Organ-specific propensity drives patterns of gene expression evolution

Kenji Fukushima[1,2*] and David D. Pollock[1,*]

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

[2]Present address: Institute for Molecular Plant Physiology and Biophysics, University of Würzburg

* Correspondence to: kenji.fukushima@uni-wuerzburg.de (K.F.) and david.pollock@cuanschutz.edu (D.D.P.)

Keywords: evolutionary transcriptomics, preadaptation, vertebrate evolution

Abstract:

The origins of multicellular physiology are tied to evolution of gene expression. Genes can shift expression as organisms evolve, but how ancestral expression influences altered descendant expression is not well understood. To examine this, we amalgamated 1,903 RNA-seq datasets from 182 research projects, including 6 organs in 21 vertebrate species. Quality control eliminated project-specific biases, and expression shifts were reconstructed using gene-family-wise phylogenetic Ornstein–Uhlenbeck models. Expression shifts following gene duplication result in more drastic changes in expression properties than shifts without gene duplication. The expression properties were tightly coupled with protein evolutionary rate, depending on whether and how gene duplication occurred. Fluxes in expression patterns among organs were nonrandom, forming modular connections which were reshaped by gene duplication. Thus, if expression shifted, ancestral expression in some organs induces a strong propensity for expression in particular organs in descendants. This supports a major role for what might be termed “preadaptive” pathways of gene expression evolution.

Introduction:

Vertebrate organs organize 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 correlates to 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). Major organ-altering evolutionary events such as development of the hominoid brain are also associated with gene expression shifts (Chen et al., 2013; Kaessmann, 2010; Zhang and Long, 2014; Zhang et al., 2011). However, although gene duplication is well-known to play an important role in expression pattern shifts (see e.g. the "ortholog conjecture" (Altenhoff et al., 2012; Castillo-Davis et al., 2004; Chen and Zhang, 2012; Rogozin et al., 2014)), the evolutionary dynamics of expression patterns with and without gene duplication remain poorly understood. An important question is whether long-term expression in one organ predisposes genes to be subsequently utilized in other organs.

A theoretical basis for such predisposition may be the idea that certain pre-existing adapted states are more conducive to evolution of specific new traits than other pre-existing states. This is known as preadaptation, and when a trait makes such a shift it is referred to as exaptation (Gould and Vrba, 1982). Evidence for preadaptation was long ago found in phenotypic traits (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). Protein sequence evolution generally involves highly epistatic interactions and context-dependent changes (Podgornaia and Laub, 2015; Starr et al., 2017) that affect preadaptation, but the modular nature of expression regulation (de-Leon and Davidson, 2007) makes it unclear whether preexisting expression patterns constrain evolutionary outcomes. By analyzing such influences on a genome-wide scale in this study, we aim to contribute some evidence towards this question.

Evolution of gene expression has been studied at genome-wide scales mainly using two distinct approaches: phylogenetic and pairwise analyses. Phylogenetic approaches model gene expression dynamics and infer ancestral expression patterns in the context of gene phylogenies. For example, Brownian motion models embody purely neutral expression evolution (Bedford and Hartl, 2009), whereas Ornstein–Uhlenbeck models are designed to detect purifying selection and adaptive evolution along with neutral

fluctuation (Brawand et al., 2011; Chen et al., 2019; Rohlfs et al., 2014). Although each gene family has a distinct evolutionary history, a species phylogeny is often used for the sake of simplicity. Because such approximations cannot be applied to gene families with lineage-specific gene duplications and losses, its application has mostly been limited to single-copy genes. In contrast, pairwise analysis compares gene expression between paralogs in single species (Assis and Bachtrog, 2015; Lan and Pritchard, 2016) or between orthologs or paralogs in pairs of species (Chen and Zhang, 2012; Guschanski et al., 2017; Kryuchkova-Mostacci et al., 2016; Rogozin et al., 2014; Warnefors and Kaessmann, 2013). Although pairwise approaches can evaluate the effect of gene duplications, ancestral expression cannot be inferred.

To infer the adaptive evolution of gene expression in diverse gene families, we applied Ornstein–Uhlenbeck models for complex gene family phylogenies containing gene duplications and losses, without assuming species phylogeny. We also developed a curation pipeline to amalgamate large amounts of transcriptome data from many studies for a better phylogenetic resolution. Using these methods, we ask how expression changes link to gene duplication and the rate of protein evolution, and whether previous patterns of gene expression constrain later evolution, what we will call the "preadaptation hypothesis".

83 **Results:**

84

85 **Duplication-permissive genome-wide analysis of gene families.** To allow
86 evolutionary expression analysis on a broad set of genes, we used a phylogenetic
87 approach that deals with the complex history of gene family trees with duplications and
88 losses, and applied it to 21 tetrapod genomes (Fig. S1). A major challenge in using gene
89 trees was divergence time estimation, a prerequisite for applying phylogenetic
90 comparative methods. We overcame this problem by incorporating phylogeny
91 reconciliation in estimating divergence time of gene trees. Gene divergence nodes were
92 constrained by the corresponding divergence times in a known species tree, and
93 duplication nodes were constrained by ancestral and descendant speciation events (see
94 Materials and Methods for details). Because we estimated individual gene phylogenies
95 rather than using a single species phylogeny, we could analyze gene families that
96 included many lineage-specific gene duplications and losses, making our study less
97 biased towards conserved genes with slow gene turnovers (Guschanski et al., 2017).
98 Use of gene family trees also allowed us to include many organ-specific genes that are
99 enriched in lineage-specific and young duplications (Guschanski et al., 2017). There
100 were only 1,377 single-copy orthologs for which the species phylogeny was applicable,
101 but we were able to include 15,475 genes per species on average (including 20,873
102 human genes, merged into 15,280 gene families). This approach eliminates problems
103 with pairwise analyses that ignore phylogenetic tree structure, and allowed us to infer
104 expression at ancestral nodes in the tree.

105

106 **Transcriptome amalgamation.** To attain high resolution in our analyses, we
107 amalgamated 1,903 RNA-seq experiments from 182 research projects (i.e., 182
108 BioProject IDs in the NCBI SRA database) and generated a dataset covering 6 organs
109 from 21 vertebrate species without missing data (Tables S1–S2; Fig. S1). In comparison,
110 other recent comparative transcriptomic analyses of vertebrates (Barbosa-Morais et al.,
111 2012; Breschi et al., 2016; Carelli et al., 2016; Chen et al., 2019; Cortez et al., 2014;
112 Guschanski et al., 2017; Julien et al., 2012; Necsulea et al., 2014; Warnefors and
113 Kaessmann, 2013) often used the same dataset containing 131 RNA-seq experiments
114 from 6 organs and 10 species (Brawand et al., 2011), with some additional data in
115 different studies. RNA-seq reads were first mapped to corresponding reference genomes

116 and then the expression level was quantified by two metrics: transcripts per million
117 (TPM) and fragments per kilobase million normalized by trimmed mean of M-values
118 (Robinson and Oshlack, 2010) (TMM-FPKM). To reduce the among-species variation,
119 the TMM normalization was applied across all 1,903 samples using the 1,377 single-
120 copy orthologs.

121 To allow rapid integrated analysis of datasets, we employed automated multi-
122 aspect quality controls, including metadata curation (Table S3), sequence read filtering
123 (Fig. S2), and iterative removal of anomalous RNA-seq samples by monitoring
124 correlations between and within data categories (Fig. 1A; Fig. S3). The metadata
125 curation enabled us to select appropriate samples from the NCBI SRA database. Data
126 that were not compatible with those from other research projects (defined by BioProject
127 ID) were removed in the correlation analysis by implementing a majority rule (Table
128 S3), resulting in a cleaned dataset. This filtering step was designed to fulfill the
129 assumption that any samples from the same organ should correlate better than samples
130 from different organs within species. When anomalous data were detected, all samples
131 belonging to the same research projects (i.e., the same BioProject ID) were discarded.

132 Finally, we applied surrogate variable analysis (SVA) (Leek and Storey, 2007) to
133 detect and correct hidden biases likely originating from heterogeneous sampling
134 conditions and sequencing procedures among experiments in both $\log_2$-scaled TPM and
135 TMM-FPKM (SVA-log-TPM and SVA-log-TMM-FPKM, respectively; Fig. S4A–B).
136 This correction greatly improved the correlation of expression levels within organs from
137 the same species, even when data were derived from different research projects (Fig.
138 1B–C; Fig. S4C–F). Among surrogate variables, BioProject IDs tend to show a high
139 predictive power, suggesting project-specific sources of bias (Fig. 1D; Fig. S4G–H).
140 Although the inclusion of many species from phylogenetically diverse lineages makes it
141 difficult to extract organ-wise characteristics from the limited number of single-copy
142 orthologs, a principal component analysis produced moderate organ-wise segregation in
143 the multispecies comparison (Fig. 1E; Fig. S4I–K), further indicating that the curated
144 dataset is sufficiently reliable for use in cross-species expression pattern analyses. The
145 previously-reported uniqueness of testis transcriptomes (Brawand et al., 2011) was
146 partly resolved as the third principal component (Fig. S4K).

147 To further evaluate the validity of amalgamated transcriptomes, we analyzed the
148 expression of community-curated cell-type-specific marker genes associated with

149 organs in PanglaoDB (Franzén et al., 2019), which organizes a number of single-cell
150 RNA-seq experiments in human and mouse. We compared median values of log-
151 transformed expression levels of >100 marker genes in each organ (Fig. S5). After SVA
152 correction, all RNA-seq samples in the both species showed the corresponding marker
153 expression values higher than those from the other organs, suggesting our amalgamated
154 transcriptomes preserve the organ-specific gene expression. In the cell-type-wise
155 analysis, a few cases, such as juxtaglomerular cells in kidney and hepatic stellate cells
156 in liver, could not resolve our organ-wise transcriptomes (Supplementary Data).
157 However, such low performance was seen in all samples rather than subsets associated
158 with particular BioProject IDs, suggesting that the dissection decisions have negligible
159 effects to cell type compositions in the organs.
160 In addition to the better phylogenetic coverage, greater accuracy of estimated
161 expression levels is another possible advantage of integrating many RNA-seq datasets.
162 This idea is supported by subsampling analysis on a housekeeping gene GAPDH
163 (glyceraldehyde-3-phosphate dehydrogenase), where, as more data are used, estimated
164 expression levels tend to quickly converge to a final value (Fig. 1F; Fig. S4L).
165
166 **Modeling expression evolution.** We next used the amalgamated transcriptomes to
167 evaluate how expression evolved along 15,280 maximum-likelihood gene family
168 phylogenies (Table S4), employing multi-optima Ornstein–Uhlenbeck (OU) models
169 (Hansen, 1997) to allow for possible adaptive shifts of optimal expression levels and
170 neutral fluctuations (Brawand et al., 2011; Chen et al., 2019; Rohlfs et al., 2014). This
171 modeling identified statistically supported expression 'regime shifts' (Hansen, 1997;
172 Khabbazian et al., 2016) on each gene tree (Fig. 2A), which were then analyzed in the
173 context of preceding duplication events. 'Speciation' events (S node; Fig. 2B) with no
174 duplication were considered the 'baseline' mode of expression evolution because
175 regulatory environments and expression patterns are more preserved among orthologous
176 genes in comparison with paralogous genes produced by gene duplication (Altenhoff et
177 al., 2012; Castillo-Davis et al., 2004; Chen and Zhang, 2012; Rogozin et al., 2014).
178 Because OU shift detection has been applied for gene expression by assuming species
179 tree phylogeny in single-copy genes, shifts in S branches are equivalent to those
180 characterized previously (Brawand et al., 2011; Chen et al., 2019) but also includes
181 many more speciation events in duplication-prone gene families. Gene tree nodes

182 associated with preceding duplication events were categorized as DNA-based
183 duplication or retrotransposition (D or R nodes, respectively) depending on complete
184 intron losses (Fig. 2B).
185 Organ-wise means of the two expression values, SVA-log-TPM and SVA-log-
186 TMM-FPKM, were separately used to model expression evolution with OU processes.
187 The two analyses resulted in similar numbers and characteristics of expression regime
188 shifts (Fig. S6), but the shift locations were sometimes inconsistent. S branches were a
189 major source of apparently inconsistent regime shifts, whereas branches following
190 duplication (D and R) showed largely consistent detection between the two metrics (Fig.
191 2C). Although the inclusion of inconsistently detected branches did not change the
192 results, we retained only consistently detected regime shifts for all downstream analysis
193 to draw a more robust conclusion. While SVA-log-TMM-FPKM values were reported
194 in the main text unless otherwise mentioned, the comparisons with SVA-log-TPM-
195 based analyses are available as Supplementary Information.
196 As an example of this analysis, an orthogroup of phosphoglycerol kinases (PGKs),
197 containing all three categories of branching events followed by expression shifts (S, D,
198 and R), is shown in Fig. 2D. This protein catalyzes the first ATP-generating step in the
199 glycolytic pathway and is required for most cell types including sperms (Danshina et al.,
200 2010; Liu et al., 2015). PGK1, the original copy on the X chromosome, is known to
201 have duplicated independently in eutherians and marsupials to produce the autosomal
202 retrocopy PGK2 that compensates the protein activity during X-inactivation (Boer et al.,
203 1987; McCarrey and Thomas, 1987; Potrzebowski et al., 2008). Our automated analysis
204 correctly recovered both retrotranspositions as well as the subsequent gains of testis-
205 specific expression in eutherian and marsupial lineages. This illustrates that our
206 automated genome-wide analysis can recover evolutionary trajectories that are
207 compatible with focused single gene family analyses (see Supplementary Data for
208 individual gene trees).
209
210 **Duplication-specific effects in expression evolution.** Across gene trees, expression
211 regime shifts were relatively rare in S branches, for a probability of 2.2% per branch,
212 and an average rate of 5.1×10^{-4} shifts/MY (million years) (Fig. 3A–B). In agreement
213 with the idea that gene duplication tends to free genes up for functional divergence and
214 enhance long-term retention of duplicated copies (Assis and Bachtrog, 2015; Huerta-

215 Cepas et al., 2011), the frequency of regime shifts in D branches was four times as
216 much per branch (9.0% per branch, at a rate of 1.9×10^{-3} shifts/MY across all genes and
217 all D branches). Thus, although far fewer branches are preceded by DNA-based
218 duplication events (72,476 branches) than speciation events (556,240 branches), D
219 branches account for over 33% of all regime shifts consistently detected by the two
220 expression measures. We note that this result reinforces previous results on the
221 "ortholog conjecture", the idea that duplicated gene copies (paralogs) are more prone to
222 expression shifts than orthologs (Altenhoff et al., 2012; Castillo-Davis et al., 2004;
223 Chen and Zhang, 2012; Rogozin et al., 2014). R branches were far more likely to result
224 in expression regime shifts (37.3% per branch, at a rate of 7.6×10^{-3} shifts/MY), but with
225 only 2,985 R branches, this resulted in only 1,107 shifts (5.6% of the total).
226 Translocated genes are more likely to shift expression than those that do not (Fig. S7),
227 in line with previous observations from the human genome (Lan and Pritchard, 2016).
228 While the expression shift frequency in S and D branches varies slightly across the
229 phylogeny, R branches showed much stronger among-lineage heterogeneity, and had a
230 particularly high frequency in the mammalian lineage (Fig. 3B). These
231 retrotransposition-related expression changes may be related to the variation in the
232 retrotransposition rate itself, which is known to vary across lineages (Marques et al.,
233 2005; Yu et al., 2007). Among-species heterogeneity in gene prediction quality may
234 also be attributed to this pattern because early-diverging species tended to show higher
235 percentages of missing single-copy orthologs than those in mammalian species (Fig. S8;
236 but see *Danio rerio*, *Oreochromis niloticus*, and *Oryctolagus cuniculus* as
237 counterexamples). In the absence of regime shifts, expression levels varied most in
238 ovary and testes, which had significantly higher average stationary variances than the
239 other four organs on the basis of tree-wise stationary variance (Fig. 3C). This supports
240 previously observed high variation of gene expression in testes (Brawand et al., 2011)
241 and extends it to the reproductive organs of both sexes.

242 If expression regime shifts are due to functional divergence, it is highly relevant to
243 characterize how expression properties changed from ancestral to derived regimes. To
244 do this, we examined changes in organ expression specificity, maximum expression
245 level, and organ expression complementarity. The specificity measure τ ranges from 0
246 for uniformly expressed genes to 1 for genes with highly specific expression (Yanai et
247 al., 2005). The distributions of regime shifts in D and R branches are shifted towards

248 greater organ specificity compared to shifts in S branches, with R branches creating the
249 most specific expression (Fig. 3D). To characterize the 'on state' transcriptional activity,
250 we analyzed the maximum fitted expression levels among the six organs (μ_{max}). D and
251 R branches appear to be enriched for down-regulation compared to S branches (Fig. 3E).
252 Complementarity of organ expression patterns was measured to evaluate the
253 differentiation between a pair of sister branches. We used a metric on the fitted organ-
254 wise expression levels (μ) called 'TEC', which ranges from 0 for completely
255 overlapping expression to 1 for mutually exclusive patterns (Huerta-Cepas et al., 2011).
256 Nearly all branches with regime shifts (95%) had complementarity values greater than
257 0.5, indicating that most regime shifts detected involve differentiation of expression
258 patterns, rather than overall up- or down-regulation. Regime shifts in D and R branches
259 often had more complementarity expression than those in S branches (Fig. 3F), further
260 supporting the role of gene duplication in functional differentiation. The more drastic
261 effect in R branches probably reflects the regular loss of regulatory elements in
262 retrotranspositions, whereas DNA-based duplication can more often retain regulatory
263 regions (Babushok et al., 2007). Jointly, these results indicate that, compared with the
264 baseline from speciation-associated shifts, gene duplication tend to produce more organ-
265 specific, more often down-regulated, and more differentiated expression patterns.
266 Although the down regulation may be explainable by a tendency to need less of the
267 newly functional expression regime, it may also be explained by either recent
268 nonfunctionalization (Balakirev and Ayala, 2003; Mighell et al., 2000) or specialized
269 expression in organs that were not part of this analysis.
270
271 **Context-dependent change in the rate of protein evolution.** Change in gene
272 expression can be accompanied by accelerated or decelerated protein evolution, which
273 may be detected by change in the ratio of nonsynonymous/synonymous substitutions
274 (*dN/dS* or ω) along branches. In D and R branches, median ω values are more than
275 double the baseline seen in S branches (Fig. 4A; Fig. S9A), again supporting the
276 ortholog conjecture and the idea that gene duplication tends to free genes up for
277 functional divergence (Assis and Bachtrog, 2015; Huerta-Cepas et al., 2011). Within
278 each of S, D, and R branch categories, branches with expression regime shifts
279 accompany an increased ω compared to sister branches (Fig. 4A). The increased rate
280 was quite small in S branches (median ω, 0.096 in branches with shifts versus 0.102 in

281 sister branches), much bigger in D branches (0.182 versus 0.244), and huge in R
282 branches (0.036 versus 0.394). If the changes in expression and the rate of protein
283 evolution are due to functional changes, this may indicate that functional divergence is
284 sometimes effected by joint changes in expression and accelerated protein evolution.
285 Although 52.3% of all branches with expression regime shifts had higher ω relative
286 to sister branches (ω ratio), 27.5% of sister branch pairs are relatively undifferentiated,
287 with differences in ω within ±5%, and 42.5% had lower ω in the branches with
288 expression shifts. This finding led us to hypothesize that the direction of the rate
289 changes in protein evolution is linked to how, rather than whether, expression is
290 changed. Strikingly, we found a context-dependent association between protein and
291 expression evolution (Fig. 4B; Fig. S9B). Increased ω ratio was linked strongly to
292 increased, rather than decreased, organ expression specificity in S and D branches,
293 potentially reflecting adaptive evolution coupled with specialized expression. However,
294 it was in turn highly associated with decreased specificity in R branches, which may be
295 explained by frequent gene decay in unsuccessful retro-copies. The change in maximum
296 expression level was overall negatively correlated with ω ratio, but this link was
297 stronger in down-regulation compared with up-regulation, except for D branches where
298 the ω ratio was smaller when $\Delta\mu_{max}$ was larger (Fig. 4B). It has been reported that high
299 expression slows protein evolution (Zhang and Li, 2004), and our results suggest that
300 DNA-based duplication creates such constraints when accompanied by up-regulation.
301 The organ expression complementarity between sister lineages was positively correlated
302 with ω ratio (Fig. 4B), and its association was strongest in R branches, suggesting that
303 protein evolution accelerates as gene expression patterns differentiate from their
304 ancestral state. Collectively, these results suggest that protein evolution rate is linked to
305 changes in expression properties through a complicated association, which masks their
306 relationships in a global, unstratified analysis, and potentially explain a previous report
307 of no strong relation (Warnefors and Kaessmann, 2013).
308
309 **Organ-specific propensity in gene expression evolution.** The preadaptation
310 hypothesis predicts that the ancestral organ expression prior to the shifts will affect
311 which organs are likely to become the target of newly specific expression. To assess this,
312 we tested whether expression shifts are random with respect to change in expression

313 from one organ to another, by characterizing the organ in which genes are most highly
314 expressed (PEO, primary-expressed organ).
315 Across vertebrates, switching from one PEO to another was detected in 6,886,
316 3,586, and 746 regime shifts in S, D, and R branches, respectively. The gain/loss ratios
317 are heterogeneous among organs (Fig. 5A), suggesting that vertebrate organs serve as
318 both sources and sinks in expression evolution, but that their relative contributions are
319 organ-specific. Although S, D, and R branches shared a global trend of relatively
320 abundant testis-related PEO shifts, their distributions are largely different. D branches
321 were moderately similar to both S and R branches (Spearman's $\rho \sim 0.6$), but S and R
322 branches were dissimilar ($\rho = 0.28$). This pattern, including the abundant shifts related
323 to testis, was robust against the correction by the organ-wise numbers of expressed
324 genes (Fig. S10). This result suggests a role for gene duplication, including by
325 retrotranspositions, in remodeling the among-organ flow of expressed genes.
326 Controlling the total number of shifts from and to each PEO, some PEO shifts are
327 significantly different from the random expectation (Fig. 5B). There are clear patterns
328 of evolutionary transitions that are statistically supported by independent OU modeling
329 of the two expression metrics. In S branches, the pairs of brain–testis and testis–ovary
330 showed strong connections, indicating a solid exchange module. Kidney and liver also
331 donate genes to one another, forming a separate module from brain–testis–ovary. D and
332 R branches showed a pronounced acceleration of PEO shifts between testis and ovary.
333 PEO shifts in S and D branches were moderately symmetric in the flow between pairs
334 of organs (Fig. 5C), meaning two organs tend to donate comparable numbers of
335 expressed genes each other. In contrast, R branches were more asymmetric. The lower
336 symmetry may have perturbed the evolutionary dynamics of gene expression.
337 To check the robustness of our analysis, we analyzed high-confidence subsets of
338 expression regime shifts. The most drastic expression changes were characterized by
339 introducing a cutoff of organ expression specificity ($\tau > 0.5$) to define organ-specific
340 genes. Although some previously significant trends were not recovered due to small
341 sample sizes, the result was largely consistent with the broader analysis (Fig. S11).
342 Because tree inference errors can bias the downstream analysis including OU modeling,
343 we also analyzed expression regime shifts found in clades which have a high support in
344 tree inference (>99% bootstrap support). Again, the results were largely consistent (Fig.
345 S11). Especially, the brain–testis–ovary and kidney–liver modules in S branches and the

346 testis–ovary connection in D and R branches were always reproduced in the analyses
347 with above thresholds in combinations with the two expression metrics (SVA-log-
348 TMM-FPKM and SVA-log-TPM; Fig. S11), demonstrating the robustness of our
349 conclusion.

