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

Kenji Fukushima¹,* and David D. Pollock¹,*

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

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

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

Abstract:
The origins of complex multicellular physiology lie in the evolution of gene expression. Genes express differentially among organs as organisms evolve, but it is not well understood whether expression in certain organs increases the probability that they will be repurposed for expression in other organs. To examine this question, we amalgamated 1,903 RNA-seq datasets from 182 research projects, including 6 organs in 21 vertebrate species. We used various automated quality controls to eliminate project-specific biases, and defined expression regime evolution using tree-based Ornstein–Uhlenbeck models, allowing us to reconstruct evolutionary pathways of gene-family-wise expression on a genome-wide scale. Fluxes in organ-specific gene expression were non-random, suggesting that some ancestral expression patterns strongly preadapted genes for expression in certain organs, while others did not. For example, brain, ovary, and testis tend to change expression from one to another, while kidney and liver form a separate loop of gene expression evolution, illustrating a strong modularity of gene exchange among vertebrate organs. More subtle but significant differences depended on the type of gene duplication. Notably, RNA-based gene duplications tended to generate asymmetric exchange fluxes between organs. This detailed view of the evolutionary dynamics of gene expression supports a major role for preadaptive pathways.
Vertebrate organs serve as functional units of physiological activities, and their identities and functions are determined by the diverse expression patterns of thousands of genes. Because of this, the evolution of gene expression patterns plays a central role in organismal evolution. The degree of organ expression specificity affects how fast amino acids substitute (Zhang and Li, 2004), how rapidly they change expression levels (Liao and Zhang, 2006), and patterns of histone modifications (She et al., 2009). Lineage-defining organs such as the hominoid brain also depend on gene expression shifts (Chen et al., 2013; Kaessmann, 2010; Zhang and Long, 2014; Zhang et al., 2011). However, evolutionary dynamics and transitions among expression patterns remain poorly understood.

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

To allow evolutionary gene expression analysis for a broad set of genes, we developed an extended phylogenetic approach tailored to gene family trees with duplications and losses, and applied it to 21 tetrapod genomes (Fig. S1). Only 1,377 genes per species (28,917 in total) in this dataset were ‘conserved orthologs’, whereas there were 15,282 gene trees in total, comprised of 15,787 genes per species on average (331,530 genes in total), including 21,242 human genes. This approach avoids problems with pairwise analyses (Assis and Bachtrog, 2015; Chen and Zhang, 2012; Guschanski et al., 2017; Kryuchkova-Mostacci et al., 2016; Lan and Pritchard, 2016; Rogozin et al.,
2014; Warnefors and Kaessmann, 2013) that ignore phylogenetic tree structure, and allows inference of expression at ancestral nodes in the tree. Because we use gene phylogenies rather than a single species phylogeny (Brawand et al., 2011), we can analyze gene families with lineage-specific gene duplications and losses, and we did not need to bias our study towards conserved genes and families with slow gene turnovers. Such biases would tend to limit the number of organ-specific genes, which are enriched in lineage-specific and young duplications (Guschanski et al., 2017).

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

We next modeled expression evolution along 15,282 maximum-likelihood gene family phylogenies (Table S4). We used Ornstein–Uhlenbeck models to allow for possible adaptive shifts in optimal expression levels (Bastide et al., 2018; Hansen, 1997; Khabbazian et al., 2016). 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., 2017; Rohlfs et al., 2014). This analysis identifies
points where statistically significant expression ‘regime shifts’ (Hansen, 1997) occur
along each gene tree (Fig. 2A). To evaluate the effect of branching events preceding
expression shifts, we categorized gene tree nodes by whether they were associated with
DNA-based or RNA-based duplication, or not (‘speciation’) (Fig. 2B). Because
regulatory environments and expression patterns are more preserved among orthologous
genes in comparison with paralogous genes (Altenhoff et al., 2012; Castillo-Davis et al.,
2004; Chen and Zhang, 2012; Rogozin et al., 2014), we consider changes following
speciation nodes to be the ‘baseline’ mode of expression evolution for comparison with
duplication events.

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

To summarize gene expression evolution, we mapped shift event totals onto the
species tree (Fig. 3A–B). During vertebrate diversification, expression regimes were
usually conserved within gene groups formed by speciation (26,158 shifts/549,012
branches, 4.8% per branch, at a rate of 0.52 shifts/kiloyear). In contrast, expression
shifts occurred at three times this rate after DNA-based duplication (7,179 shifts/72,162
branches, 9.9% per branch, at a rate of 1.49 shifts/kiloyear). RNA-based duplications
occurred four times faster than DNA duplications (1,176 shifts/3,345 branches 35.2%
per branch, at a rate of 6.04 shifts/kiloyear). This may be explained by the idea that
functional divergence alters expression patterns and enhances long-term retention of
duplicated copies (Assis and Bachtrog, 2015; Huerta-Cepas et al., 2011). Genes that
shift genome location tend to have higher expression shift frequencies than those that do not (Fig. S9), in line with previous observations from the human genome (Lan and Pritchard, 2016). Expression levels varied most in ovary and testes, which had significantly higher average stationary variances than the other four organs (Fig. 3C). This is consistent with the observation of strong sex-related divergence of gene expression in testes (Brawand et al., 2011), but shows strong sex-related divergence in the reproductive organs of both sexes.

