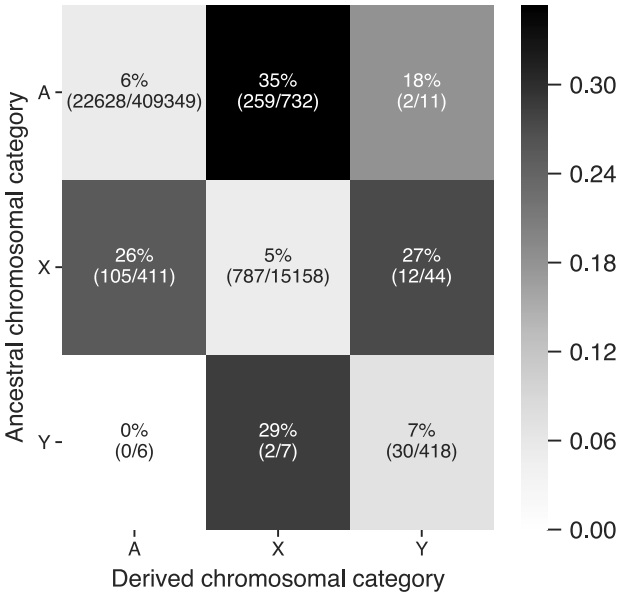


Fig. S9. The relationships between expression shifts and chromosomal location. The heatmap shows the frequency of expression shifts observed among the branches with or without a change in the chromosomal category (non-diagonal or diagonal, respectively). Chromosomal locations were categorized into autosomes (A), X chromosome (X) and Y chromosome (Y), and the ancestral locations were inferred by stochastic character mapping.