350 The effect of gene trees was further examined by replicating the analysis with
351 alternative tree topologies inferred by species tree reconciliation, which takes into
352 account duplication–loss rates (Morel et al., 2019). This reconciliation step is expected
353 to correct erroneous tree topology, while possibly introducing another bias derived from
354 over-correction of biological signals such as incomplete lineage sorting. With the
355 reconciled trees, the OU modeling with SVA-log-TMM-FPKM values resulted in
356 equivalent numbers of expression shifts in S and D branches compared with those with
357 non-reconciled trees (97% [23,231/23,985] and 104% [9,407/9,018], respectively) (Fig.
358 S12A). In contrast, the phylogeny reconciliation substantially reduced the number of
359 shifts in R branches (39% [481/1,238]). This could be explained by the correction of
360 erroneous tree topology caused by the fast-evolving retro-copies (Fig. 4A), although the
361 differences in shift numbers did not correlate with the topological differences measured
362 by Robinson-Foulds distance (Robinson and Foulds, 1981) (Fig. S12B). Nevertheless,
363 resulting PEO shift distributions were largely similar (Fig. S12C), with the reproduced
364 accelerations in the brain–testis–ovary and kidney–liver modules (Fig. S12D) and the
365 asymmetric PEO shifts in R branches (Fig. S12E), suggesting the robustness of detected
366 modules against gene tree topology.

367 To obtain insight into the biological relevance of the among-organ modules, we
368 characterized human genes involved in PEO shifts from all branch categories using
369 the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis.
370 The genes descended from the testis–ovary PEO shifts enriched only one KEGG
371 pathway term "Cell cycle" (Table S5; adjusted P value < 0.05), likely reflecting their
372 function in meiosis. Although the adjusted P value was not statistically significant, it is
373 noteworthy that the top-ranked term for the brain–testis connection was "Endocrine and
374 other factor-regulated calcium reabsorption" annotated to four genes including GNAQ,
375 which has been implicated to tumor formation in neuronal tissues (Gessi et al., 2013;
376 Van Raamsdonk et al., 2009). In the kidney–liver module, 15 terms were significantly
377 enriched, many of which appear to be related to the functions and diseases in those
378 organs, for example, "Bile secretion", "Phagosome", "Lysosome", "ABC transporters",

379 “Sphingolipid metabolism”, and “Type I diabetes mellitus” (Table S5). These results
380 suggest that the among-organ modules in the PEO shifts played a role in supplying
381 functionally important genes.
382 Because ancestral expression was shown to orient new expression by the analysis
383 of PEO shifts, we concluded that there was prevalent organ-specific propensity, which
384 supports the presence of preadaptation in gene expression.
385
386 **Discussion:**
387 Our results suggest that the landscape of expression evolution is strongly shaped by
388 mechanisms of gene birth. Expression shifts are more pronounced following gene
389 duplication in agreement with the results of pairwise gene expression analyses (Assis
390 and Bachtrog, 2015; Guschanski et al., 2017; Huerta-Cepas et al., 2011), and shifts in
391 patterns of primary-expressed organs strongly depend on the expression state in the
392 ancestral organism. Thus, by analyzing such influences on a genome-wide scale for a
393 moderately large number of species, the question whether long-term expression in one
394 organ predisposes genes to be subsequently utilized in other organs has been answered
395 in the affirmative. There are preadaptive propensities in the evolution of vertebrate gene
396 expression, and the propensity varies with the presence and type of gene duplication.
397 Furthermore, the approach developed in this study, using complex gene family
398 phylogenies including gene duplications and losses that do not assume perfect match to
399 the species phylogeny, and incorporating a curation pipeline to amalgamate large
400 amounts of transcriptome data from many studies, was essential to obtain the necessary
401 species density and phylogenetic resolution to answer this question. The extensibility of
402 this method will allow for more species and more organs to be incorporated as further
403 studies come into the literature from diverse laboratories.
404 The mechanisms responsible for the preadaptive propensities that influence
405 expression shifts among organs are, however, unknown. A key question in
406 understanding these shifts may be the role of adaptation in the shift, and in subsequent
407 evolution. We have been careful so far to simply describe the shifts, but adaptive
408 possibilities include subfunctionalization, escape from adaptive conflict (EAC), and
409 neofunctionalization (Conant and Wolfe, 2008; Des Marais and Rausher, 2008; Sikosek
410 et al., 2012). The increased number of shifts following duplication suggests that drift
411 alone is not the explanation, but subfunctionalization easily could be.

412 Subfunctionalization is the idea that, if a gene has multiple functions prior to
413 duplication, they may be segregated among the duplicates following gene duplication.
414 Thus, the expression shifts may be simply a shift in focus of a duplicated copy on a
415 subset of the necessary expression profile needed at the organismal level. In this
416 scenario, any accompanying acceleration of amino acid substitution would be caused by
417 a loss of constraint and reduced purifying selection in one expression environment or
418 the other.

419 EAC involves more adaptation by adding the simple idea that prior to duplication,
420 the multiple functions and expression regimes were at least partially in conflict. Such
421 conflicts could clearly occur at the amino acid level, but could also occur at the
422 expression level. For example, if expression levels were focused on a most-important
423 tissue or most-sensitive tissue prior to duplication, but after duplication could be more
424 tailored to what is better for the new expression regime. Finally, neofunctionalization
425 would occur at the sequence or expression level, if the loss of selection on a duplicate
426 allowed mutations that were previously harmful to the old function, but now are not,
427 and are able to carry out some novel functional aspect that was previously prohibited.
428 Neofunctionalization is perhaps the most interesting and extraordinary possible cause
429 for the expression regime shifts we see, but it requires strong evidence and it is not a
430 necessary explanation for what we observe.

431 In this context, the patterns of expression regime shifts we observed may be
432 explained at different levels of biological organization, from the tissues and cells that
433 make up organs, to subcellular compartments, chromatin structure, promoter usage, and
434 protein biochemistry. Part of the propensity shifts we observed can be explained by the
435 "out of the testis" hypothesis, which posits that accelerated gains of testis expression are
436 based on the permissive chromatin state, abundant transcriptional machinery, relatively
437 simple promoters required for the expression in spermatogenic cells and following gains
438 of new expression patterns (Kaessmann, 2010; Kleene, 2005). This theory fits to the
439 accelerated testis-related PEO shifts, and could fit with any of the adaptive scenarios
440 discussed above, but the other detected patterns (e.g., kidney–liver module) require
441 other explanations.

442 One potential mechanism of preferences in expression regime shifts is a cell-type
443 or sub-cellular component mechanism. In such a mechanism, if two organs tend to share
444 cell types or usage of sub-cellular components, they may be prone to expropriate genes

445 between the two organs. It is known that gene expression levels in the kidney and liver
446 tend to change jointly, possibly reflecting their similar physiology including
447 detoxifications and waste excretion (Brawand et al., 2011). Such functional similarity
448 may also explain the presence of the kidney–liver module of gene exchange.

449 Another possibility is a regulatory mechanism whereby frequent gene-exchanging
450 organs use similar sets of regulatory elements. Altered expression between such organs
451 could occur with relatively few mutations in regulatory sequences. *Cis*-regulation is
452 indeed a major source of expression evolution, as it explains a certain fraction of
453 expression variability, for example, in budding yeast (30% in duplicates and 19% in
454 singletons) and undergoes a more rapid divergence than *trans*-regulation (Dong et al.,
455 2011).

456 Finally, another possible mechanism for expression regime shifts following gene
457 duplication is at the protein level. If frequently interacting organs have similar
458 environmental requirements for expressed proteins, a few amino acid substitutions may
459 tend to be required to adjust biochemical properties. Protein reusability may be
460 determined by cellular environments such as pH and temperature or by functional
461 categories of proteins. Our analysis indicates that the regime shifts that drastically
462 differentiate the expression tend to be coupled with accelerated protein evolution (Fig.
463 4B), and this result can be viewed as a support for a protein-level mechanism. In such a
464 mechanism, synergistic resolution of EAC may be a driving force for changes in both
465 amino acid composition and expression regimes. We note that these mechanistic
466 hypotheses are not mutually exclusive, and varying combinations of factors may
467 contribute to generate preadaptive patterns of gene expression.

468 In this study, we established a method to standardize RNA-seq data from disparate
469 research projects and developed a pathway for data amalgamation. Thanks to multiple
470 rounds of innovations in sequencing technology, transcriptome data are being produced
471 at an unprecedented rate in a greater variety of organisms and samples, such as those for
472 multi-species multi-organ developmental series (Cardoso-Moreira et al., 2019). The
473 transcriptome amalgamation will expand the use of such resources to study gene
474 expression evolution.

475 By reconstructing gene expression in gene family phylogenies, our analysis
476 revealed non-randomness and directionality of expression evolution. This suggests
477 prevalent preadaptation in gene expression, and that adaptation to expression in certain

organs is more conducive to future expression in other organs. This provides further details on how gene duplication has helped to reshape the dynamics of expression evolution that contributed to the vertebrate diversification.

482 **Materials and Methods:**

483

484 **Species selection.** A total of 105 species included in the Ensembl release 91 (Yates et
485 al., 2016) were searched for data availability in the NCBI SRA database (Leinonen et al.,
486 2011) (final search on May 1, 2018) and 22 species were found to have RNA-seq data
487 for 6 organs: brain, heart, kidney, liver, ovary, and testis. *Lepisosteus oculatus* was
488 excluded due to an insufficient quality of available expression data, and therefore
489 remaining 21 species were selected for further analysis.

490

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

495

496 **Gene sets.** Coding sequences (CDS) were retrieved from the Ensembl database. The
497 longest transcript was retained when multiple transcripts were annotated to the gene.
498 The quality of gene sets was evaluated using BUSCO 4.0.5 (Simão et al., 2015) with the
499 single-copy ortholog set "vertebrata_odb10" (Fig. S8).

500

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

510

511 **Transcriptome quantification.** Fastq files were extracted from downloaded SRA files
512 using parallel-fastq-dump 0.6.2 (https://github.com/rvalieris/parallel-fastq-dump) with
513 the minimum read length of 25 bp and the quality filter (-E option) (Leinonen et al.,
514 2011). The fastq sequences were then subjected to a quality filtering by fastp 0.12.3

515 (Chen et al., 2018). The filtered reads were mapped to genomic features annotated as
516 non-messenger RNAs in the Ensembl GTF files using bowtie2 2.3.4 (Langmead and
517 Salzberg, 2012) and resultant unmapped reads were used for expression level
518 quantification using kallisto 0.43.1 with the sequence-based bias correction (Bray et al.,
519 2016). Samples were removed if 20% or smaller percentages of reads were mappable
520 (Fig. S2). Estimated mapped read counts and transcript lengths were used to calculate
521 TPM and FPKM values. For the latter, the "trimmed mean of M-values (TMM)"
522 normalization method was applied (Robinson and Oshlack, 2010). Sample-wise TMM
523 scaling factors were obtained across all RNA-seq samples using the FPKM values of
524 1,377 single-copy orthologs. Because TMM and TPM transformations are incompatible
525 with each other, the TMM scaling was applied only to FPKM values. TPM and TMM-
526 FPKM values were subsequently transformed to $log(N + 1)$ values (log-TPM and log-
527 TMM-FPKM, respectively). Paralogous genes that haven't diverged in their nucleotide
528 sequences could not be distinguished well in the quantification step. Although our scope
529 is to characterize gene expression in the timescale of vertebrate evolution, this difficulty
530 likely leads to an underestimation of expression regime shifts in young duplicates.
531
532 **Iterative anomalous sample removal coupled with surrogate variable analysis.**
533 Anomalous RNA-seq samples were iteratively removed by a correlation analysis
534 coupled with an expression bias correction. In each iteration, unwanted biases in
535 expression level were removed by surrogate variable analysis (SVA, 'sva' function in
536 an R package 'sva') (Leek and Storey, 2007). SVA analysis was applied at the
537 beginning of each iteration so that it is not influenced by anomalous samples removed
538 in previous iterations. Subsequently, Pearson's correlation coefficients were calculated
539 for every RNA-seq data against mean expression level in each organ generated by
540 averaging all other data excluding those from the same BioProject (Fig. S3). We assume
541 that the sample's correlation coefficient against the same organ is higher than any of the
542 values against the other organs, and we removed all samples from the same BioProject
543 when violations were found. These steps were repeated until no violations were left and
544 SVA-corrected expression levels were finally reported (SVA-log-TMM-FPKM and
545 SVA-log-TPM). The curation steps were skipped if multiple samples were unavailable
546 in the species and hence SVA analysis was inapplicable. The final dataset was

547 comprised of 1,903 RNA-seq experiments from 182 BioProjects that cover 6 organs
548 from 21 vertebrate species without missing data (Tables S1–S2).
549
550 **Orthogroup classification.** Orthogroups, which contain all genes descended from one
551 gene in the common ancestor, were inferred from coding sequences of the 21 species
552 using OrthoFinder 2.1.2 (Emms and Kelly, 2015) guided by the species tree. In total,
553 17,896 orthogroups were generated. The largest orthogroup, which is comprised of
554 7,893 olfactory receptor genes, was removed from the analysis because of
555 computational burden. After sequence alignment processing (see "Multiple sequence
556 alignment"), we removed small orthogroups which retained less than four genes and
557 orthogroups that showed no parsimony informative sites, because phylogenetic
558 relationships cannot be inferred. As the result of filtering, 15,280 orthogroups were left
559 for OU modeling, with the largest one containing 3,796 zinc-finger proteins (Table S4).
560
561 **Multiple sequence alignment.** Multiple fasta files containing coding sequences were
562 generated for each orthogroup. Stops and ambiguous codons were masked as gaps (for
563 implementation, see https://github.com/kfuku52/cdskit). In-frame multiple codon
564 sequence alignments were generated using MAFFT 7.394 with the 'auto' option (Katoh
565 and Standley, 2013) and tranalign in EMBOSS 6.5.7.0 (Rice et al., 2000). Anomalous
566 genes were excluded by MaxAlign (Gouveia-Oliveira et al., 2007) which decreased the
567 largest orthogroup size from 3,796 to 2,382 genes. Spurious codons were removed in-
568 frame using pgtrimal in Phylogears2-2.0.2016.09.06
569 (https://www.fifthdimension.jp/products/phylogears/) with the 'gappyout' option
570 (Capella-Gutierrez et al., 2009).
571
572 **Gene tree reconstruction.** Maximum-likelihood trees were reconstructed using IQ-
573 TREE 1.6.5 (Nguyen et al., 2015) with the best-fit nucleotide substitution models
574 selected by ModelFinder with the Bayesian Information Criterion (Kalyaanamoorthy et
575 al., 2017). Larger orthogroups and longer genes tended to fit more complex substitution
576 matrices and larger numbers of categories of rate heterogeneity (Fig. S13A–B; Table
577 S4). Ultrafast bootstrapping with 1,000 replicates was performed to evaluate the
578 credibility of tree topology (Minh et al., 2013) with a further optimization of each
579 bootstrapping tree (-bnni option) (Hoang et al., 2018). To evaluate the effect of

580 alternative gene tree topologies, we performed phylogeny reconciliation using GeneRax
581 1.0.0 (Morel et al., 2019) with the maximum-likelihood gene trees and the species tree
582 as input. Because rooted trees were generated in this step, the tree rooting (described
583 below) was skipped for the reconciled trees.
584
585 **Reconciliation-assisted gene tree rooting.** Candidate rooting positions were inferred
586 with different methods. Using the dated species tree, all rooting branches with the
587 minimum duplication-loss score were identified using the rooting mode of NOTUNG
588 2.9 with the default parameters (duplication score = 1.5, loss score = 1.0) (Chen et al.,
589 2000). The midpoint of the longest path (Farris, 1972) and the position with the minimal
590 ancestor deviations (MAD) (Tria et al., 2017) were also considered as candidates. The
591 final rooting position was reported based on overlaps among those rooting positions
592 (Figs. S13C,E; Table S4).
593
594 **Reconciliation-assisted divergence time estimation.** To prepare dated gene trees, we
595 first matched species tree nodes with corresponding gene tree nodes using the
596 reconciliation mode of NOTUNG 2.9 (Chen et al., 2000) and created time constraints of
597 speciation nodes (Fig. S13F). Duplication nodes were constrained with the upper and
598 lower age limits derived from corresponding speciation nodes. If the root node is a
599 duplication node and is not covered by the range of the species tree, the upper age limit
600 was set to 1,105 million years ago, which corresponds to the split of animals and fungi
601 (Hedges et al., 2006). Divergence time was then estimated by a penalized likelihood
602 method (Sanderson, 2002) implemented in an R package 'ape' ('chronos' function with
603 'discrete' model) (Popescu et al., 2012) with time constraints on speciation, duplication,
604 and root nodes. When reasonable initial parameters were not found after 1,000 trials, the
605 above constraints were partly relaxed (Fig. S13D; Table S4). The implementation is
606 provided on GitHub (https://github.com/kfuku52/RADTE).
607
608 **Modeling and shift detection of expression evolution.** Using the dated gene trees and
609 organ-wise mean values of SVA-log-TMM-FPKM and SVA-log-TPM, regime shifts in
610 gene expression were detected as shifts of optimal trait values in Ornstein–Uhlenbeck
611 (OU) models determined by a Lasso-based model selection with AICc in an R package

612 '*l*1ou' (Khabbazian et al., 2016). Because there is no available software to handle
613 within-species variation in phylogenetic OU shift detection without predefined
614 hypotheses on the number and place of regime shifts, we used mean expression level as
615 the input. It is shown by simulation that the species mean and species variance models
616 show comparable power in the regime shift detection (Rohlfs et al., 2014), suggesting
617 that our species mean model is expected to perform as good as the species variance
618 model. In the model, α and σ^2 parameters were assumed unchanged in the tree
619 (Khabbazian et al., 2016), and therefore only the global, rather than branch- or clade-
620 wise, stationary variance (γ) were obtained. Expression levels in the six organs were
621 treated as multivariate traits where α and σ^2 were estimated for each organ but regime
622 shifts were assumed to occur jointly in the same set of branches (Khabbazian et al.,
623 2016). The phylogenetic mean (expression level at the root node) was estimated with
624 the "OUfixedRoot" model. To handle gene trees recalcitrant to this analysis (especially
625 those with a large number of genes), we skimmed gene trees by collapsing clades with
626 small changes in expression level (Fig. S14). Specifically, we first calculated all-vs-all
627 Pearson's correlation coefficients of gene expression level among all genes that belong
628 the clade. The clade was collapsed into a single tip if the expression patterns were
629 almost identical (minimum correlation coefficient between genes > 0.99). Phylogenetic
630 means of the collapsed clade were calculated by assuming the Brownian motion and
631 were used as expression level at the new tip. The upper limit of regime shifts was set as
632 $\max\left[\min(N/2, 100), \sqrt[2]{N}\right]$, except for the largest tree with 2,382 genes where the upper
633 limit was decreased to 10 to cope with an unrealistically large number of branch
634 combinations to consider. The number of detected regime shifts was always smaller
635 than the upper limit in the SVA-log-TMM-FPKM analysis, whereas three out of 15,820
636 trees reached the upper limit in the SVA-log-TPM analysis (Fig. S6A).