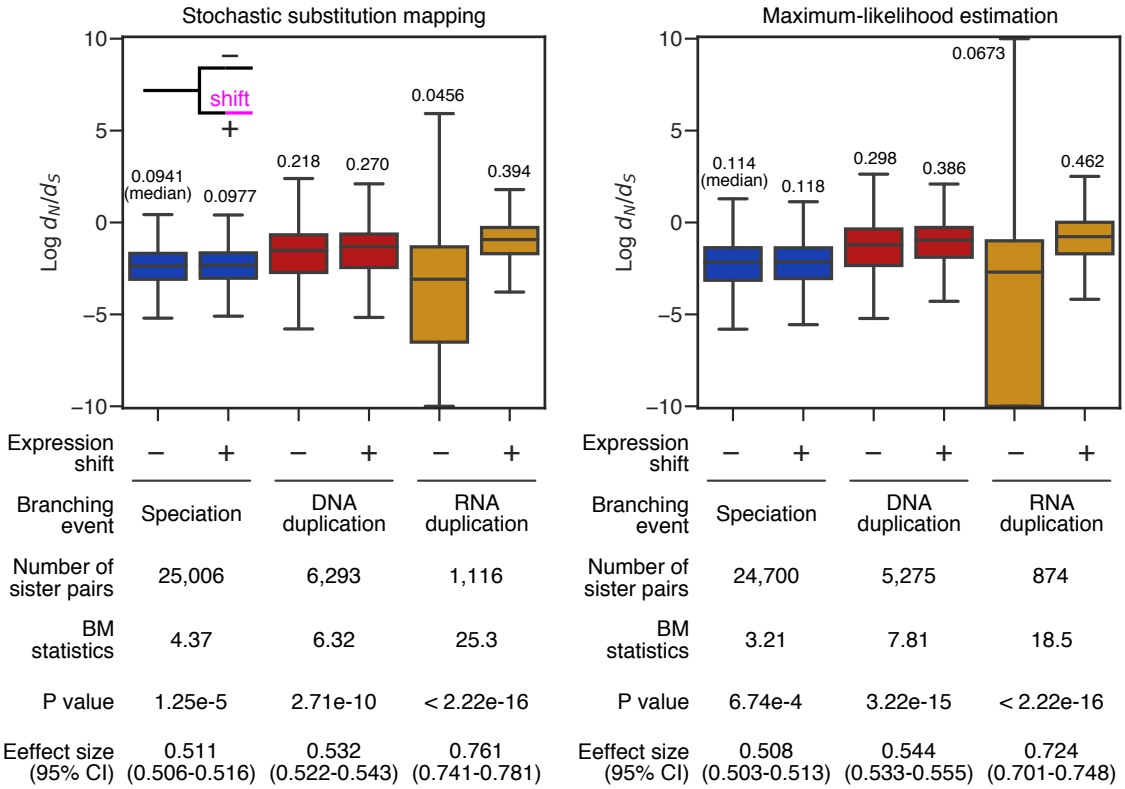


Fig. S10. Acceleration of protein evolution rate in correlation with expression regime shifts. Results from two different approaches are shown: stochastic substitution mapping (*mapdNdS*) and maximum-likelihood estimation (HyPhy). A plus (+) indicates branches with expression shifts, whereas minus (−) branches are sisters to the 'plus' branches. Statistical differences between pairs of distributions were tested using a two-sided Brunner–Munzel test (Brunner and Munzel, 2000). Non-log-transformed median values are shown above the boxplots. For visualization purpose, extreme values were clipped (minimum = −10; maximum = 10).

764

765 **Fig. S11. Frequency of organs in which highest expression level is observed before**
766 **and after expression shifts.** The stacked bar plot was generated from regime shifts
767 involving altered primary-expressed organs. The numbers of regime shifts are shown in
768 Fig. 4. S, speciation; D, DNA-based duplication; R, RNA-based duplication.

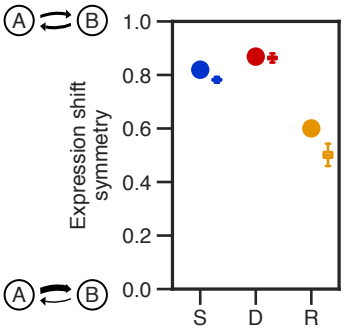


Fig. S12. The global polarity of transitions of primary-expressed organs. The global polarity is defined by the sum of asymmetry between two primary-expressed organs. Boxplots show 95% confidence intervals estimated by 10,000 bootstrap resamplings. The numbers of regime shifts are shown in Fig. 4. S, speciation; D, DNA-based duplication; R, RNA-based duplication.

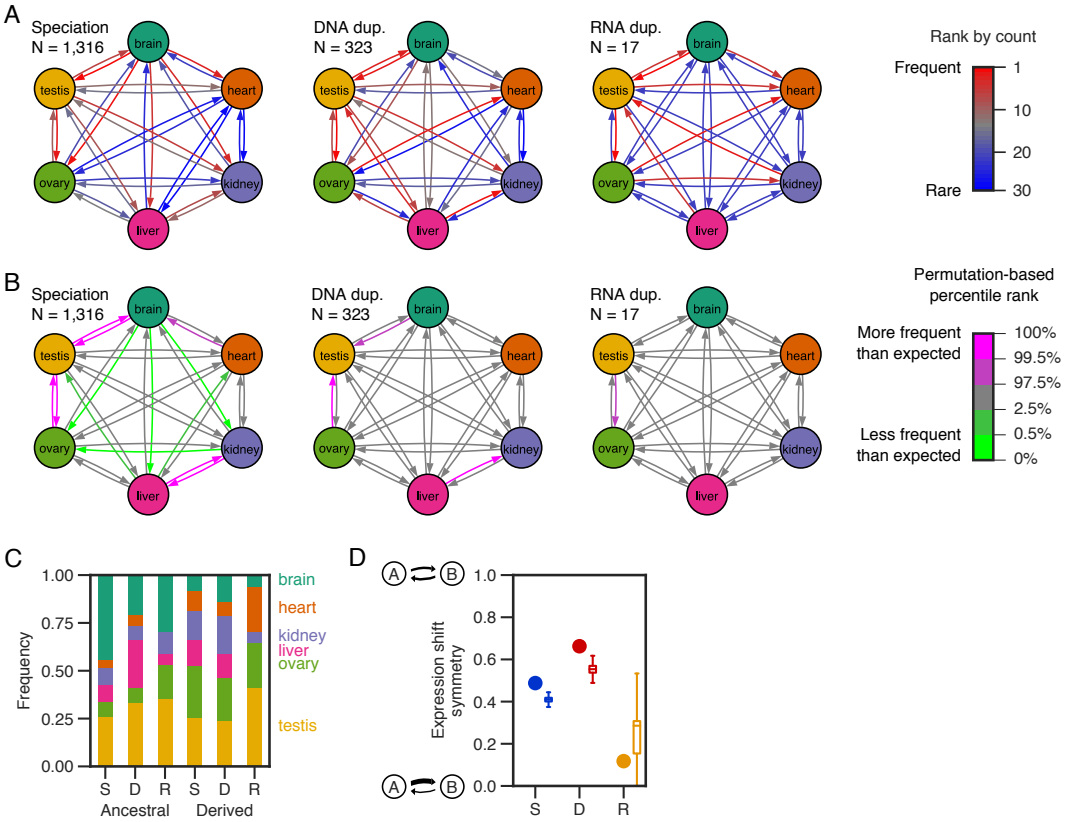


Fig. S13. Evolutionary dynamics of specifically expressed genes. Expression shifts with high ancestral and derived specificity (τ>0.67) are analyzed. The panels **A**, **B**, **C**, **D** correspond to Fig. 4A, Fig. 4B, Fig. S11, Fig. S12, respectively.

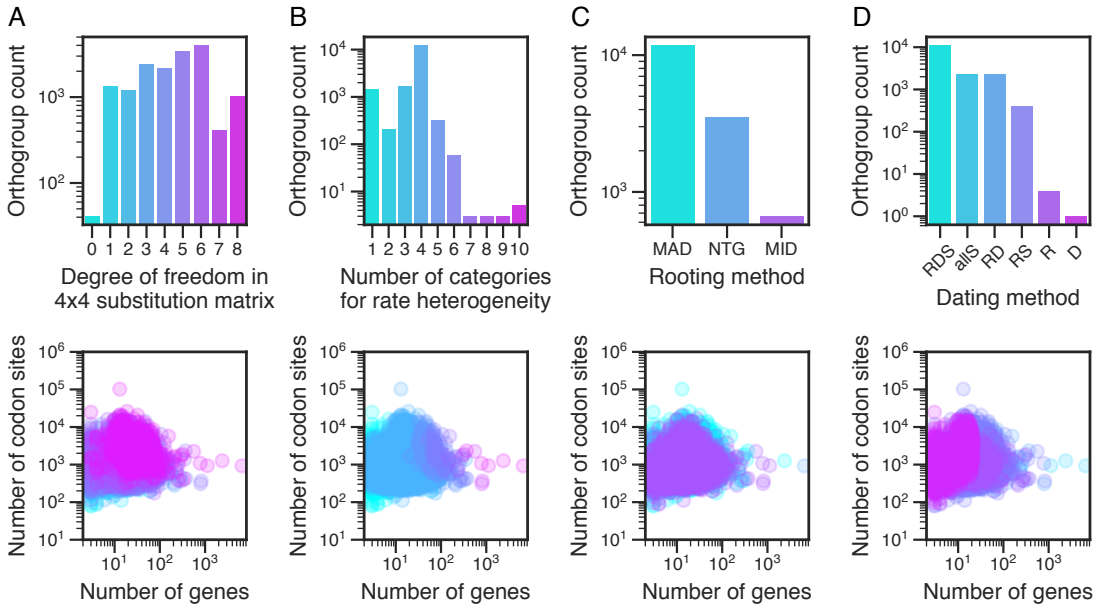


Fig. S14. Best-fit nucleotide substitution models, rooting methods, and constrained nodes for divergence time estimation in gene tree reconstructions. (**A**) Complexity of best-fit nucleotide substitution matrices. (**B**) Complexity of rate heterogeneity among nucleotide sites. Both discrete Gamma models (Yang, 1994) and FreeRate models (Soubrier et al., 2012; Yang, 1995) were included to count the number of categories for rate heterogeneity. (**C**) Selected rooting positions in reconciliation-assisted gene tree rooting. MAD, minimal ancestor deviation; NTG, 'rooting mode' of NOTUNG; MID, midpoint between the longest path. (**D**) Time-constrained nodes in tree dating. R, root node; S, speciation node; D, duplication node. All available constraints (RDS) are used in the first trial and then successively relaxed if estimation fails. In the category 'allS', all nodes are speciation nodes and therefore no divergence time estimation was performed for the trees.