637

638 **Analysis of expression patterns.** We characterized expression patterns of extant and
639 ancestral genes by calculating different metrics from fitted values (μ) in the OU models.
640 Organ specificity was measured by τ (Yanai et al., 2005), which outperformed other
641 methods in a benchmark for tissue specificity (Kryuchkova-Mostacci and Robinson-
642 Rechavi, 2016). Expression complementarity between sister lineages was measured by

643 the metrics called 'TEC' (Huerta-Cepas et al., 2011). Because μ was estimated from
644 log-transformed expression levels, these expression metrics were calculated with unlog-
645 transformed μ values. PEOs were defined as the organ in which the highest expression
646 levels were observed among the six organs we analyzed.
647
648 **Estimation of protein evolution rate.** Parameters for codon substitution matrix, shape
649 parameter of discrete gamma distribution for rate heterogeneity (α), equilibrium
650 transition/transversion rate ratio (κ), equilibrium nonsynonymous/synonymous
651 substitution rate ratio (ω) were estimated using IQ-TREE 1.6.5 (Nguyen et al., 2015) by
652 fitting GY+F3X4+G4 models to each gene tree. Equilibrium base composition (θ) was
653 estimated from empirical codon state frequencies which are calculated from the
654 alignment by counting. To obtain θ at the root node, we calculated θ at subroot nodes by
655 taking advantage of IQ-TREE's empirical Bayesian method for ancestral sequence
656 reconstruction. Considering subroot branch length, a weighted average of the subroot θ
657 values were calculated as the θ value at the root node. Using all those parameters,
658 branch-wise nonsynonymous/synonymous substitution rate ratios were estimated by
659 stochastic substitution mapping (*mapdNdS*) (Guéguen and Duret, 2018) using bio++
660 library (Guéguen et al., 2013). To examine the robustness of the *mapdNdS*-based ω
661 estimation, we compared the results with those obtained by maximum-likelihood ω
662 estimation by fitting MG94W9 models in HyPhy 2.3.11 (Pond et al., 2005). The two
663 methods yielded consistent results on the effect of branching events and expression
664 shifts (Fig. 4A; Fig. S9A), suggesting methodological robustness. We reported
665 *mapdNdS*-based results in the main text.
666
667 **Analysis of gene structure and location.** The number of introns and chromosomal
668 location were obtained for each gene from the Ensembl gene models (GFF3 files). The
669 intron numbers were subsequently converted to binary values that represent intronless
670 and intron-containing states. Chromosomal locations were categorized into autosome, X
671 chromosome, and Y chromosome. Genes from non-therian species were treated as
672 missing data because mechanisms of their sex determination are not homologous to the
673 mammalian XY system (Veyrunes et al., 2008). Genes from *Chinchilla lanigera* were
674 also treated as missing data because their sequenced genomes are not anchored to
675 chromosomes in the Ensembl release 91. The posterior probabilities of ancestral

676 character states were inferred by the stochastic character mapping of discrete traits
677 (Bollback, 2006) implemented in an R package 'phytools' (Revell, 2012). Because
678 functional retrotranspositions (Jun et al., 2009; Marques et al., 2005) and inter-
679 chromosomal duplications (Pace et al., 2009) are rare events relative to the timescale of
680 the vertebrate evolution on the per-gene basis, we set the transition rate parameters to a
681 sufficiently small value (1×10^{-3} per gene per million years). Since intron gain occurs
682 few orders of magnitude less frequently than its loss caused by retrotransposition and
683 others (Roy and Gilbert, 2005), the rate of intron gain is set to be lower (1×10^{-4} per
684 gene per million years).
685
686 **Analysis of branching events.** Speciation and duplication nodes (S and D/R nodes,
687 respectively) were classified by a species-overlap method (Huerta-Cepas et al., 2007)
688 and were mapped to the species tree on the basis of species coverages of the gene tree
689 clades. A transition from intron-containing to intronless states with a posterior
690 probability greater than 0.5 was classified as a retrotransposition event (R node). The
691 branches that correspond to the original copy of a retrotransposition event were not
692 included in R branches. Although our classification cannot detect retrotranspositions
693 from originally intronless genes, we expect such situations would be rare because most
694 vertebrate genes contain at least one intron (e.g., 20,160/21,242 human genes). Inter-
695 chromosomal translocation was detected by considering chromosomal locations with
696 the highest posterior probability as the ancestral states. Because of the difficulty in
697 determining rooting positions of deep phylogenies, gene tree nodes older than the root
698 node of the species tree were removed from the analysis.
699
700 **KEGG pathway enrichment analysis.** Human genes that descend from the shift
701 branches were pooled for each specific PEO shift. The gene lists were examined for
702 enrichment against an Enrichr library "KEGG_2019_Human" (Kuleshov et al., 2016)
703 using a python package GSEApy (https://github.com/zqfang/GSEApy). Statistically
704 significant (adjusted P value < 0.05) KEGG pathway terms were reported in Table S5.
705
706 **Data visualization.** Phylogenetic trees were visualized using a python package 'ETE 3'
707 (Huerta-Cepas et al., 2016) and an R package 'ggtree' (Yu et al., 2017). A part of
708 animal silhouettes in Fig. 3A were obtained from PhyloPic (http://phylopic.org).

Box plot elements of all figures are defined as follows: center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers. Box-plot outliers are suppressed in Figs. 3, 4, S9, S11, and S12.

Code and data availability. Scripts, parameter values, gene expression values including SVA-log-TMM-FPKM and SVA-log-TPM, and other data used in this study are available as Supplementary Data (Mendeley Data ID available on publication).

Acknowledgements:

We acknowledge the following sources for funding: MEXT/JSPS KAKENHI 18J00178 (K.F.), Sofja Kovalevskaja programme by the Alexander von Humboldt Foundation (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.

729 **References:**

730 Altenhoff, A.M., Studer, R.A., Robinson-Rechavi, M., Dessimoz, C., and Couto, F.
731 (2012). Resolving the ortholog conjecture: orthologs tend to be weakly, but
732 significantly, more similar in function than paralogs. PLoS Comput. Biol. *8*, e1002514.

733 Assis, R., and Bachtrog, D. (2015). Rapid divergence and diversification of mammalian
734 duplicate gene functions. BMC Evol. Biol. *15*, 138.

735 Babushok, D. V., Ostertag, E.M., and Kazazian, H.H. (2007). Current topics in genome
736 evolution: Molecular mechanisms of new gene formation. Cell. Mol. Life Sci. *64*, 542–
737 554.

738 Balakirev, E.S., and Ayala, F.J. (2003). Pseudogenes: Are they "junk" or functional
739 DNA? Annu. Rev. Genet. *37*, 123–151.

740 Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J.,
741 Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., et al. (2012). The evolutionary
742 landscape of alternative splicing in vertebrate species. Science *338*, 1587–1593.

743 Bedford, T., and Hartl, D.L. (2009). Optimization of gene expression by natural
744 selection. Proc. Natl. Acad. Sci. U. S. A. *106*, 1133–1138.

745 Boer, P.H., Adra, C.N., Lau, Y.F., and McBurney, M.W. (1987). The testis-specific
746 phosphoglycerate kinase gene pgk-2 is a recruited retroposon. Mol. Cell. Biol. *7*, 3107–
747 3112.

748 Bollback, J. (2006). SIMMAP: Stochastic character mapping of discrete traits on
749 phylogenies. BMC Bioinformatics *7*, 88.

750 Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier,
751 M., Liechti, A., Aximu-Petri, A., Kircher, M., et al. (2011). The evolution of gene
752 expression levels in mammalian organs. Nature *478*, 343–348.

753 Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic
754 RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

755 Breschi, A., Djebali, S., Gillis, J., Pervouchine, D.D., Dobin, A., Davis, C.A., Gingeras,
756 T.R., and Guigó, R. (2016). Gene-specific patterns of expression variation across organs
757 and species. Genome Biol. *17*, 151.

758 Brunner, E., and Munzel, U. (2000). The nonparametric Behrens-Fisher problem:
759 asymptotic theory and a small-sample approximation. Biom. J. *42*, 17–25.

760 Budd, G.E. (2007). On the origin and evolution of major morphological characters. Biol.
761 Rev. *81*, 609–628.

762 Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009). trimAl: a tool for
763 automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*,
764 1972–1973.

765 Cardoso-Moreira, M., Halbert, J., Valloton, D., Velten, B., Chen, C., Shao, Y., Liechti,
766 A., Ascenção, K., Rummel, C., Ovchinnikova, S., et al. (2019). Gene expression across
767 mammalian organ development. Nature *571*, 505–509.

768 Carelli, F.N., Hayakawa, T., Go, Y., Imai, H., Warnefors, M., and Kaessmann, H.
769 (2016). The life history of retrocopies illuminates the evolution of new mammalian
770 genes. Genome Res. *26*, 301–314.

771 Castillo-Davis, C.I., Hartl, D.L., and Achaz, G. (2004). cis-Regulatory and protein
772 evolution in orthologous and duplicate genes. Genome Res. *14*, 1530–1536.

773 Chen, X., and Zhang, J. (2012). The ortholog conjecture is untestable by the current
774 gene ontology but is supported by RNA sequencing data. PLoS Comput. Biol. *8*,
775 e1002784.

776 Chen, J., Swofford, R., Johnson, J., Cummings, B.B., Rogel, N., Lindblad-Toh, K.,
777 Haerty, W., Palma, F. di, and Regev, A. (2019). A quantitative framework for
778 characterizing the evolutionary history of mammalian gene expression. Genome Res. *29*,
779 53–63.

780 Chen, K., Durand, D., and Farach-Colton, M. (2000). NOTUNG: a program for dating
781 gene duplications and optimizing gene family trees. J. Comput. Biol. *7*, 429–447.

782 Chen, S., Krinsky, B.H., and Long, M. (2013). New genes as drivers of phenotypic
783 evolution. Nat. Rev. Genet. *14*, 645–660.

784 Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ
785 preprocessor. Bioinformatics *34*, i884–i890.

786 Conant, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job: How duplicated
787 genes find new functions. Nat. Rev. Genet. *9*, 938–950.

788 Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P.D.,
789 Grützner, F., and Kaessmann, H. (2014). Origins and functional evolution of Y
790 chromosomes across mammals. Nature *508*, 488–493.

791 Danshina, P. V., Geyer, C.B., Dai, Q., Goulding, E.H., Willis, W.D., Kitto, G.B.,
792 McCarrey, J.R., Eddy, E.M., and O'Brien, D.A. (2010). Phosphoglycerate kinase 2
793 (PGK2) is essential for sperm function and male fertility in mice. Biol. Reprod. *82*,
794 136–145.

795 de-Leon, S.B.-T., and Davidson, E.H. (2007). Gene regulation: gene control network in
796 development. Annu. Rev. Biophys. Biomol. Struct. *36*, 191–212.

797 Des Marais, D.L., and Rausher, M.D. (2008). Escape from adaptive conflict after
798 duplication in an anthocyanin pathway gene. Nature *454*, 762–765.

799 Dong, D., Yuan, Z., and Zhang, Z. (2011). Evidences for increased expression variation
800 of duplicate genes in budding yeast: from *cis*- to *trans*- regulation effects. Nucleic Acids
801 Res. *39*, 837–847.

802 Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole
803 genome comparisons dramatically improves orthogroup inference accuracy. Genome
804 Biol. *16*, 157.

805 Farris, J.S. (1972). Estimating phylogenetic trees from distance matrices. Am. Nat. *106*,
806 645–668.

807 Franzén, O., Gan, L.-M., and Björkegren, J.L.M. (2019). PanglaoDB: a web server for
808 exploration of mouse and human single-cell RNA sequencing data. Database *2019*.

809 Gessi, M., Hammes, J., Lauriola, L., Dörner, E., Kirfel, J., Kristiansen, G., Muehlen, A.
810 zur, Denkhaus, D., Waha, A., and Pietsch, T. (2013). GNA11 and N-RAS mutations:
811 alternatives for MAPK pathway activating GNAQ mutations in primary melanocytic
812 tumours of the central nervous system. Neuropathol. Appl. Neurobiol. *39*, 417–425.

813 Gould, S.J., and Vrba, E.S. (1982). Exaptation—a missing term in the science of form.
814 Paleobiology *8*, 4–15.

815 Gouveia-Oliveira, R., Sackett, P.W., and Pedersen, A.G. (2007). MaxAlign:
816 maximizing usable data in an alignment. BMC Bioinformatics *8*, 312.

817 Guéguen, L., and Duret, L. (2018). Unbiased estimate of synonymous and
818 nonsynonymous substitution rates with nonstationary base composition. Mol. Biol. Evol.
819 *35*, 734–742.

820 Guéguen, L., Gaillard, S., Boussau, B., Gouy, M., Groussin, M., Rochette, N.C., Bigot,
821 T., Fournier, D., Pouyet, F., Cahais, V., et al. (2013). Bio++: efficient extensible
822 libraries and tools for computational molecular evolution. Mol. Biol. Evol. *30*, 1745–
823 1750.

824 Guschanski, K., Warnefors, M., and Kaessmann, H. (2017). The evolution of duplicate
825 gene expression in mammalian organs. Genome Res. *27*, 1461–1474.

826 Hansen, T.F. (1997). Stabilizing selection and the comparative analysis of adaptation.
827 Evolution *51*, 1341–1351.

828 Hedges, S.B., Dudley, J., and Kumar, S. (2006). TimeTree: a public knowledge-base of
829 divergence times among organisms. Bioinformatics *22*, 2971–2972.

830 Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018).
831 UFBoot2: Improving the ultrafast bootstrap approximation. Mol. Biol. Evol. *35*, 518–
832 522.

833 Huerta-Cepas, J., Dopazo, H., Dopazo, J., Gabaldón, T., McPherson, J., Marra, M.,
834 Hillier, L., Waterston, R., Chinwalla, A., Wallis, J., et al. (2007). The human phylome.
835 Genome Biol. *8*, 934–941.

836 Huerta-Cepas, J., Dopazo, J., Huynen, M.A., and Gabaldón, T. (2011). Evidence for
837 short-time divergence and long-time conservation of tissue-specific expression after
838 gene duplication. Brief. Bioinform. *12*, 442–448.

839 Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, analysis, and
840 visualization of phylogenomic data. Mol. Biol. Evol. *33*, 1635–1638.

841 Julien, P., Brawand, D., Soumillon, M., Necsulea, A., Liechti, A., Schütz, F., Daish, T.,
842 Grützner, F., and Kaessmann, H. (2012). Mechanisms and evolutionary patterns of
843 mammalian and avian dosage compensation. PLoS Biol. *10*, e1001328.

844 Jun, J., Ryvkin, P., Hemphill, E., Mandoiu, I., and Nelson, C. (2009). The birth of new
845 genes by RNA- and DNA-mediated duplication during mammalian evolution. J.
846 Comput. Biol. *16*, 1429–1444.

847 Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes.
848 Genome Res. *20*, 1313–1326.

849 Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S.
850 (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat.
851 Methods *14*, 587–589.

852 Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software
853 version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

854 Khabbazian, M., Kriebel, R., Rohe, K., and Ané, C. (2016). Fast and accurate detection
855 of evolutionary shifts in Ornstein-Uhlenbeck models. Methods Ecol. Evol. *7*, 811–824.

856 Kleene, K.C. (2005). Sexual selection, genetic conflict, selfish genes, and the atypical
857 patterns of gene expression in spermatogenic cells. Dev. Biol. *277*, 16–26.

858 Kryuchkova-Mostacci, N., and Robinson-Rechavi, M. (2016). A benchmark of gene
859 expression tissue-specificity metrics. Brief. Bioinform. *44*, bbw008.

860 Kryuchkova-Mostacci, N., Robinson-Rechavi, M., Shmoish, M., Chalifa-Caspi, V.,
861 Shklar, M., and Ophir, R. (2016). Tissue-specificity of gene expression diverges slowly
862 between orthologs, and rapidly between paralogs. PLOS Comput. Biol. *12*, e1005274.

863 Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z.,
864 Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a
865 comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids
866 Res. *44*, W90–W97.

867 Lan, X., and Pritchard, J.K. (2016). Coregulation of tandem duplicate genes slows
868 evolution of subfunctionalization in mammals. Science *352*, 1009–1013.

869 Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2.
870 Nat. Methods *9*, 357–359.

871 Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies
872 by surrogate variable analysis. PLoS Genet. *3*, e161.

873 Leinonen, R., Sugawara, H., and Shumway, M. (2011). The Sequence Read Archive.
874 Nucleic Acids Res. *39*, D19–D21.

875 Liao, B.-Y., and Zhang, J. (2006). Low rates of expression profile divergence in highly
876 expressed genes and tissue-specific genes during mammalian evolution. Mol. Biol. Evol.
877 *23*, 1119–1128.

878 Liu, X.-X., Zhang, H., Shen, X.-F., Liu, F.-J., Liu, J., and Wang, W.-J. (2015).
879 Characteristics of testis-specific phosphoglycerate kinase 2 and its association with
880 human sperm quality. Hum. Reprod. *126*, dev301.

881 Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A., and Kaessmann, H.
882 (2005). Emergence of young human genes after a burst of retroposition in primates.
883 PLoS Biol. *3*, e357.

884 McCarrey, J.R., and Thomas, K. (1987). Human testis-specific PGK gene lacks introns
885 and possesses characteristics of a processed gene. Nature *326*, 501–505.

886 Mighell, A.J., Smith, N.R., Robinson, P.A., and Markham, A.F. (2000). Vertebrate
887 pseudogenes. FEBS Lett. *468*, 109–114.

888 Minh, B.Q., Nguyen, M.A.T., and von Haeseler, A. (2013). Ultrafast approximation for
889 phylogenetic bootstrap. Mol. Biol. Evol. *30*, 1188–1195.

890 Morel, B., Kozlov, A.M., Stamatakis, A., and Szöllősi, G.J. (2019). GeneRax: A tool
891 for species tree-aware maximum likelihood based gene tree inference under gene
892 duplication, transfer, and loss. BioRxiv 779066.

893 Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker,
894 J.C., Grützner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and
895 expression patterns in tetrapods. Nature *505*, 635–640.

896 Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: A
897 fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies.
898 Mol. Biol. Evol. *32*, 268–274.

899 Pace, J.K., Sen, S.K., Batzer, M.A., and Feschotte, C. (2009). Repair-mediated
900 duplication by capture of proximal chromosomal DNA has shaped vertebrate genome
901 evolution. PLoS Genet. *5*, e1000469.

902 Podgornaia, A.I., and Laub, M.T. (2015). Pervasive degeneracy and epistasis in a
903 protein-protein interface. Science *347*, 673–677.

904 Pond, S.L.K., Frost, S.D.W., and Muse, S. V. (2005). HyPhy: hypothesis testing using
905 phylogenies. Bioinformatics *21*, 676–679.

906 Popescu, A.-A., Huber, K.T., and Paradis, E. (2012). ape 3.0: New tools for distance-
907 based phylogenetics and evolutionary analysis in R. Bioinformatics *28*, 1536–1537.

908 Potrzebowski, L., Vinckenbosch, N., Marques, A.C., Chalmel, F., Jégou, B., and
909 Kaessmann, H. (2008). Chromosomal gene movements reflect the recent origin and
910 biology of therian sex chromosomes. PLoS Biol. *6*, e80.

911 Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and
912 other things). Methods Ecol. Evol. *3*, 217–223.

913 Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular
914 Biology Open Software Suite. Trends Genet. *16*, 276–277.

915 Robinson, D.F., and Foulds, L.R. (1981). Comparison of phylogenetic trees. Math.
916 Biosci. *53*, 131–147.

917 Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for
918 differential expression analysis of RNA-seq data. Genome Biol *11*, 2010–2011.

919 Rogozin, I.B., Managadze, D., Shabalina, S.A., and Koonin, E. V (2014). Gene family
920 level comparative analysis of gene expression in mammals validates the ortholog
921 conjecture. Genome Biol. Evol. *6*, 754–762.

922 Rohlfs, R. V., Harrigan, P., and Nielsen, R. (2014). Modeling Gene Expression
923 Evolution with an Extended Ornstein–Uhlenbeck Process Accounting for Within-
924 Species Variation. Mol. Biol. Evol. *31*, 201–211.

925 Roy, S.W., and Gilbert, W. (2005). Rates of intron loss and gain: implications for early
926 eukaryotic evolution. Proc. Natl. Acad. Sci. U. S. A. *102*, 5773–5778.

927 Sanderson, M.J. (2002). Estimating absolute rates of molecular evolution and
928 divergence times: a penalized likelihood approach. Mol. Biol. Evol. *19*, 101–109.

929 She, X., Rohl, C.A., Castle, J.C., Kulkarni, A. V, Johnson, J.M., and Chen, R. (2009).
930 Definition, conservation and epigenetics of housekeeping and tissue-enriched genes.
931 BMC Genomics *10*, 269.

932 Sikosek, T., Chan, H.S., and Bornberg-Bauer, E. (2012). Escape from Adaptive Conflict
933 follows from weak functional trade-offs and mutational robustness. Proc. Natl. Acad.
934 Sci. *109*, 14888–14893.

935 Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M.
936 (2015). BUSCO: assessing genome assembly and annotation completeness with single-
937 copy orthologs. Bioinformatics *31*, 3210–3212.

938 Soubrier, J., Steel, M., Lee, M.S.Y., Der Sarkissian, C., Guindon, S., Ho, S.Y.W., and
939 Cooper, A. (2012). The influence of rate heterogeneity among sites on the time
940 dependence of molecular rates. Mol. Biol. Evol. *29*, 3345–3358.

941 Starr, T.N., Picton, L.K., and Thornton, J.W. (2017). Alternative evolutionary histories
942 in the sequence space of an ancient protein. Nature *549*, 409–413.

943 Tria, F.D.K., Landan, G., and Dagan, T. (2017). Phylogenetic rooting using minimal
944 ancestor deviation. Nat. Ecol. Evol. *1*, 0193.

945 Van Raamsdonk, C.D., Bezrookove, V., Green, G., Bauer, J., Gaugler, L., O'Brien,
946 J.M., Simpson, E.M., Barsh, G.S., and Bastian, B.C. (2009). Frequent somatic
947 mutations of GNAQ in uveal melanoma and blue naevi. Nature *457*, 599–602.

948 Veyrunes, F., Waters, P.D., Miethke, P., Rens, W., McMillan, D., Alsop, A.E., Grutzner,
949 F., Deakin, J.E., Whittington, C.M., Schatzkamer, K., et al. (2008). Bird-like sex
950 chromosomes of platypus imply recent origin of mammal sex chromosomes. Genome
951 Res. *18*, 965–973.

952 Warnefors, M., and Kaessmann, H. (2013). Evolution of the Correlation between
953 Expression Divergence and Protein Divergence in Mammals. Genome Biol. Evol. *5*,
954 1324–1335.

955 Wilson, B.A., Foy, S.G., Neme, R., and Masel, J. (2017). Young genes are highly
956 disordered as predicted by the preadaptation hypothesis of de novo gene birth. Nat. Ecol.
957 Evol. *1*, 0146.

958 Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-
959 Even, A., Horn-Saban, S., Safran, M., Domany, E., et al. (2005). Genome-wide
960 midrange transcription profiles reveal expression level relationships in human tissue
961 specification. Bioinformatics *21*, 650–659.

962 Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences
963 with variable rates over sites: Approximate methods. J. Mol. Evol. *39*, 306–314.

964 Yang, Z. (1995). A space-time process model for the evolution of DNA sequences.
965 Genetics *139*, 993–1005.

966 Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D.,
967 Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2016). Ensembl 2016. Nucleic
968 Acids Res. *44*, D710–D716.

969 Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.-Y. (2017). ggtree : an R package
970 for visualization and annotation of phylogenetic trees with their covariates and other
971 associated data. Methods Ecol. Evol. *8*, 28–36.

972 Yu, Z., Morais, D., Ivanga, M., and Harrison, P.M. (2007). Analysis of the role of
973 retrotransposition in gene evolution in vertebrates. BMC Bioinformatics *8*, 308.

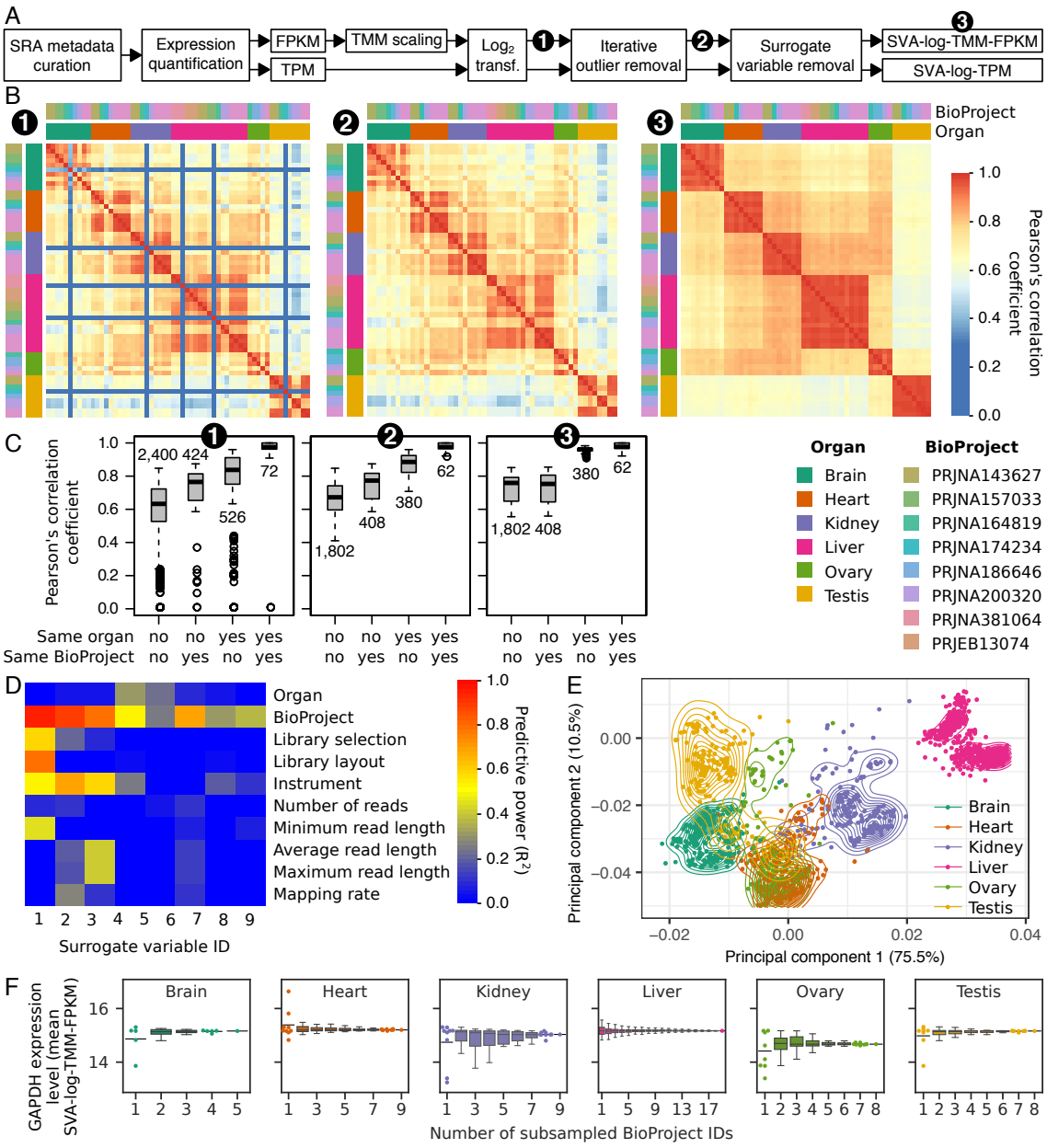
974 Zhang, L., and Li, W.-H. (2004). Mammalian housekeeping genes evolve more slowly
975 than tissue-specific genes. Mol. Biol. Evol. *21*, 236–239.

976 Zhang, Y.E., and Long, M. (2014). New genes contribute to genetic and phenotypic
977 novelties in human evolution. Curr. Opin. Genet. Dev. *29*, 90–96.

978 Zhang, Y.E., Landback, P., Vibranovski, M.D., Long, M., and Stark, A. (2011).
979 Accelerated recruitment of new brain development genes into the human genome. PLoS
980 Biol. *9*, e1001179.

981

982



983

984 **Figure 1. Transcriptome amalgamation to integrate heterogeneous RNA-seq**
985 **samples.** (**A**) A simplified flow chart of the transcriptome amalgamation. The full chart
986 is available in Fig. S1. (**B–D**) Transcriptome curation within species. Data from
987 *Monodelphis domestica* with SVA-log-TMM-FPKM metrics are shown as an example.
988 The heatmaps show Pearson's correlation coefficients among RNA-seq samples (**B**).
989 Each row and column correspond to one RNA-seq sample. The expression levels of all
990 genes were used to calculate the correlation coefficients. Note that anomalous samples
991 contaminated in the curated metadata (low correlation samples in 1) are successfully
992 removed, and that project-specific correlations visible in the uncorrected data (marked

993 2) are absent in the corrected data (marked 3). The box plots show distinct distributions
994 of the correlation coefficients depending on whether a pair of samples are the same
995 organ or whether they are from the same research project (**C**). The numbers of
996 comparisons are provided in the plot. The correlation coefficients are largely improved
997 in within-organ comparisons when surrogate variables are removed, while within-
998 project biases are attenuated. In this species, nine surrogate variables were detected
999 against 52 RNA-seq data from eight projects (**D**). Analysis of those variables by linear
1000 regression highlights the BioProject feature as the strongest source of removed biases.
1001 For full description of predictors, see Fig. S4. (**E**) A principal component analysis using
1002 expression levels of 1,377 single-copy orthologs from 21 species. Points correspond to
1003 RNA-seq samples. Curves show the estimated kernel density. Explained variations in
1004 percentages are indicated in each axis. (**F**) Estimated organ-wise expression levels of a
1005 housekeeping gene. Since data from relatively many BioProjects are available,
1006 glyceraldehyde-3-phosphate dehydrogenase gene (GAPDH, Ensembl gene ID:
1007 ENSGALG00000014442) in *Gallus gallus* is shown as an example. Points corresponds
1008 to the average expression level calculated by random subsampling. All data points and
1009 the median value (bar), rather than a box plot, are shown if the number of subsampled
1010 BioProject combinations is less than 10.

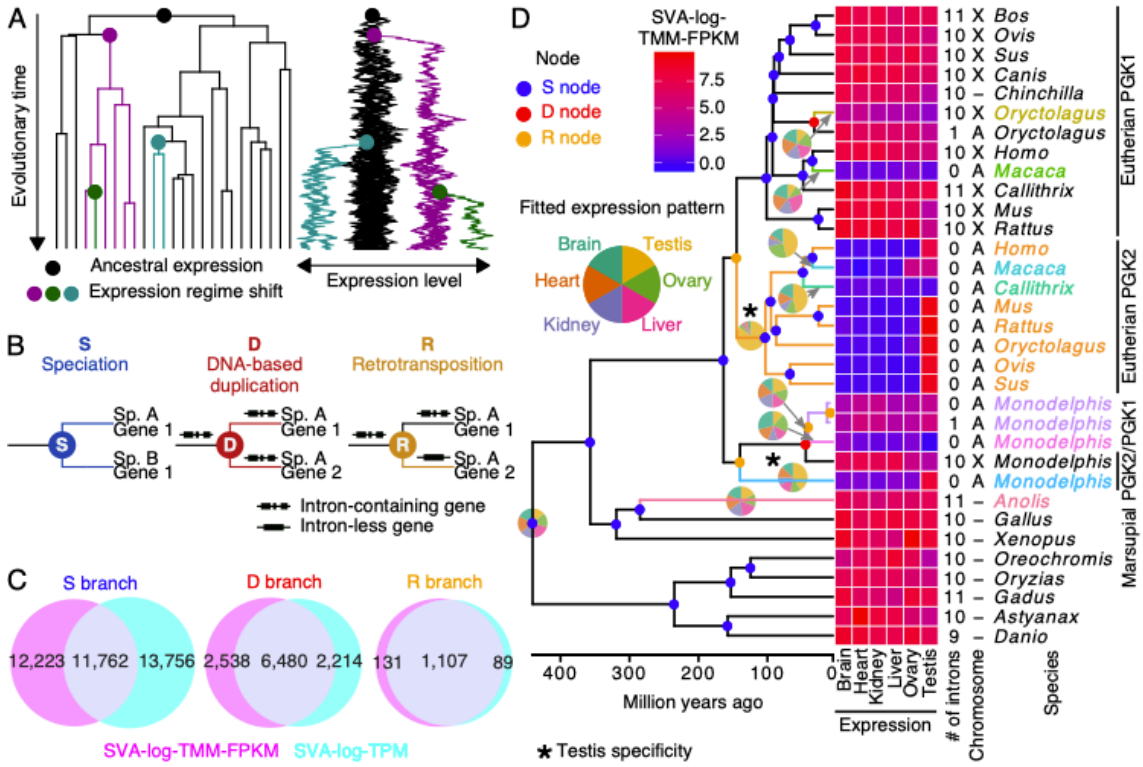


Figure 2. Expression evolution in a complex history of gene family evolution. (**A**) Modeling expression evolution with multi-optima Ornstein–Uhlenbeck process. A phylogenetic simulation is shown. Colors show branches belonging to different regimes. Regime shifts (change of color) appear as a substantial change in optimal trait values. (**B**) Nodes and branches of gene family phylogeny were categorized into S, D, or R based on the branching events, i.e., speciation, DNA-based duplication, or retrotransposition, respectively. (**C**) Venn diagrams of expression regime shifts. Circles represent the sets of branches where regime shifts were detected with SVA-log-TMM-FPKM or SVA-log-TPM. (**D**) 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 and lineage-specific gene duplication. Tips correspond to genes. The colors of branches and tip labels indicate expression regimes. Node colors match to the categorization in **B**. The heatmap shows expression levels and among-organ expression patterns are shown as a pie chart for each regime. To the right, the number of introns and located chromosomes (A, autosome; X, X chromosome; Y, Y chromosome) are indicated. For full information including complete tip labels and bootstrap branch supports, see Supplementary Data.

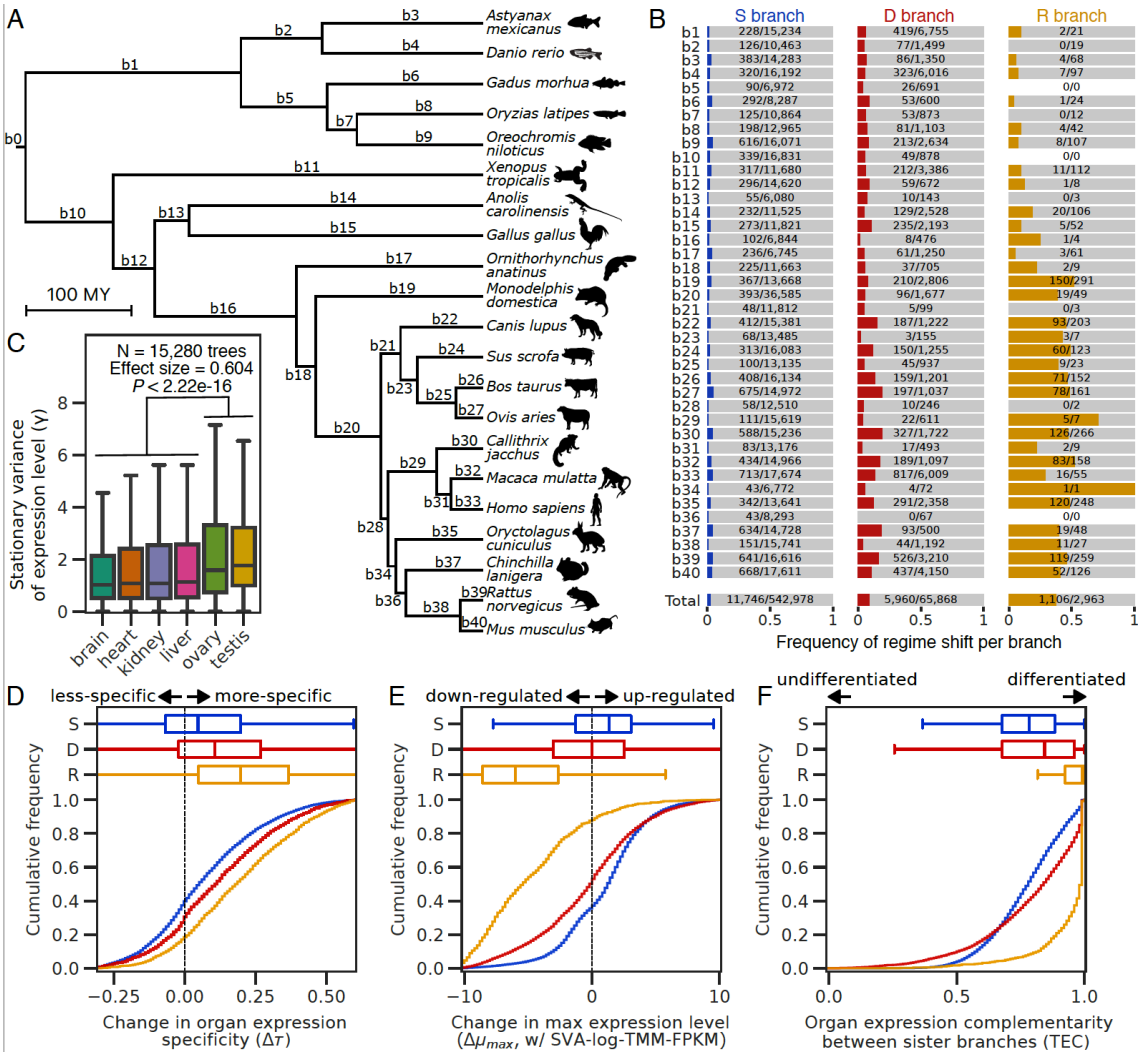


Figure 3. Characteristics of expression shifts in 15,280 gene trees. (**A**) The species tree showing analyzed genomes and their divergence time. (**B**) Mapping of 18,812 expression shifts in the species tree. The number and proportion of expression regime shifts in S, D, and R branches are shown. Corresponding branches in the species tree are indicated in **A**. (**C**) Organ-specific stationary variances (γ) of expression level evolution in vertebrates. The distribution of γ between reproductive and non-reproductive organs were compared by a Brunner-Munzel test (Brunner and Munzel, 2000). (**D–F**) Cumulative frequency of change in organ expression specificity (**D**), change in maximum expression level (**E**), and expression complementarity between sister lineages (**F**) among detected expression shifts. Number of analyzed regime shifts are shown in **B**.

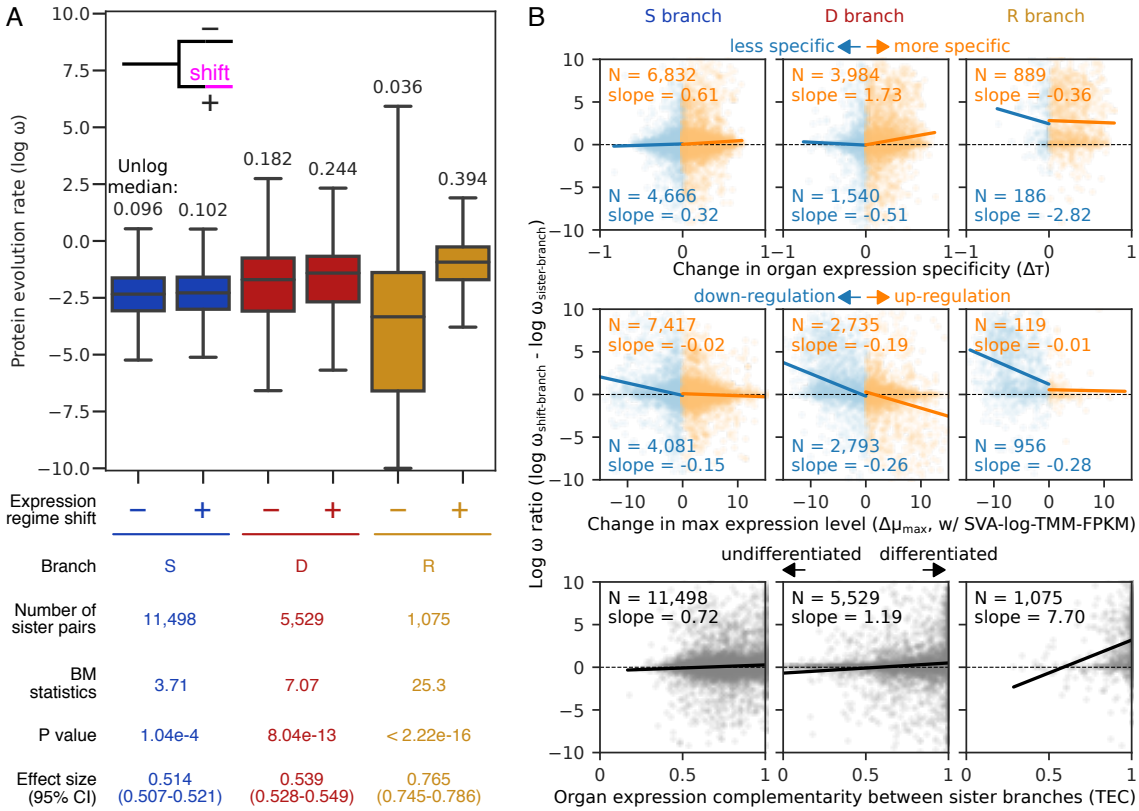


Figure 4. Context-dependent changes of protein evolution rate coupled with expression regime shifts. (**A**) Distribution of ω values. A plus (+) indicates branches with expression shifts, whereas minus (−) branches are sisters to the 'plus' branches. Statistical differences between pairs of distributions 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 exceeding ± 10 were clipped. (**B**) Relationships between protein evolution rate and change in expression properties. Points correspond to expression regime shifts. Dashed lines indicate no between-branch difference in ω. Solid lines show a linear regression. Its slope and number of regime shifts are provided in the plot. Regime shifts with negative and positive changes were separately analyzed for organ expression specificity (upper) and maximum expression level (middle).

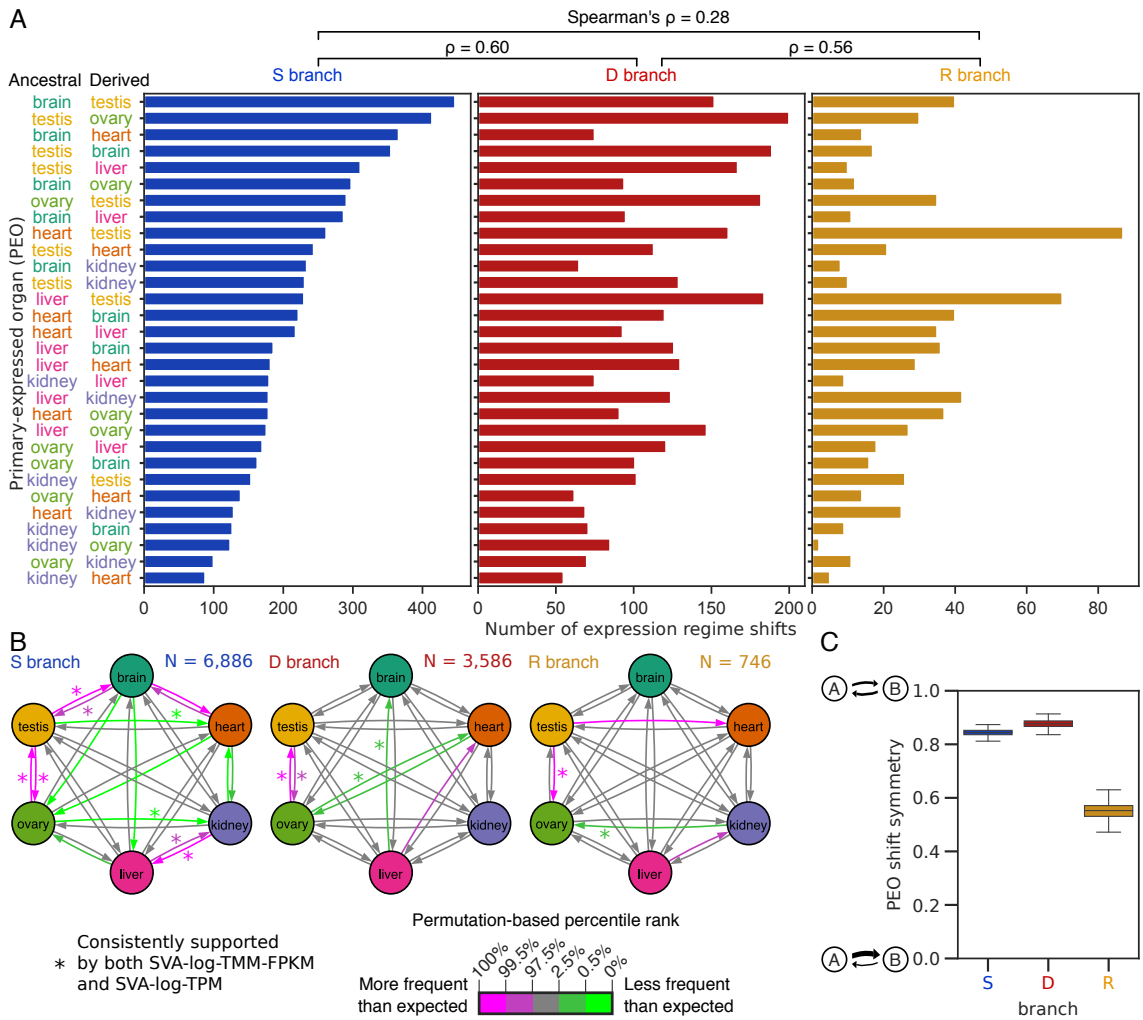


Figure 5. Evolutionary dynamics of gene expression. (**A**) Shift distributions of primary-expressed organs (PEOs). Y axis was sorted by abundance in S branches. Spearman's correlation coefficients among S, D, and R branches are shown above the plots. (**B**) Preadaptation networks in organ expression. Arrows represent transitions from ancestral PEOs to derived PEOs, and its color shows statistical significance based on 10,000 permutations. The results were obtained with SVA-log-TMM-FPKM, and an asterisk (*) indicates the statistical significance supported also by SVA-log-TMM-based analysis (Fig. S11). (**C**) The global polarity of PEO shifts. The global polarity is defined by the scaled sum of differences between two opposite PEO shifts. Boxplots show the distribution estimated by 1,000 bootstrap resampling.