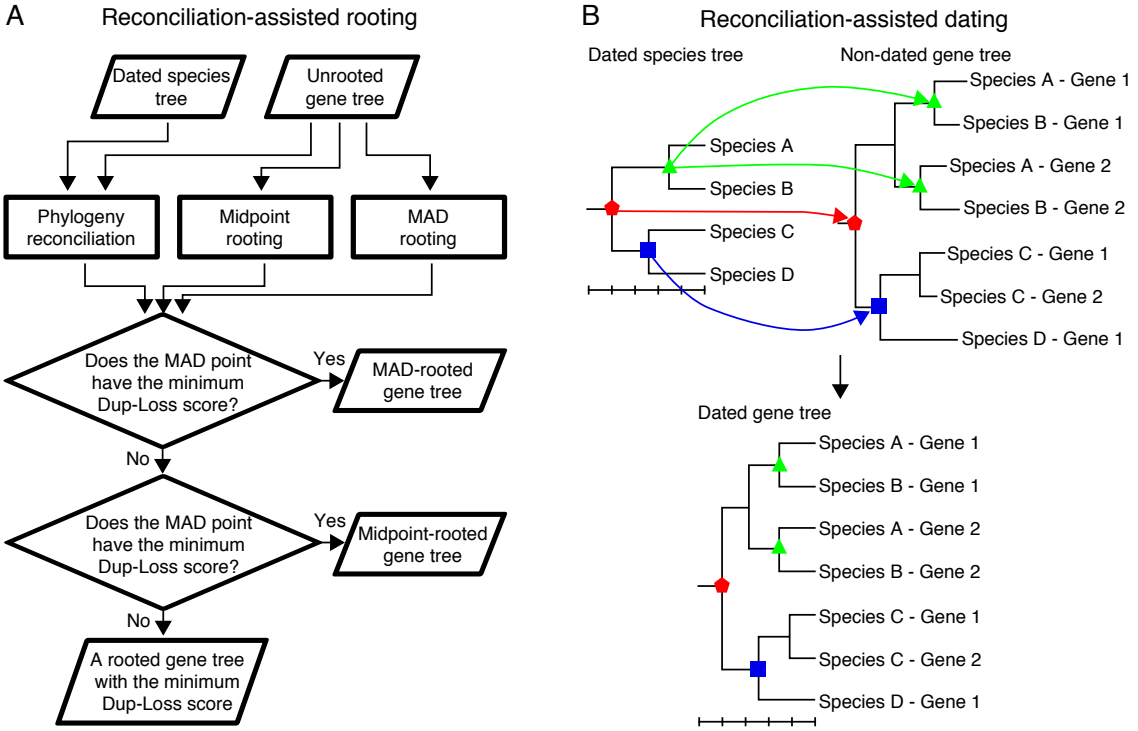


Fig. S15. Use of phylogeny reconciliation for gene tree rooting and dating. (**A**) Reconciliation-assisted gene tree rooting. Rooting points were estimated with two different methods: the minimum ancestor deviation (MAD) method and midpoint rooting. If they were compatible with the event parsimony involving gene duplication and loss, the MAD- or midpoint-rooted tree was reported in sequence. If not, one of event parsimony trees was reported. (**B**) Reconciliation-assisted gene tree dating. Speciation nodes in the dated species tree were mapped onto the non-dated gene tree by phylogeny reconciliation. By using those nodes as calibration points, the other node ages were estimated using the penalized likelihood method.

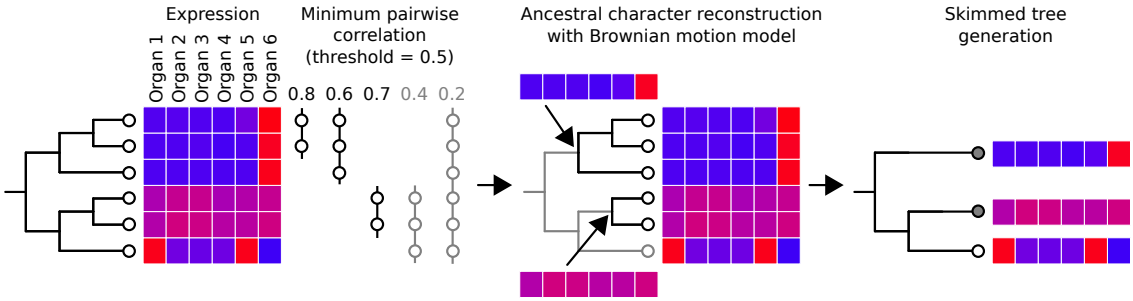


Fig. S16. Gene tree skimming. Gene tree clades are collapsed if character states are highly correlated. Resultant trees contain a smaller number of leaves than the original trees while preserving drastic changes in character evolution.

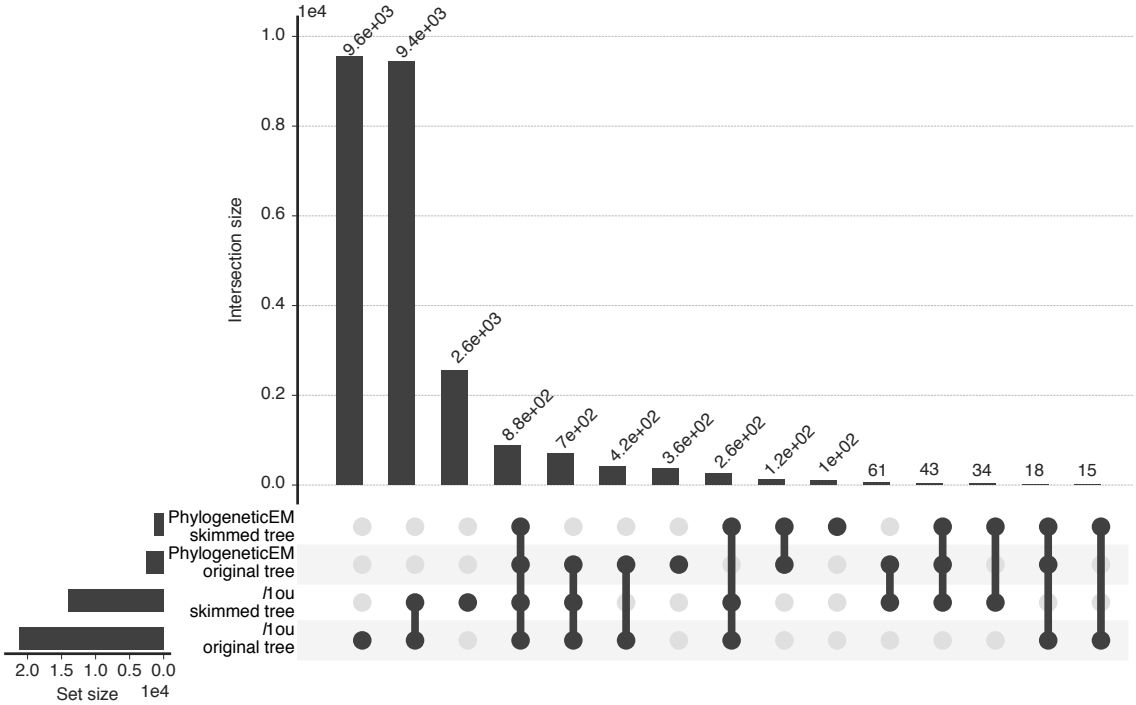


Fig. S17. The UpSet plot of four methods for expression shift detection. The overlap of expression shift branches is shown among those detected by *l*1ou or PhylogeneticEM in combination with the original trees or skimmed trees (Fig. S16). Results are shown for expression shift detection that was successfully completed on 3,862 orthogroups.

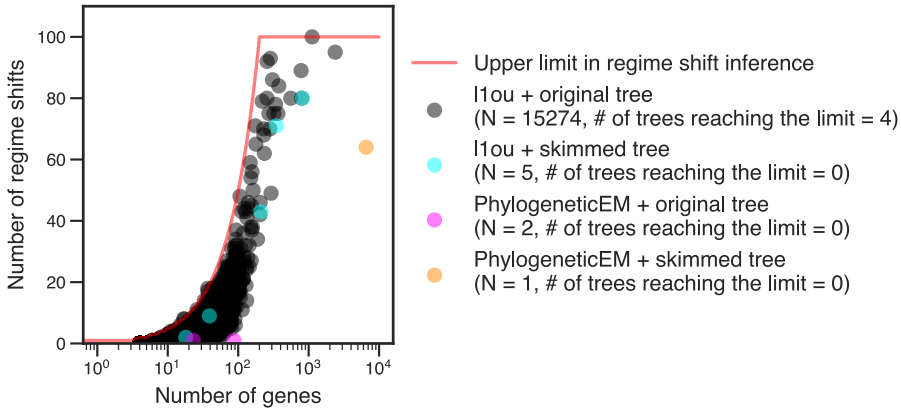


Fig. S18. The number of detected regime shifts in gene expression evolution. Each point corresponds to one gene tree. The red line shows the upper limits in the regime shift search.