Supplementary Materials for

Organ-specific propensity drives patterns of gene expression evolution

Kenji Fukushima and David D. Pollock

List of Supplementary Materials:

Tables S1–S5 (separate file)

Figs. S1–S14

Supplementary Data (separate file)

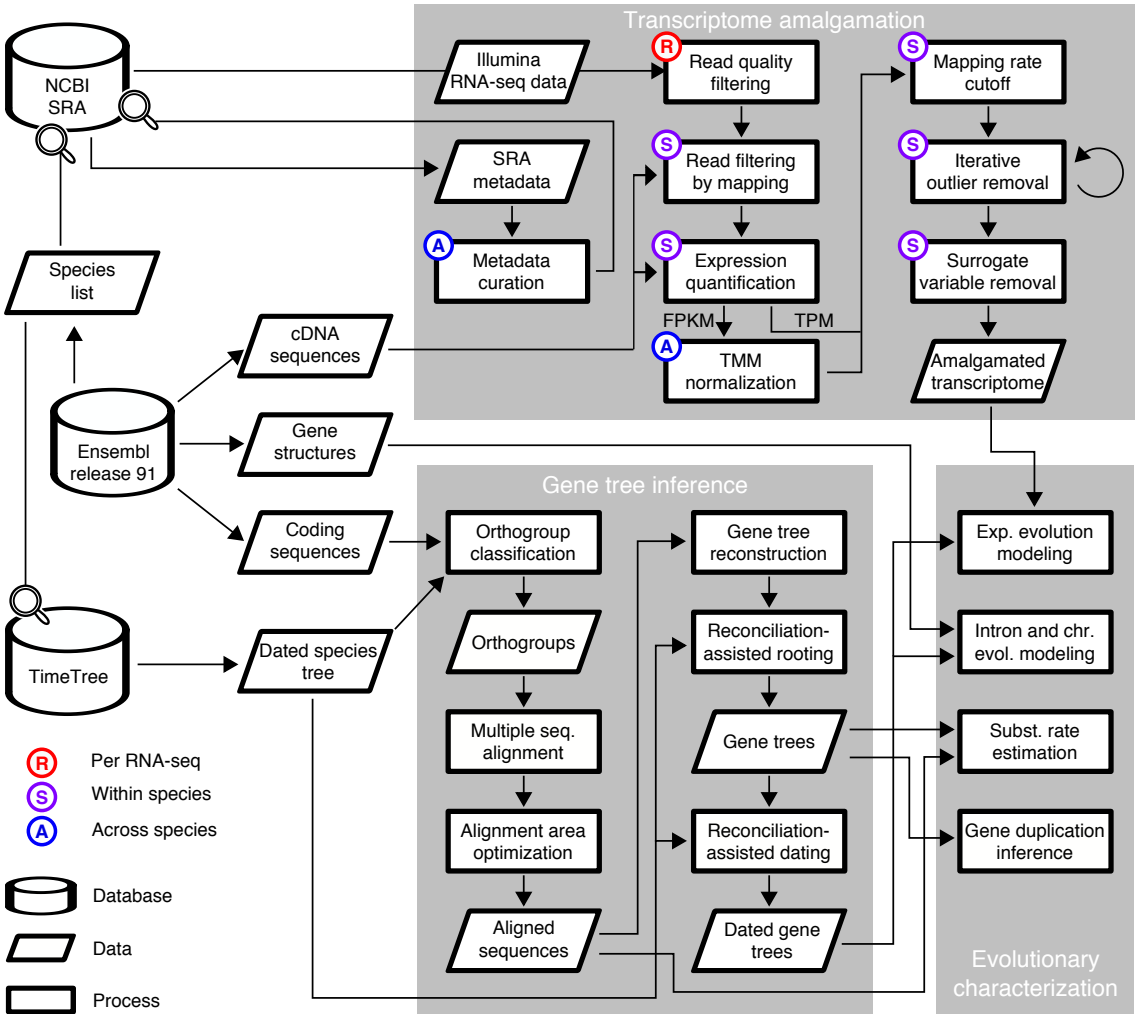


Fig. S1. A flow-chart of transcriptome amalgamation, gene tree inference, and evolutionary characterization in this study.

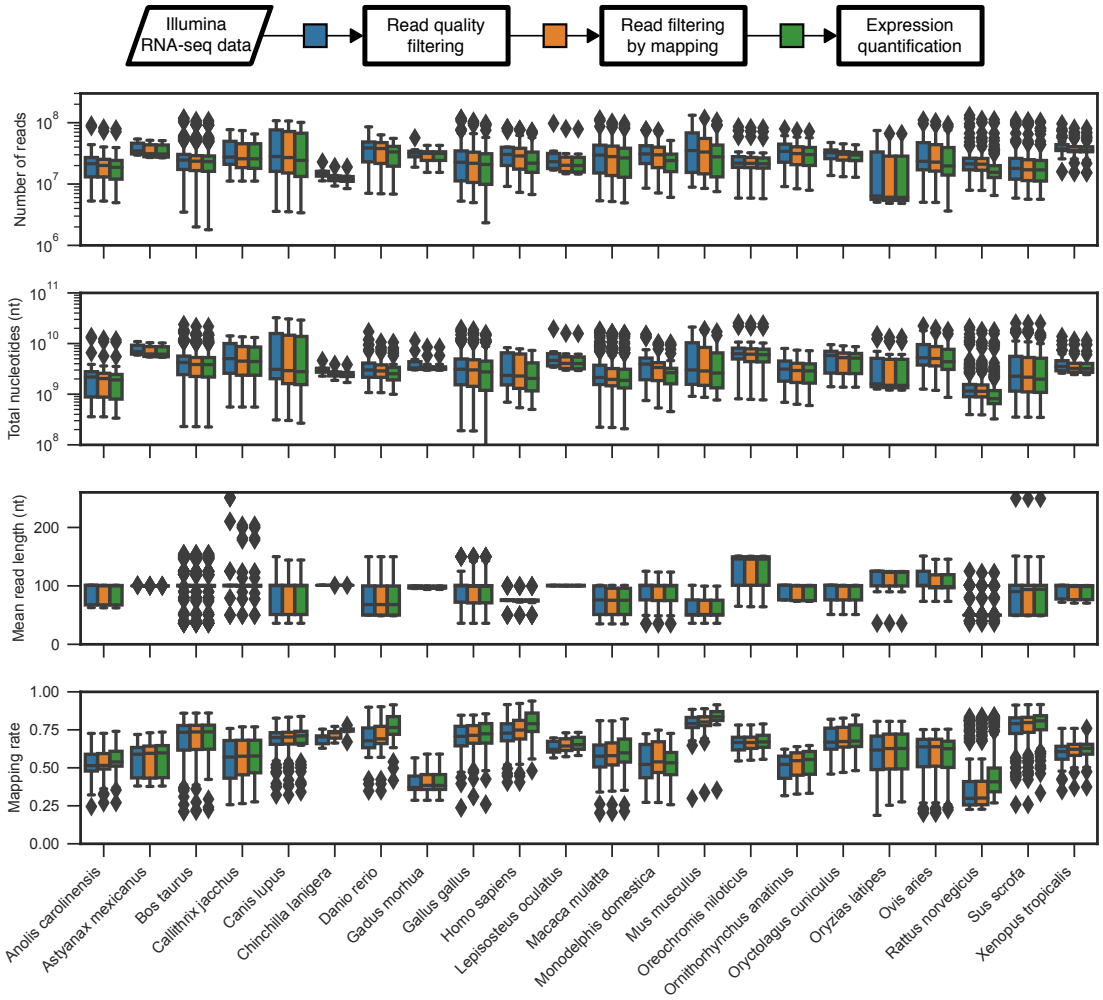


Fig. S2. Changes in number of reads, total nucleotide length, mean read length, and mapping rate by RNA-seq read filtering. Mapping rates were, in almost every case, improved by both read quality filtering and read filtering by mapping to miscellaneous genomic features.

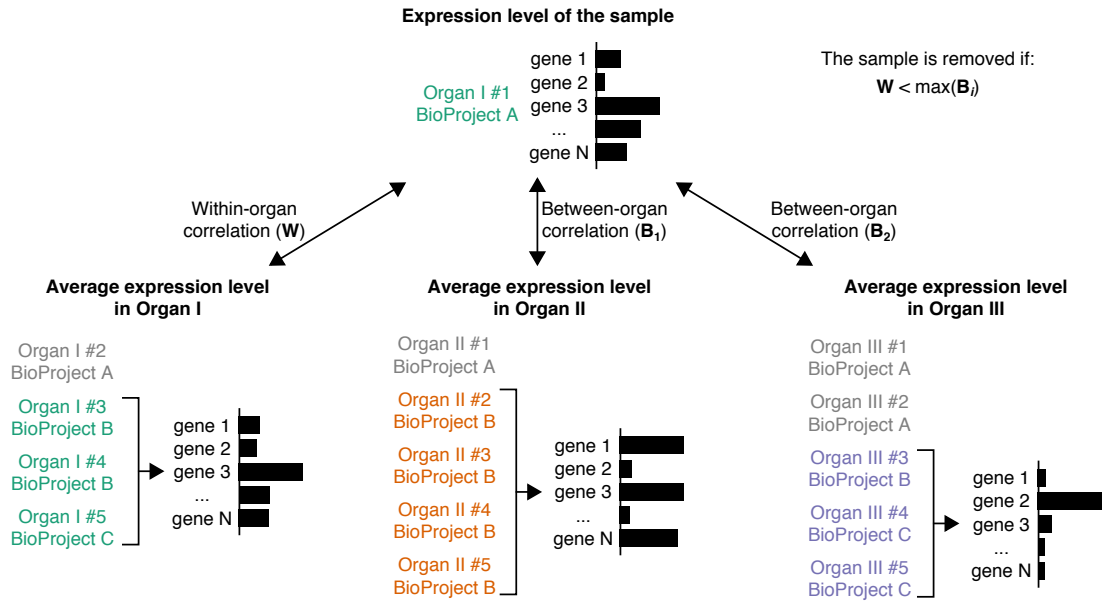


Fig. S3. A correlation analysis for the detection and removal of anomalous RNA-seq samples. Expression levels of all genes were compared between the sample and the organ averages. A sample was removed if any between-organ comparisons yielded a correlation coefficient higher than the within-organ comparison.

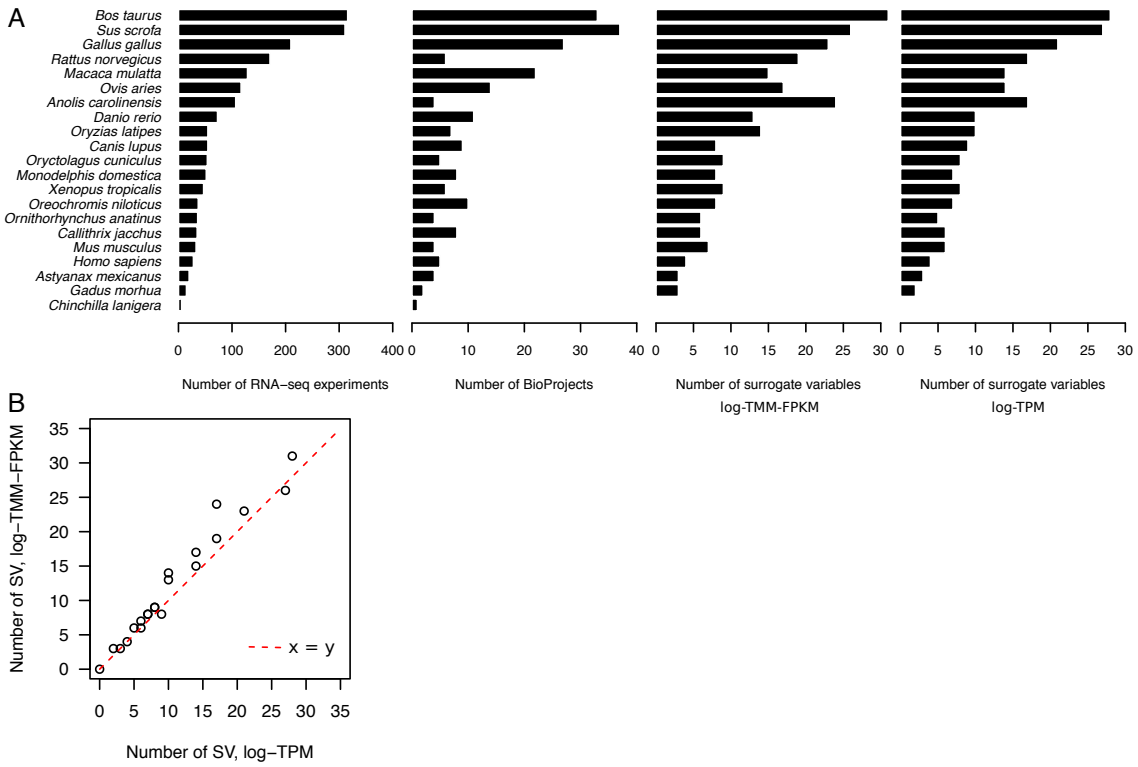


Fig. S4. Characteristics of amalgamated transcriptome. (A) The number of RNA-seq experiments, BioProjects, and detected surrogate variables in each species. The counts were derived from the final dataset. The numbers of surrogate variables (SV) are correlated with the numbers of RNA-seq samples and BioProjects. (**B**) Relationships of the number of SVs with log-TPM and log-TMM-FPKM. Points correspond to species. (**C–D**) Correlation heatmaps of corrected transcriptomes. See Table S1 for full descriptions including RNA-seq sample IDs and BioProject IDs. (**E–F**) 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. (**G–H**) 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. (**I–J**) Multispecies correlation analysis of averaged organ expression. Corrected expression levels of 1,377 single-copy orthologs were used to calculate pairwise Pearson's correlation coefficients. (**K**) A principal component analysis using expression levels of 1,377 single-copy orthologs from 21 species. Points correspond to RNA-seq samples. Curves show the estimated kernel density. Explained variations in percentages are indicated in each axis. (**L**) Estimated organ-wise expression levels of a housekeeping gene. Since data from relatively many BioProjects are available, glyceraldehyde-3-phosphate dehydrogenase gene (GAPDH, ENSGALG00000014442) in *Gallus gallus* is shown. Points correspond to the average expression level calculated by a random subsampling. All data points and the median value (bar), rather than a box plot, are shown if the number of subsampled BioProject combinations is less than 10.

C SVA-log-TMM-FPKM

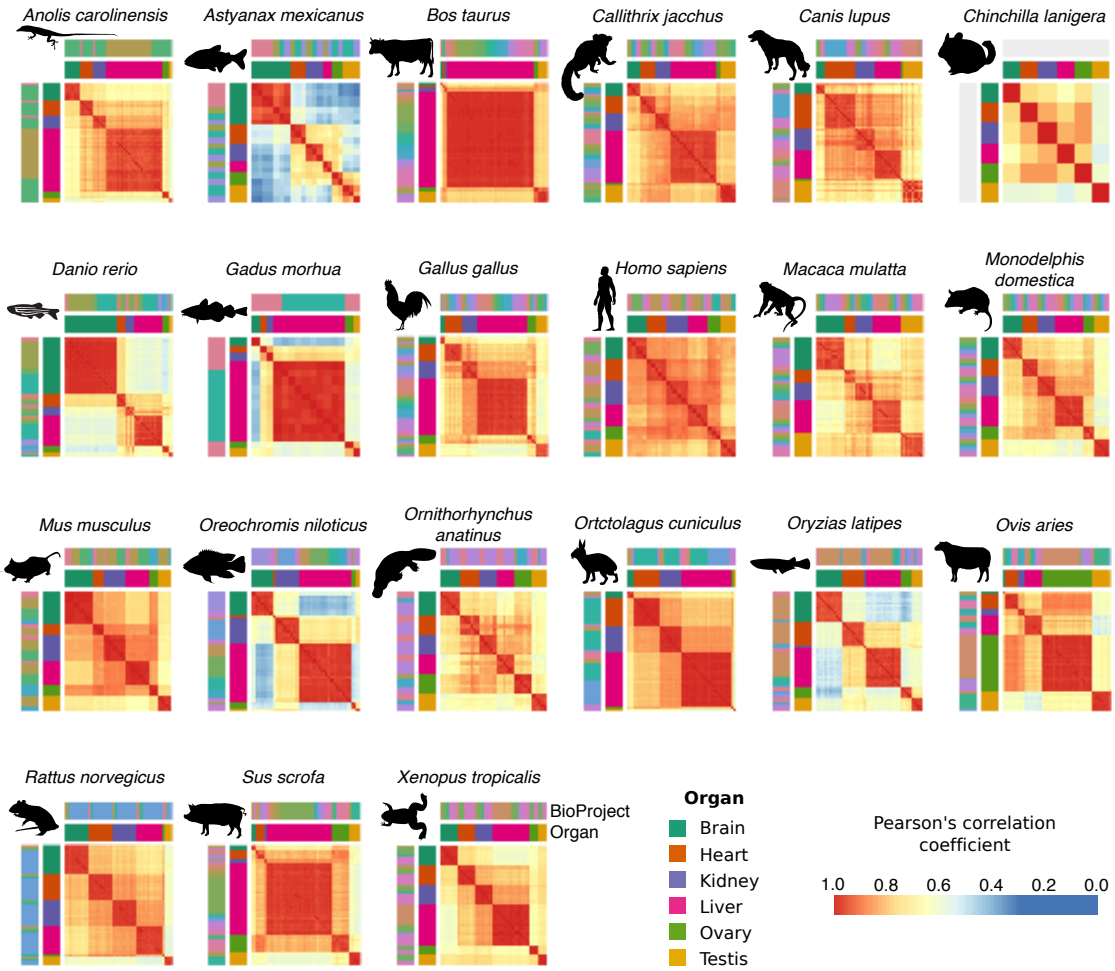


Fig. S4 (continued)

D SVA-log-TPM

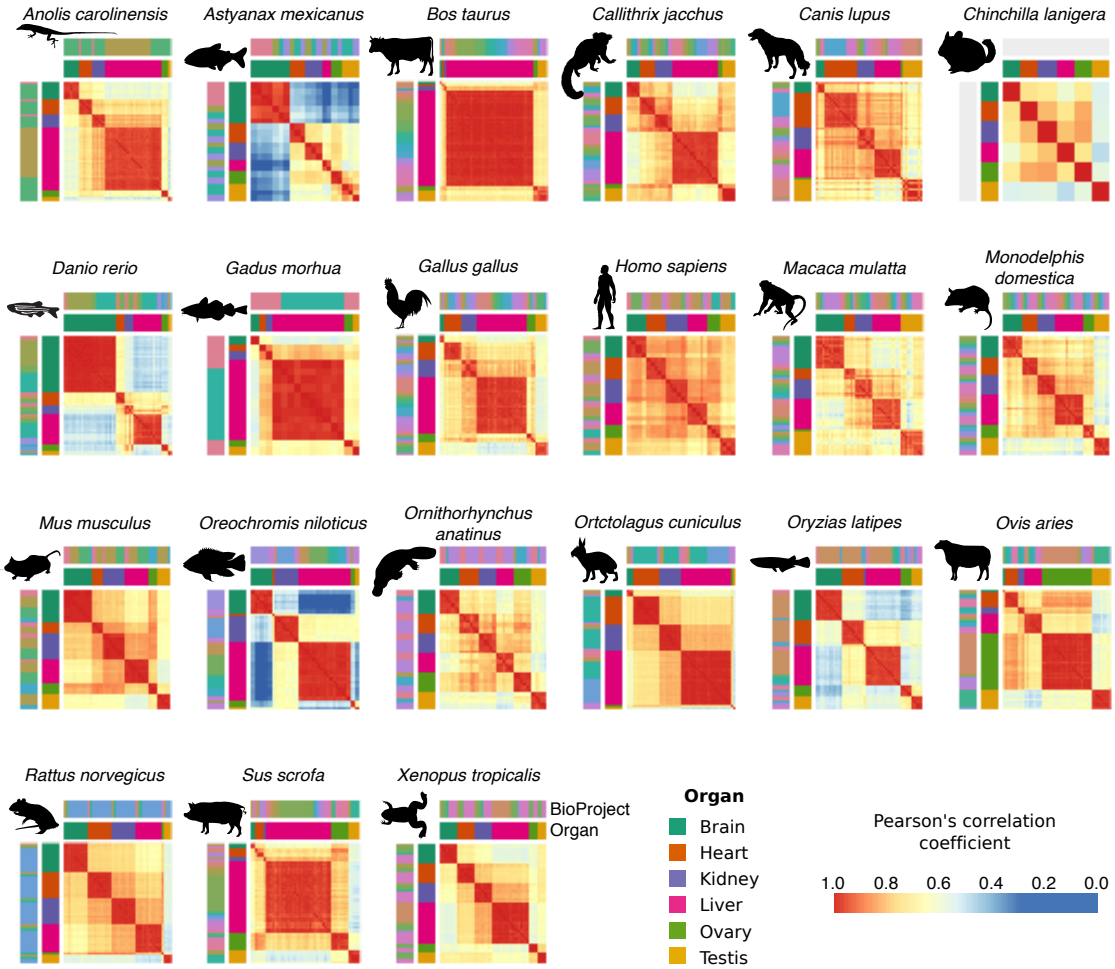


Fig. S4 (continued)

E SVA-log-TMM-FPKM

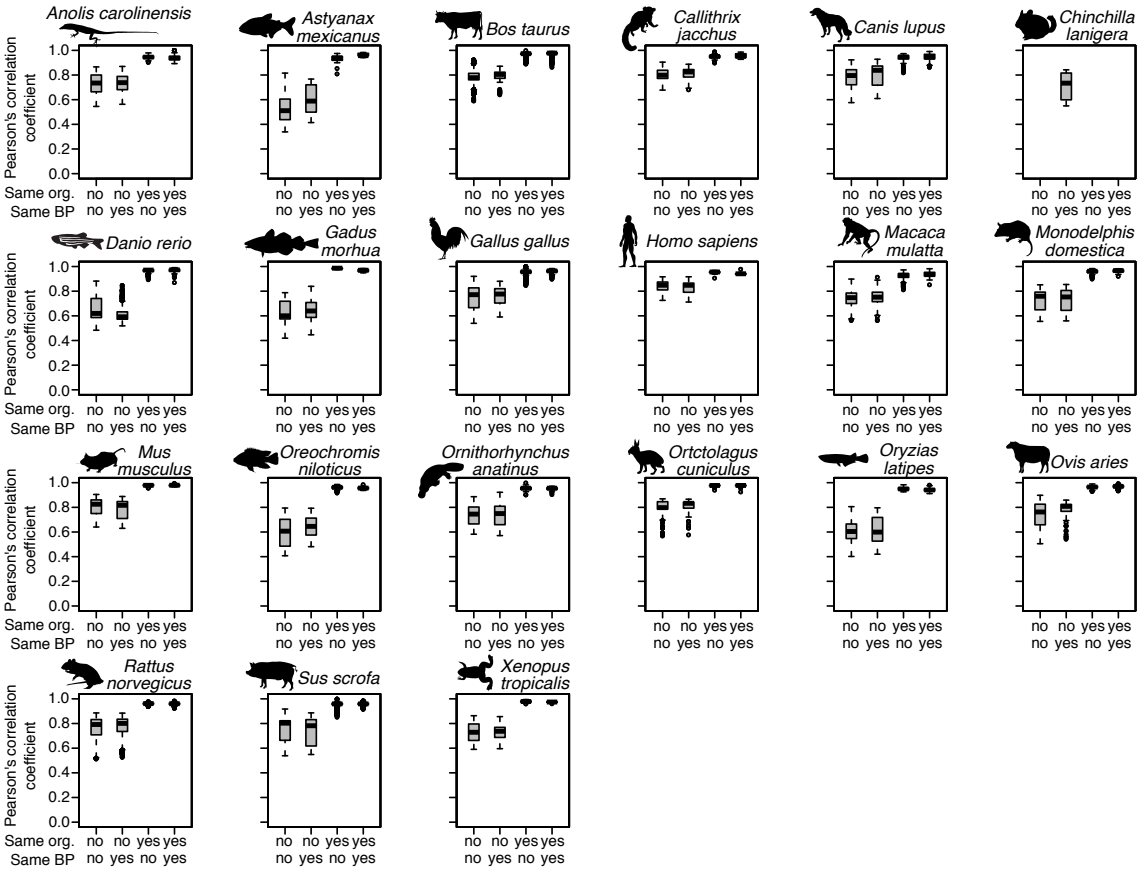


Fig. S4 (continued)

F SVA-log-TPM

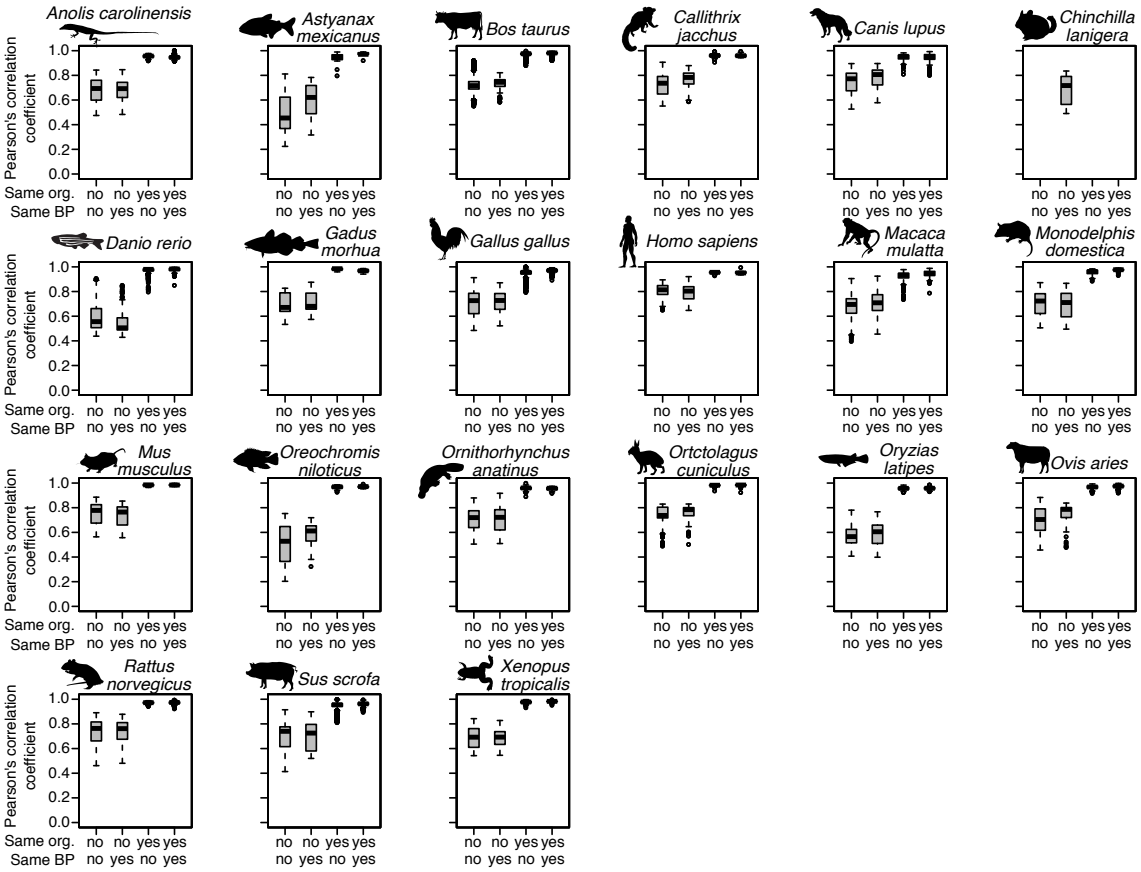


Fig. S4 (continued)

G SVA-log-TMM-FPKM

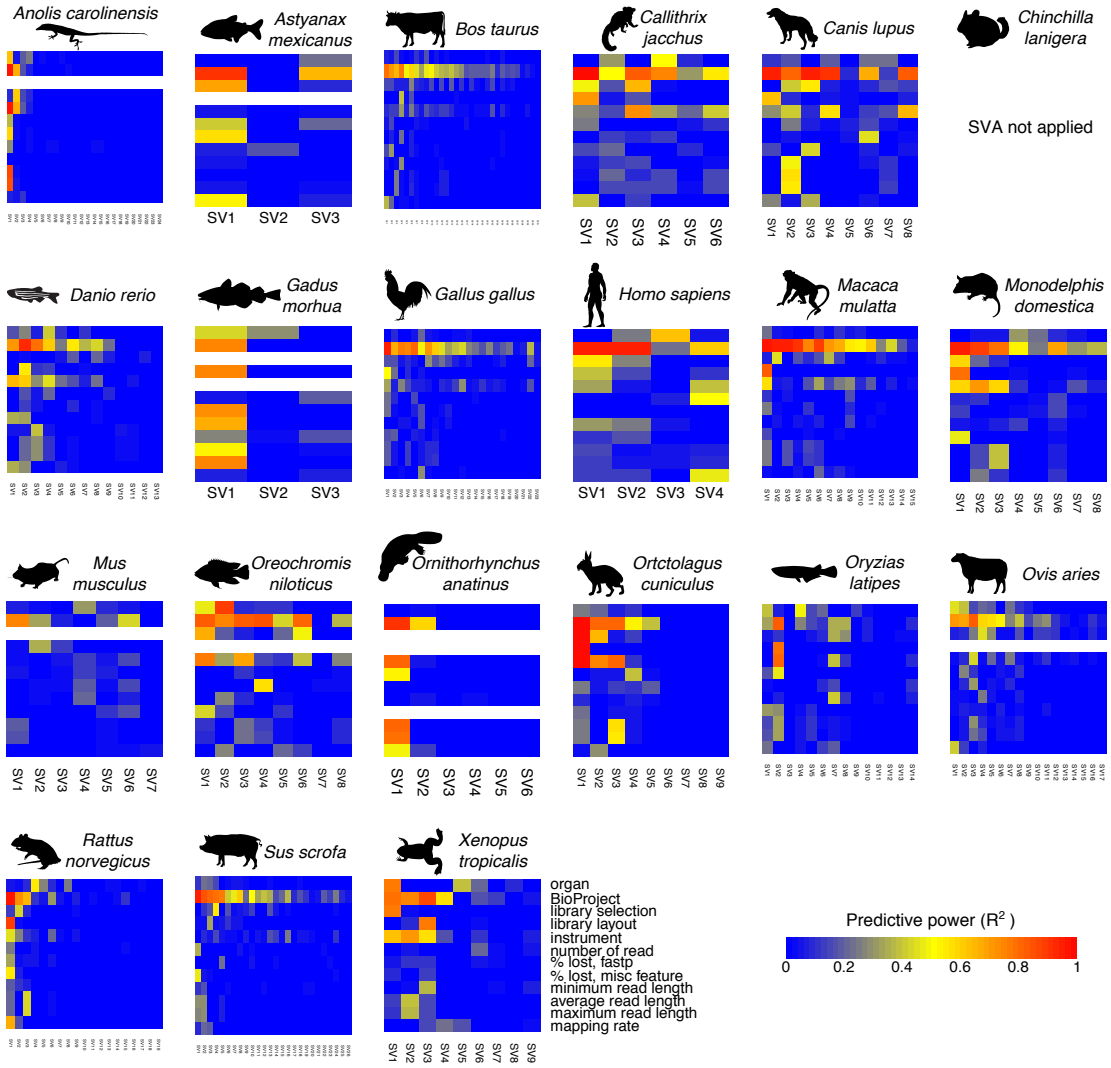


Fig. S4 (continued)

H SVA-log-TPM

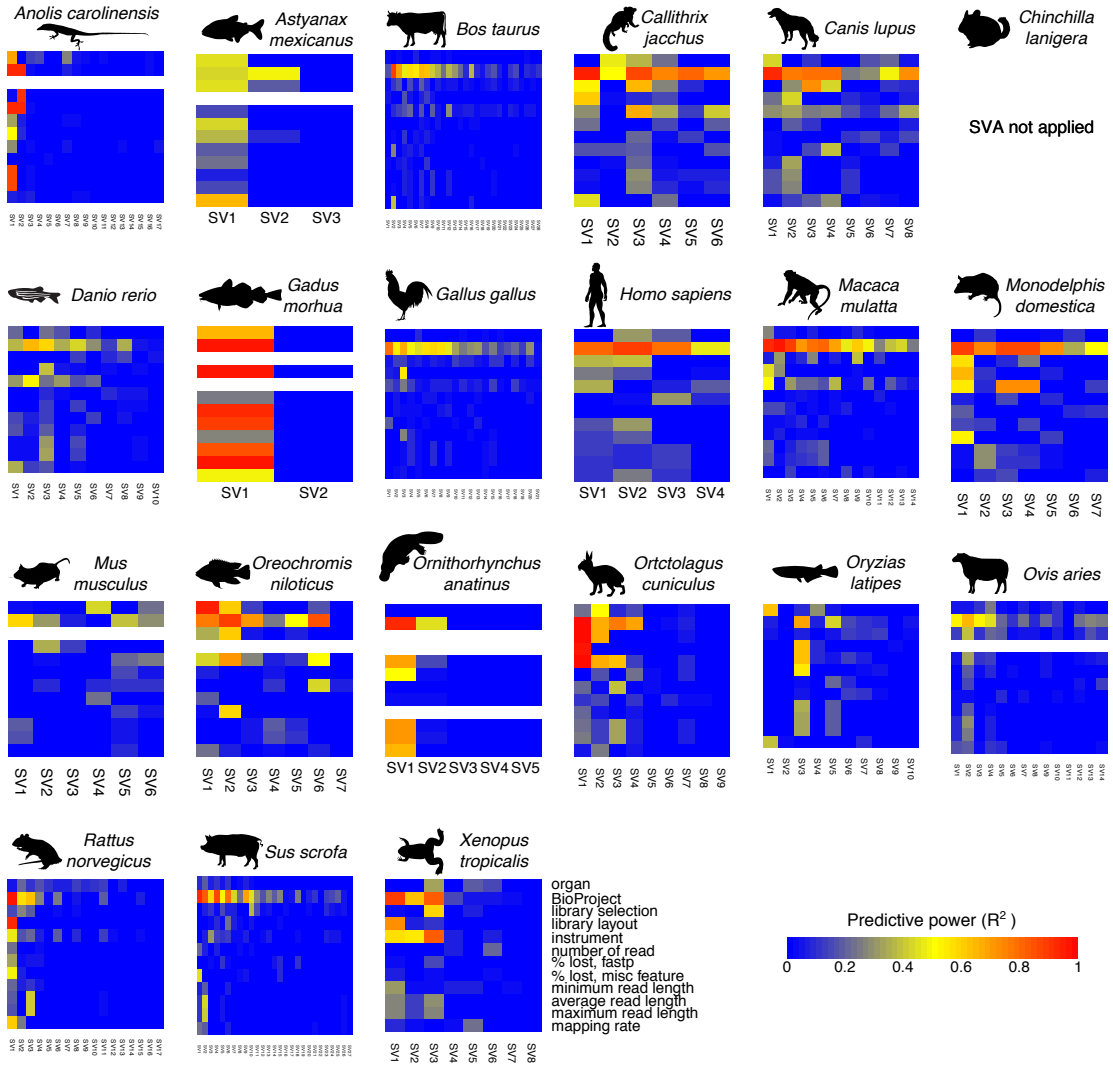
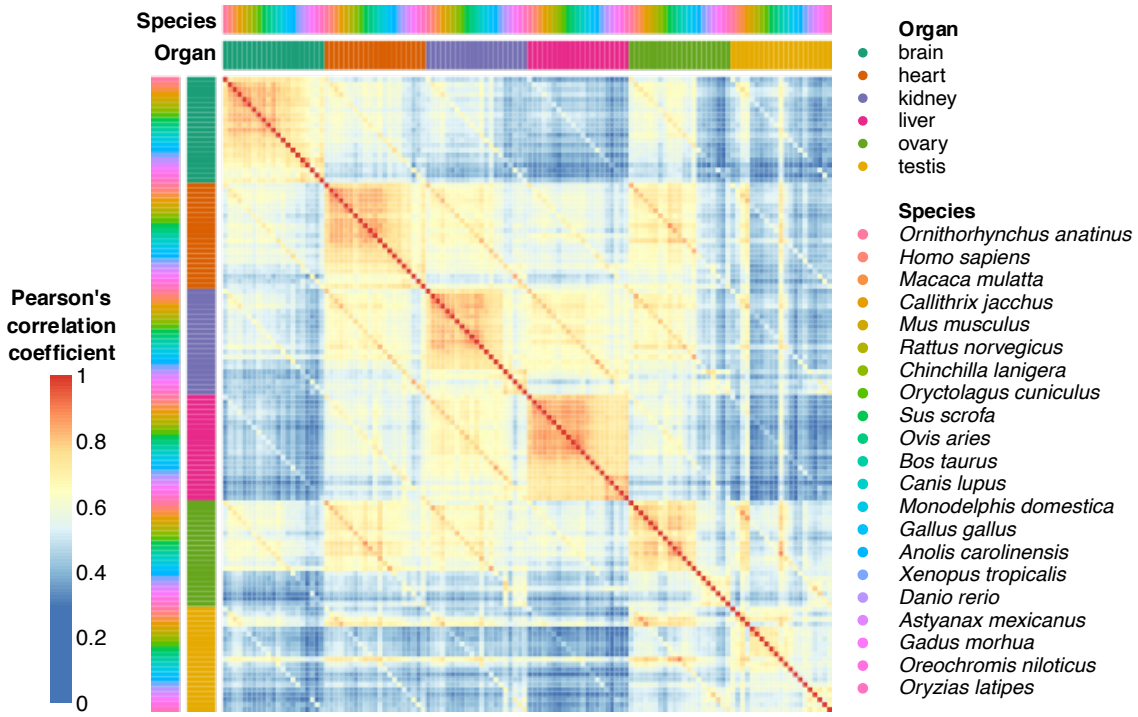


Fig. S4 (continued)

I SVA-log-TMM-FPKM



J SVA-log-TPM

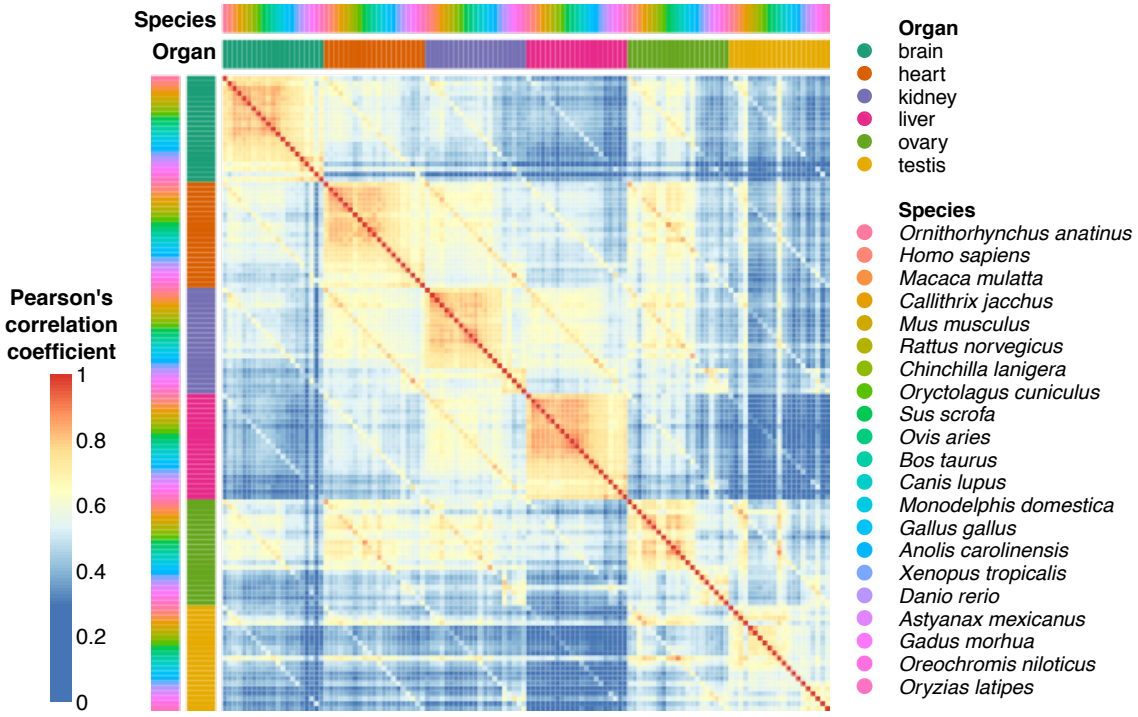
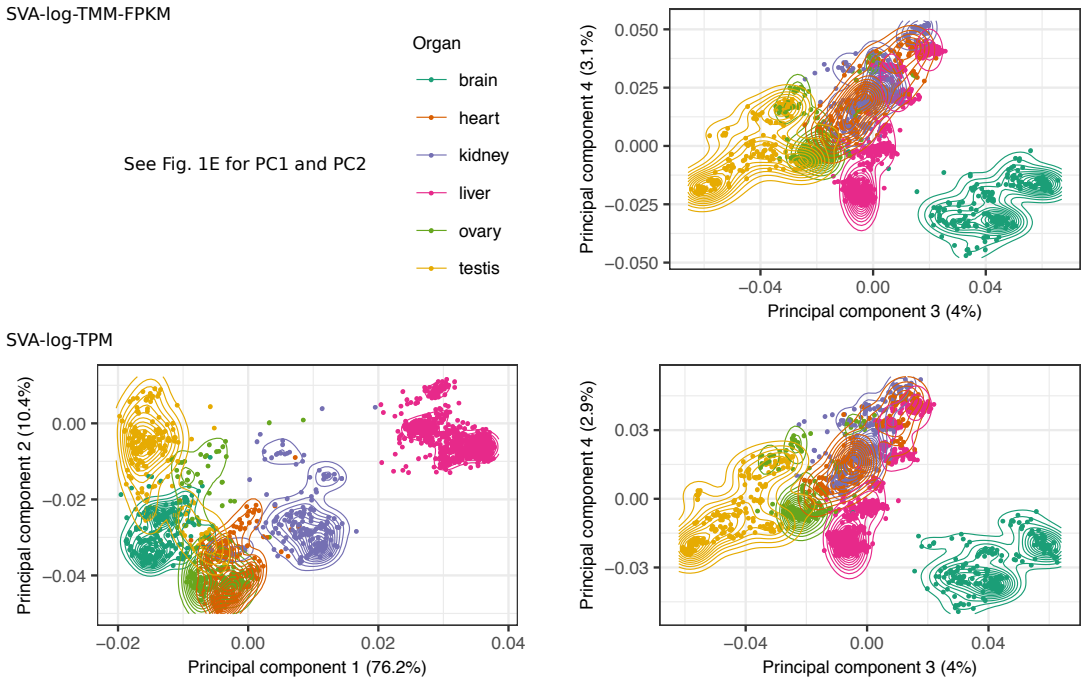


Fig. S4 (continued)

K



L

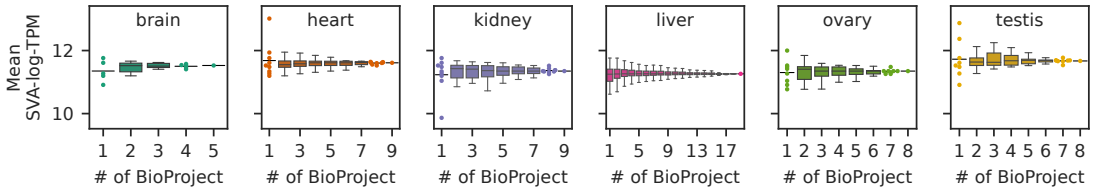


Fig. S4 (continued)

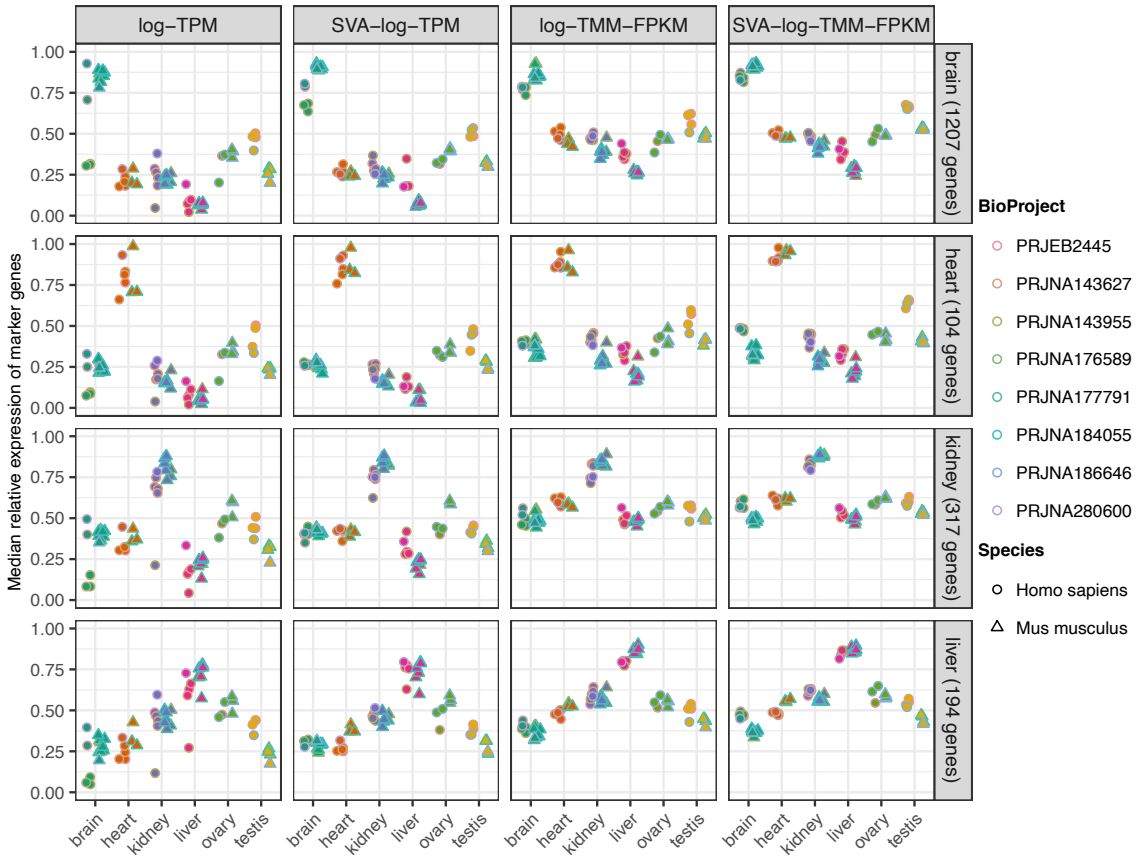


Fig. S5. Expression of organ-specific marker genes in human and mouse. Marker genes were retrieved from PanglaoDB (Franzén et al., 2019), and its median expression values were obtained for each RNA-seq sample. A cell-type-wise analysis is provided in Supplementary Data. Cell types in ovary and testis were not included in PanglaoDB (access date: April 1, 2020).

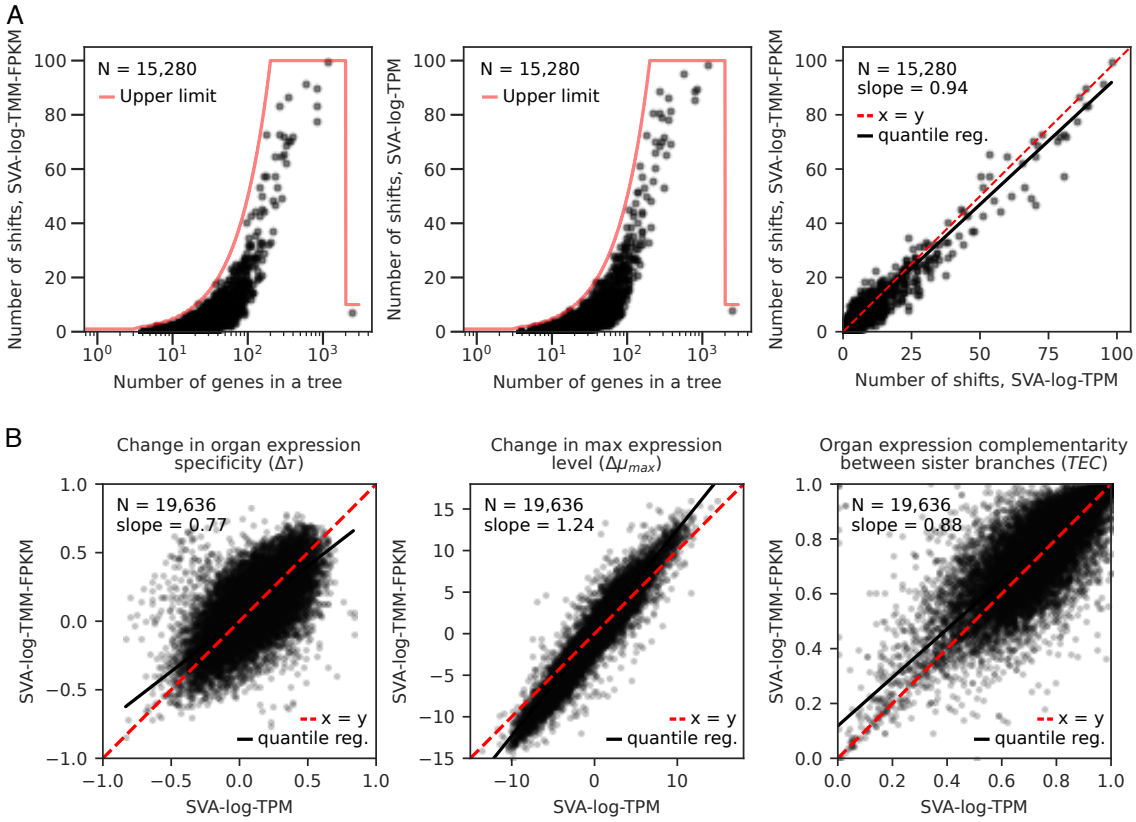


Fig. S6. Comparison of SVA-log-FPKM and SVA-log-TPM in expression regime shift detection. (**A**) The numbers of expression regime shifts. Each point corresponds to a gene tree. The solid red lines show the upper limits in the regime shift search. The numbers of shifts detected with SVA-log-TMM-FPKM (left) and SVA-log-TMM (center) are well correlated (right). The black line is a quantile regression, and the dashed red line shows a slope of 1. (**B**) Expression properties. Each point corresponds to an expression regime shift which is consistently detected by SVA-log-TMM-FPKM and SVA-log-TPM. (**C–D**) The branch-wise numbers of detected shifts in the species tree. The shifts were categorized by primary-expressed organs (PEOs) in ancestral and derived states. The order of the species tree branches (y axis) corresponds to the total number of corresponding gene tree branches in the dataset. See Fig. 3A for branch ids.

C SVA-log-TMM-FPKM

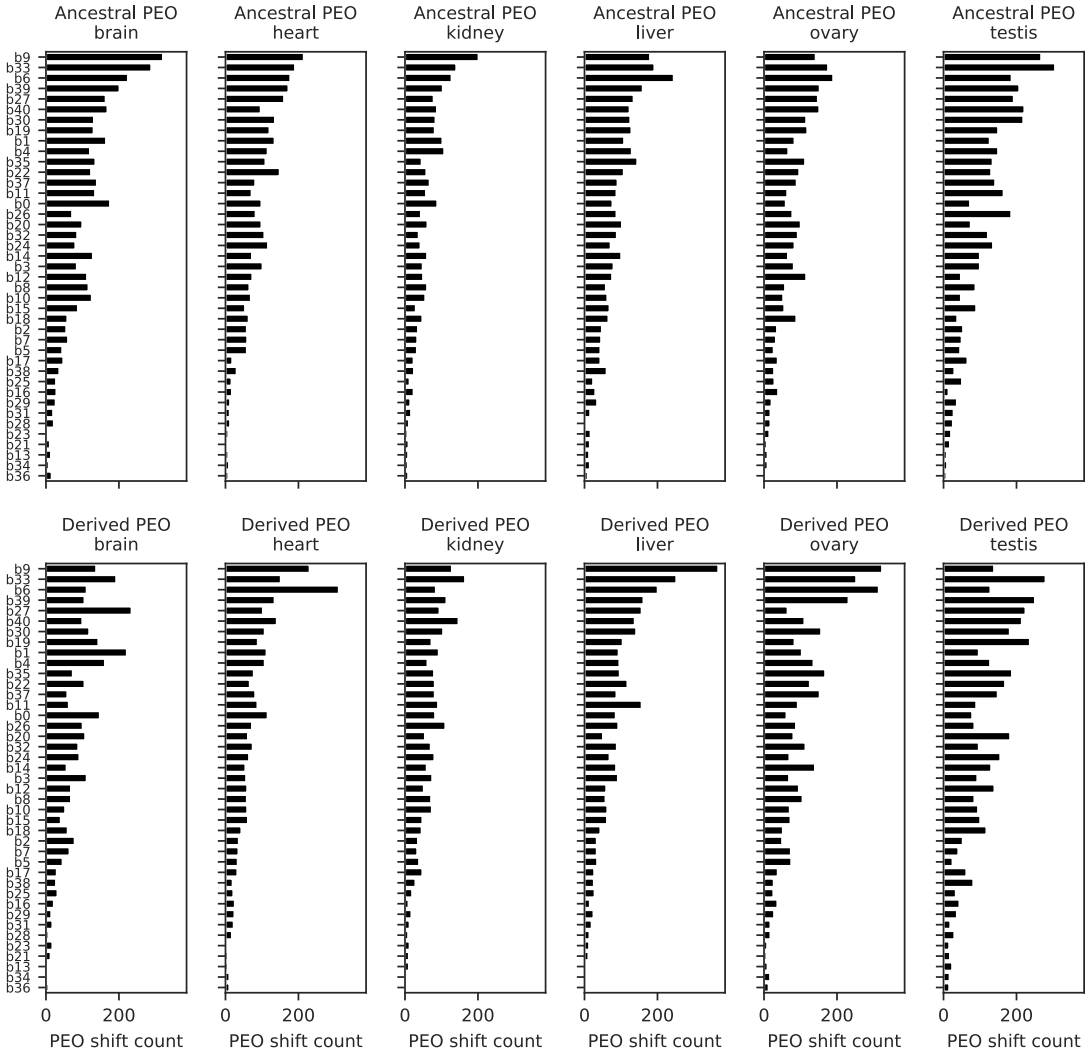


Fig. S6 (continued)

D SVA-log-TPM

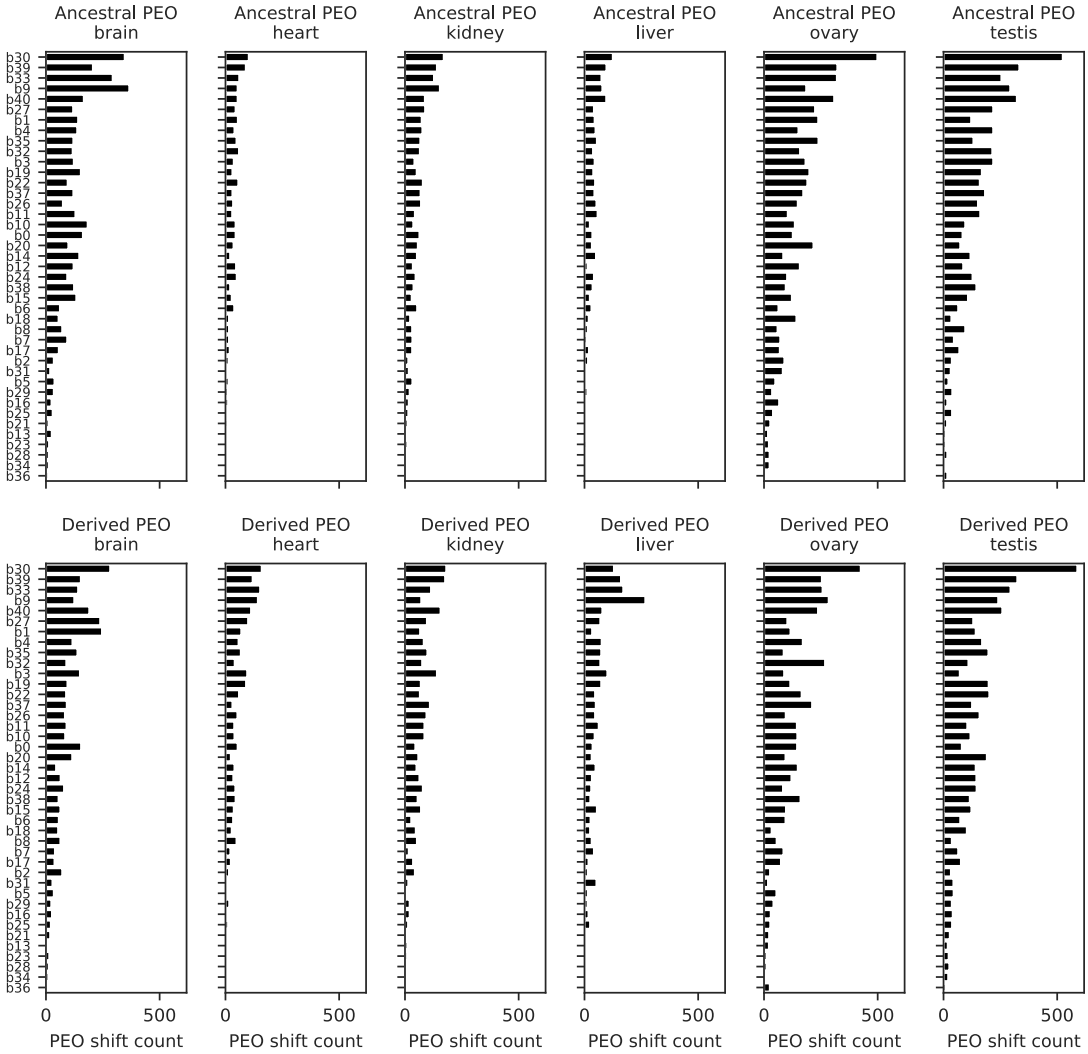


Fig. S6 (continued)

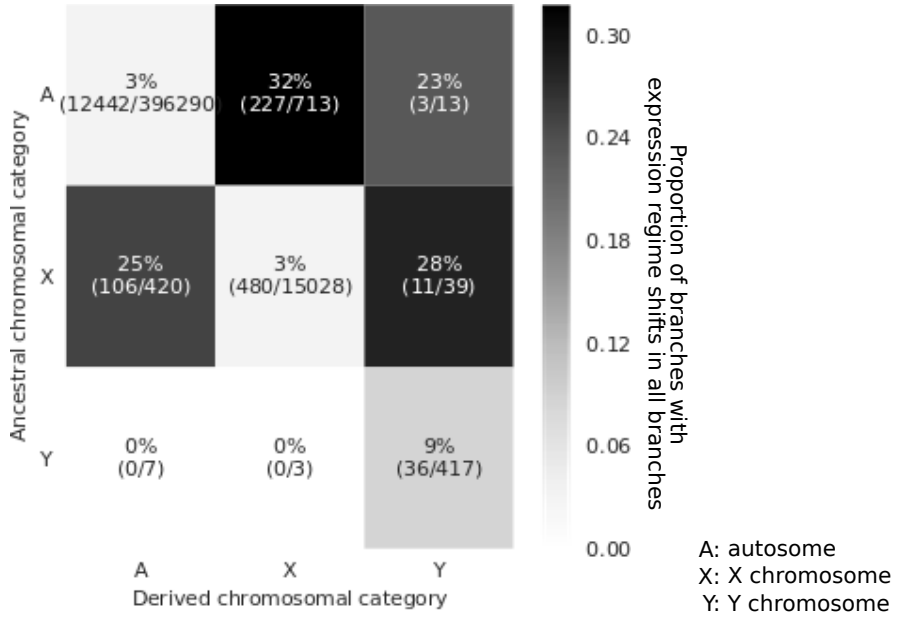


Fig. S7. The relationships between expression shifts and chromosomal location. The heatmap shows the frequency of expression shifts observed among the branches with or without a change in the chromosomal category (non-diagonal or diagonal, respectively). Chromosomal locations were categorized into autosomes (A), X chromosome (X) and Y chromosome (Y), and the ancestral locations were inferred by stochastic character mapping.

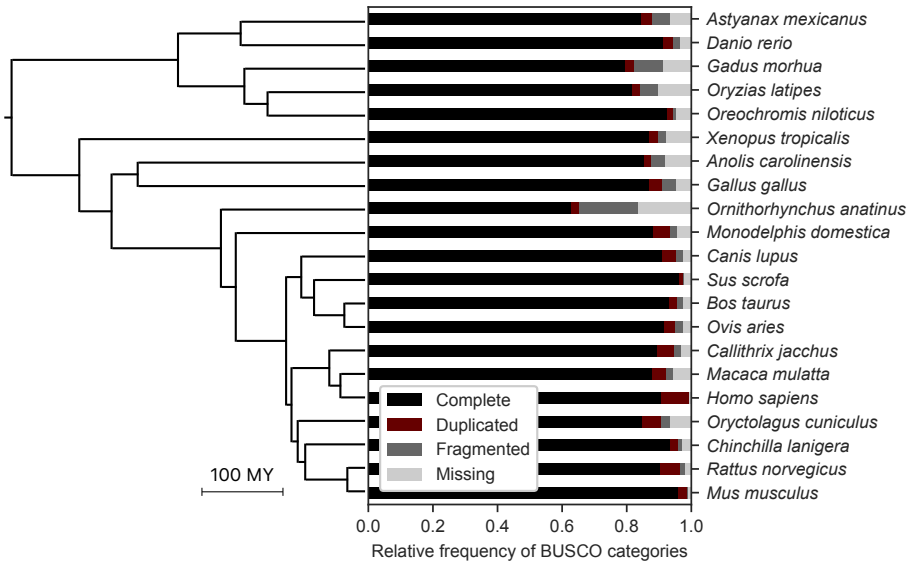


Fig. S8. Evaluation of gene set completeness. BUSCO analysis was performed for gene sets from 21 species using 3,407 single-copy orthologs in the dataset "vertebrata_odb10". The species tree is shown to visualize lineage-specific trends.

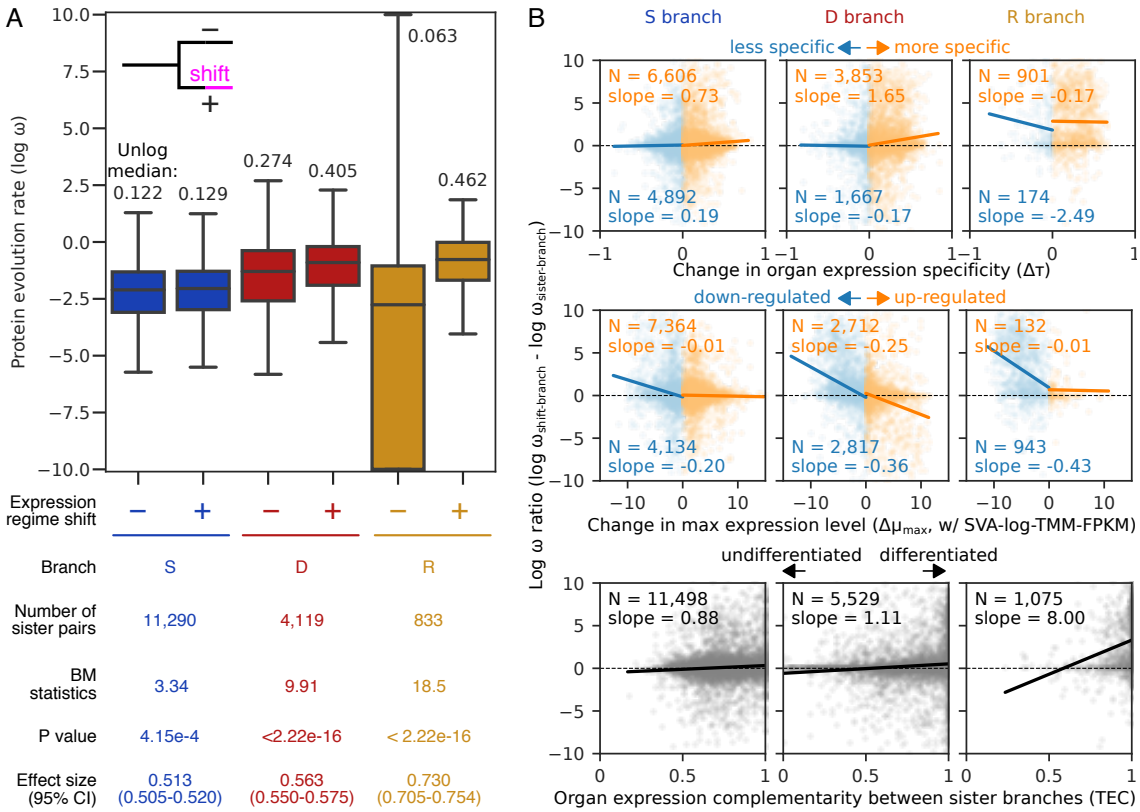


Fig. S9. Alternative analysis supporting non-linear change in protein evolution rate in correlation with expression regime shifts. (**A**) Distribution of ω values. While the results from stochastic character mapping (map*dNdS*) are reported in the main text, results from maximum-likelihood estimation (HyPhy) are shown here. 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 exceeding ±10 were clipped. (**B**) Relationships between protein evolution rate and change in expression properties. While SVA-log-TMM-FPKM-based analysis is reported in the main text, results from SVA-lot-TPM-based analysis is shown here. Stochastic character mapping was used to obtain branch-wise ω values. Points correspond to expression regime shifts (log ω ratio = 0). Dashed lines indicate no between-branch difference in protein evolution rate. Solid lines show a linear regression. Its slope and number of regime shifts are also provided. Regime shifts with negative and positive changes were separately analyzed for organ specificity (upper) and expression level (middle).

1225

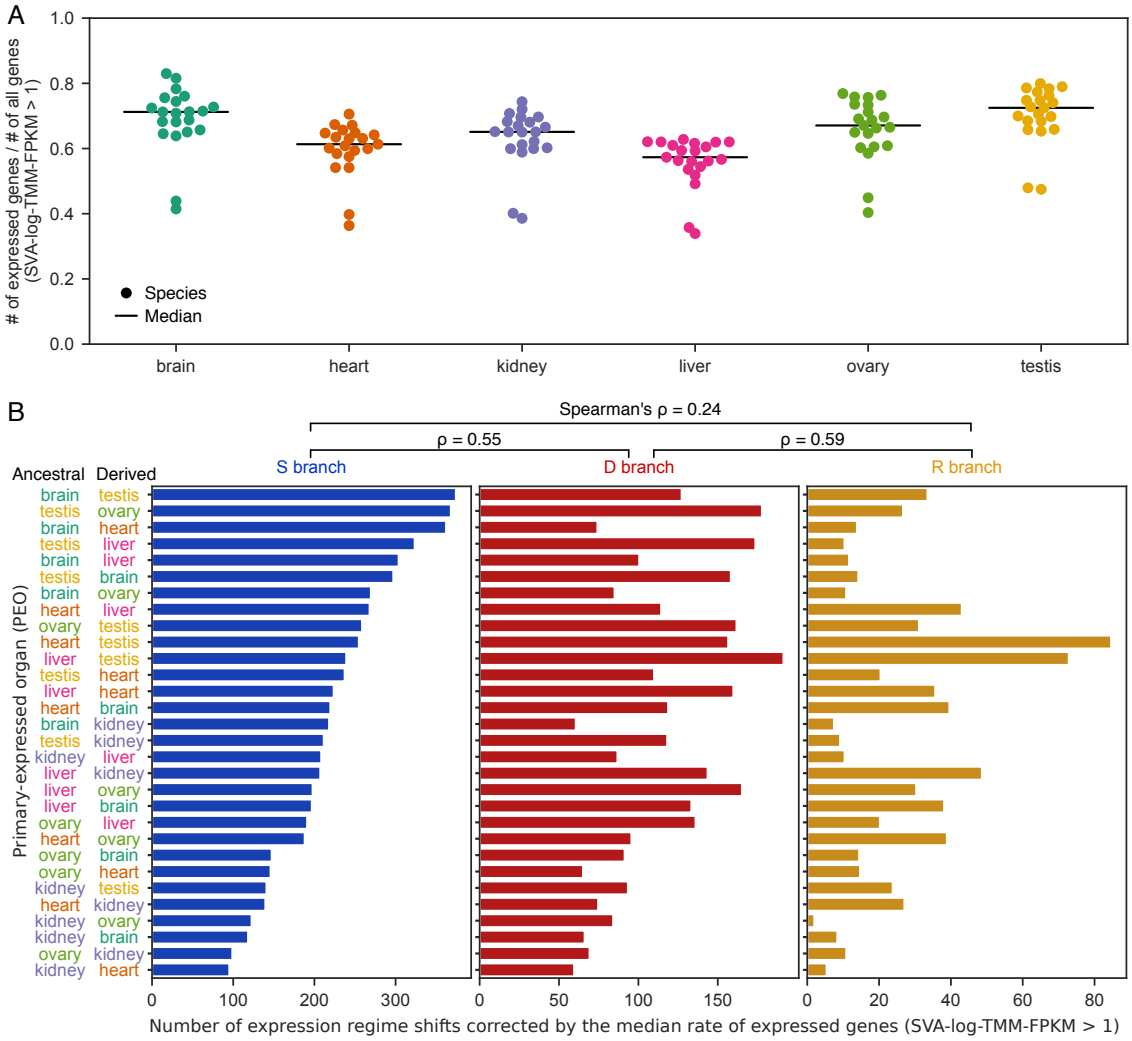


Fig. S10. The number of expressed genes does not explain the organ-wise abundance of PEO shifts. (**A**) The organ-wise numbers of expressed genes in the 21 species. In this analysis, expressed genes are defined as genes with >1 SVA-log-TMM-FPKM. (**B**) The PEO shift distributions corrected by the median numbers of expressed genes. The data in Fig. 5A were corrected as follows: $S_{corrected} = S_{original} \div (M_{ancestral}/\bar{M}) \div (M_{derived}/\bar{M})$. $S_{original}$ corresponds to the numbers shown in Fig. 5A. $M_{ancestral}$ and $M_{derived}$ are median numbers of expressed genes in ancestral and derived PEOs, respectively. $\bar{M}$ indicates the across-organ average of expressed gene numbers.

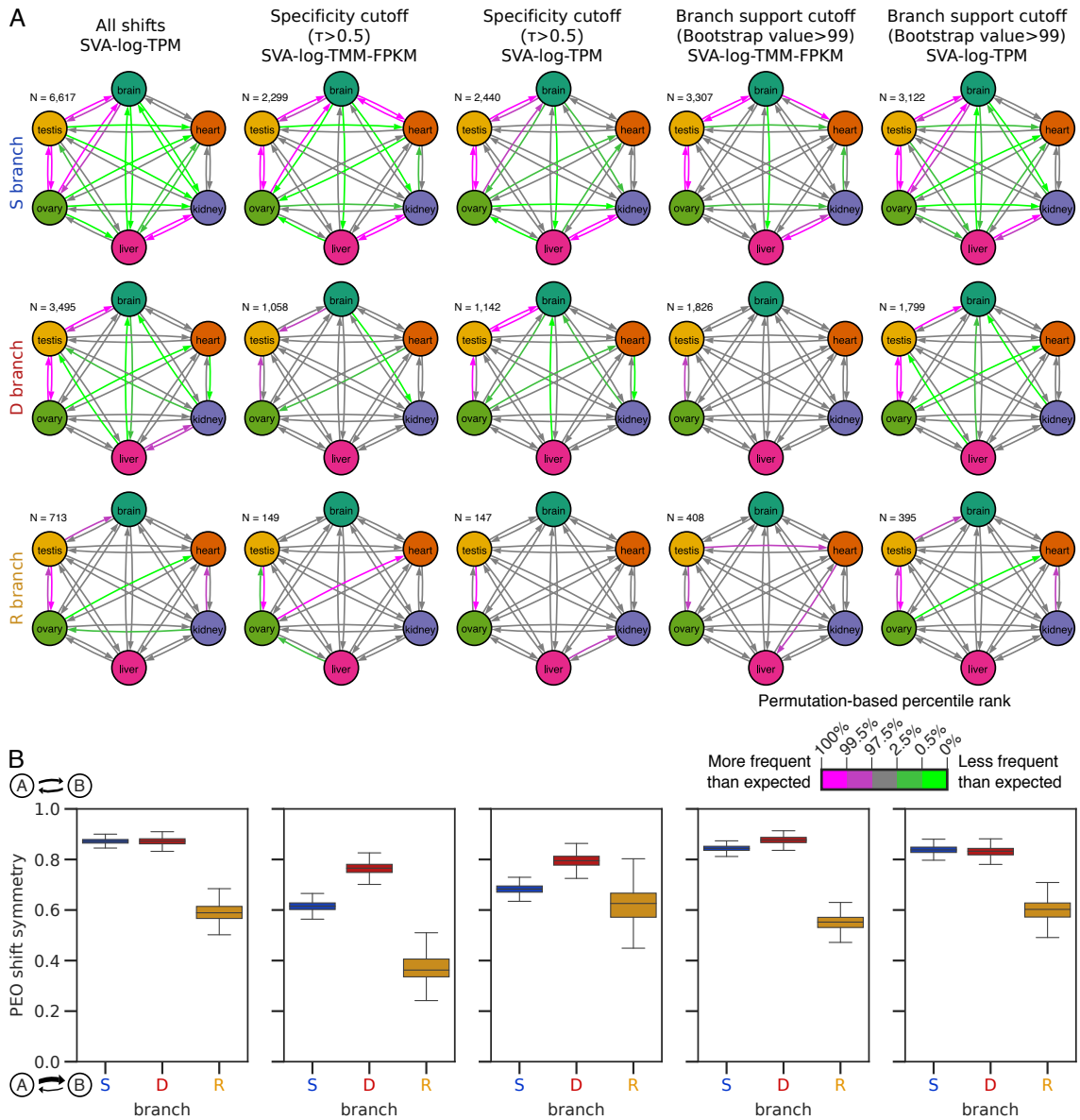


Fig. S11. Evolutionary dynamics of gene expression analyzed with conservative datasets. Both SVA-log-TPM and SVA-log-TMM FPKM values were analyzed. Organ-specific genes were examined by selecting PEO shifts with high ancestral and derived organ expression specificity ($\tau > 0.5$). To examine the effect of inference errors in gene tree reconstruction, branches with high support values (ultrafast bootstrap percentage > 99) were analyzed. **A** and **B** correspond to Fig. 5B and C, respectively.

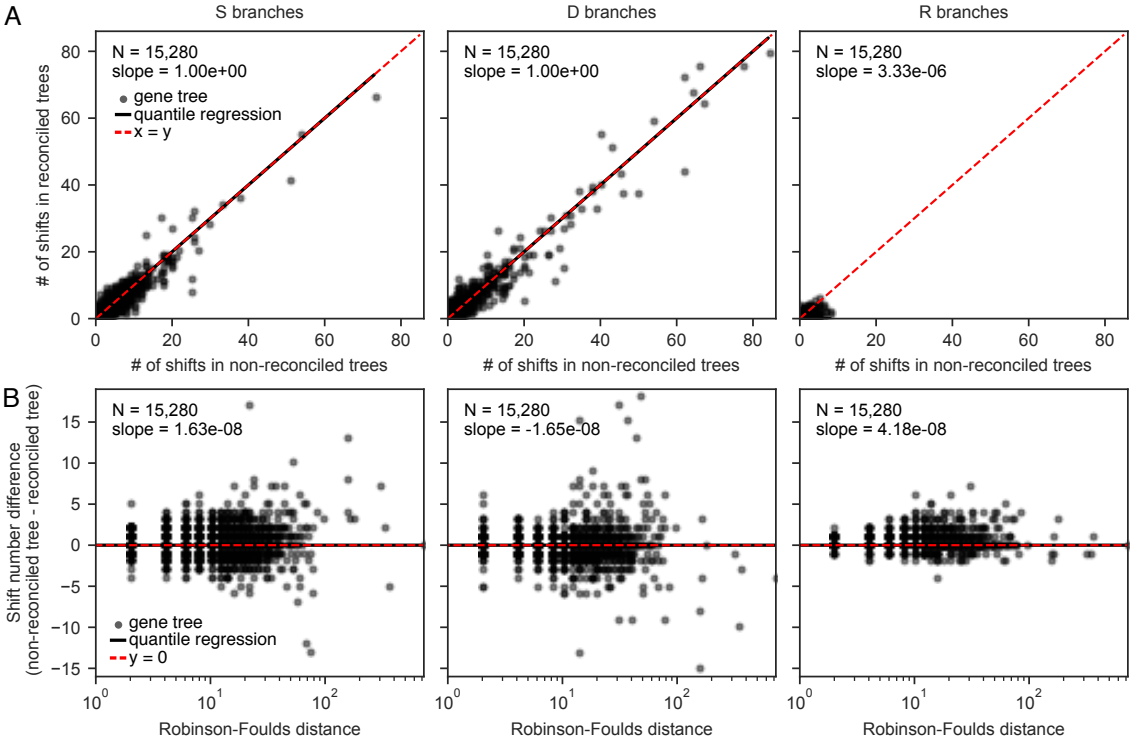


Fig. S12. Effects of phylogeny reconciliation. (**A**) The numbers of detected shifts in non-reconciled and reconciled maximum-likelihood gene trees. (**B**) The relationship between shift number difference and tree topology difference measured by Robinson-Foulds distance [51]. Trees with no detected shifts were removed from the analysis. (**C–E**) Reproduction of PEO shift distributions using reconciled gene trees. C, D, and E correspond to Fig. 5A, Fig. 5B, and Fig. 5C, respectively.

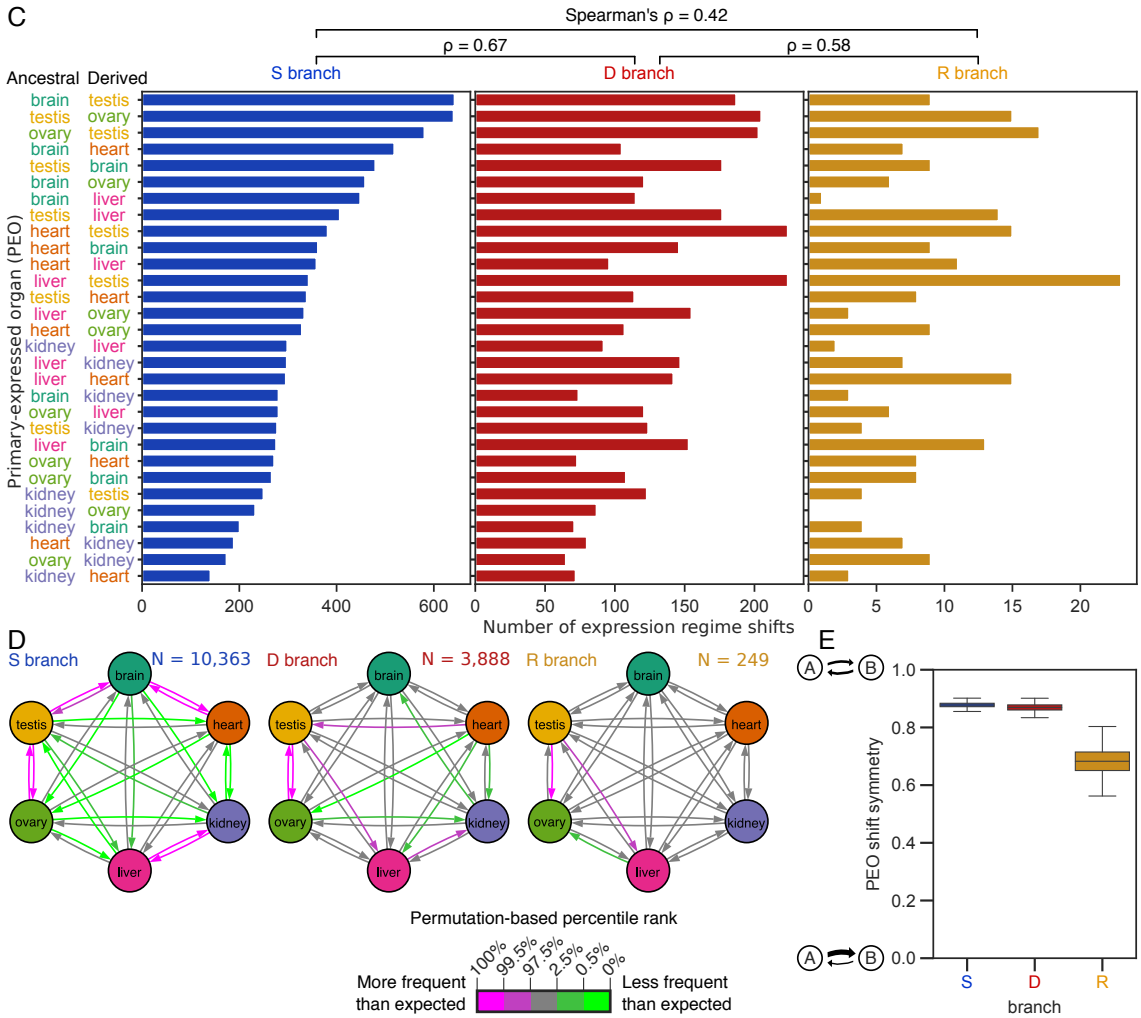


Fig. S12 (continued)

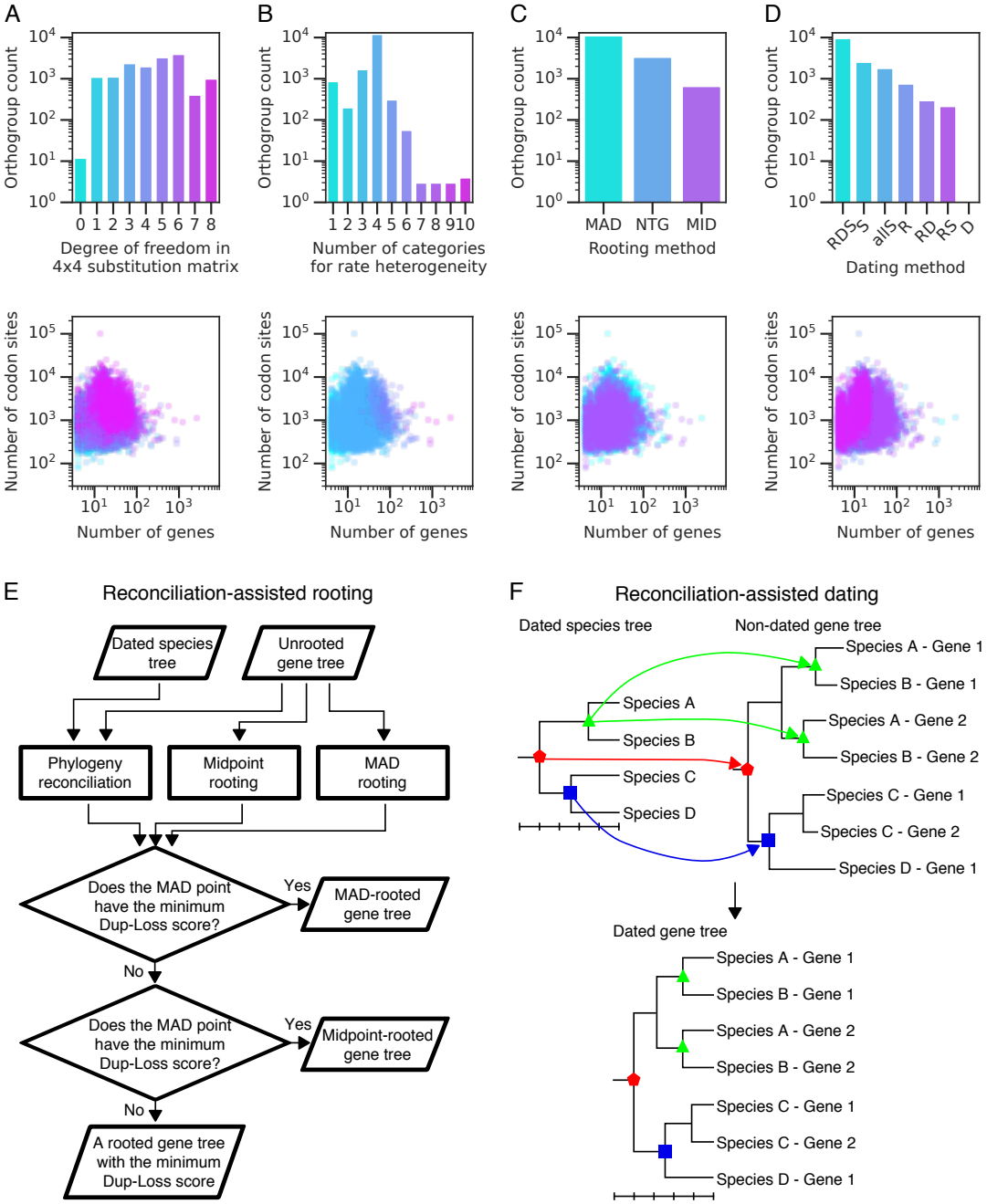


Fig. S13. Gene tree reconstruction. (**A**) Complexity of best-fit nucleotide substitution matrices. (**B**) Complexity of rate heterogeneity among nucleotide sites. Both discrete Gamma models (Yang, 1994) and FreeRate models (Soubrier et al., 2012; Yang, 1995) were included to count the number of categories 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

1264 successively relaxed if estimation fails. In the category 'allS', all nodes are speciation
1265 nodes and therefore no divergence time estimation was performed for the trees. (**E**)
1266 Reconciliation-assisted gene tree rooting. Rooting points were estimated with two
1267 different methods: the minimum ancestor deviation (MAD) method and midpoint
1268 rooting. If they were compatible with the event parsimony involving gene duplication
1269 and loss, the MAD- or midpoint-rooted tree was reported in sequence. If not, one of
1270 event parsimony trees was reported. (**F**) Reconciliation-assisted gene tree dating.
1271 Speciation nodes in the dated species tree were mapped onto the non-dated gene tree by
1272 phylogeny reconciliation. By using those nodes as calibration points, the other node
1273 ages were estimated using the penalized likelihood method.
1274

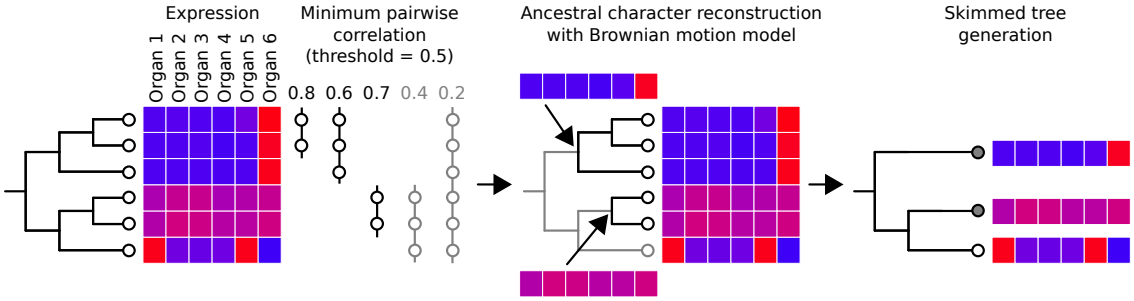


Fig. S14. Gene tree skimming in phylogenetic comparative analysis. Gene tree clades are collapsed if character states are highly correlated. Resultant trees contain a smaller number of leaves than the original trees while preserving drastic changes in character evolution. Note that, unlike this example, we used extremely stringent threshold in the analysis (Pearson's correlation coefficient > 0.99).