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

To further characterize how expression properties change from ancestral to derived regimes, we examined changes in expression specificity, measured by $\tau$ (Yanai et al., 2005). Genes with highly specific expression have $\tau$ values close to one, whereas uniformly expressed genes yield $\tau$ values close to zero. Compared with speciation, expression patterns following RNA-based duplications are more specific, and DNA-based duplications were in between (Fig. 3D). We next examined the maximum expression level among the six organs, because maximum rather than average values are expected to reflect the ‘on state’ transcriptional activity. Duplication-associated shifts tended to have bigger changes in maximum expression level, with a pronounced enrichment of down-regulation (Fig. 3E). Down-regulation related to duplication events may be best explained by either recent nonfunctionalization (Balakirev and Ayala, 2003; Mighell et al., 2000) or specialized expression in unanalyzed organs. Finally, we compared expression complementarity (Huerta-Cepas et al., 2011) to evaluate the differentiation in expression patterns between sister lineages. Duplication events gave rise to greater complementarity in expression more often than speciation events (Fig. 3F), further supporting the role of gene duplication in functional specialization. Changes
in expression properties were more drastic in RNA-based duplications, followed by DNA-based duplications and speciation events, probably reflecting the different modes of inheritance of regulatory environments. RNA-based duplication accompanies the elimination of regulatory elements outside transcribed regions, whereas DNA-based duplication can retain certain regulatory regions depending on the range of duplicated genomic segments (Babushok et al., 2007).

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

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

To eliminate the effect of subtle expression changes, we also analyzed shifts with drastic expression changes by introducing a cutoff of expression specificity ($\tau$>0.67) to define organ-specific genes; with this cutoff, the second-highest expression level is no
greater than two-thirds of the highest expression in both ancestral and descendent states. Although some previously significant trends were not recovered due to small sample sizes, especially in RNA-based duplication, the result was largely consistent with the broader analysis (Fig. S13; Table S5), demonstrating the robustness of our conclusion. Following RNA-based duplication, genes often become expressed in testis (Fig. S11). This result is somewhat concordant with the “out of the testis” hypothesis, which explains the accelerated gains of testis expression based on the permissive chromatin state, abundant transcriptional machinery, relatively simple promoters required for the expression in spermatogenic cells and following gains of new expression patterns (Kaessmann, 2010; Kleene, 2005).

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

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

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

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

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

Transcriptome quantification. Fastq files were extracted from downloaded SRA files using parallel-fastq-dump 0.6.2 (https://github.com/rvalieris/parallel-fastq-dump) with the minimum read length of 25 bp and the quality filter (-E option) (Leinonen et al., 2011). The fastq sequences were then subjected to a quality filtering by fastp 0.12.3 (Chen et al., 2018). The filtered reads were mapped to genomic features annotated as non-messenger RNAs in the Ensembl GTF files using bowtie2 2.3.4 (Langmead and Salzberg, 2012) and resultant unmapped reads were used for expression level quantification using kallisto 0.43.1 with the sequence-based bias correction (Bray et al., 2016). Samples were removed if 20% or smaller percentages of reads were mappable (Fig. S2). Log2-transformed TPM (transcripts per million) values were subjected to downstream analyses.

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

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

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

**Gene tree reconstruction.** Maximum-likelihood trees were reconstructed using IQ-TREE 1.6.5 (Nguyen et al., 2015) with the best-fit nucleotide substitution models selected by ModelFinder with the Bayesian Information Criterion (Kalyaanamoorthy et
Larger orthogroups and longer genes tended to fit more complex substitution matrices and larger numbers of categories of rate heterogeneity (Fig. S14A–B; Table S4). Ultrafast bootstrapping with 1,000 replicates was performed to evaluate the credibility of tree topology (Minh et al., 2013) with a further optimization of each bootstrapping tree (-bnni option) (Hoang et al., 2018).

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

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

**Modeling and shift detection of expression evolution.** Using the dated gene trees and SVA-corrected expression matrices, expression shifts were detected as shifts of optimal trait values in Ornstein–Uhlenbeck (OU) models determined by a Lasso-based model selection with AICc in an R package ‘l1ou’ (Khabbazian et al., 2016) or by an expectation–maximization algorithm with a penalized likelihood in an R package ‘PhylogeneticEM’ (Bastide et al., 2018). To handle gene trees recalcitrant to this
analysis (especially those with a large number of genes), we skimmed gene trees by collapsing clades with small changes in expression level (Fig. S16). Specifically, we first calculated all-vs-all Pearson’s correlation coefficients of gene expression level among all genes that belong the clade. The clade was replaced with a single tip if the minimum correlation coefficient was greater than 0.5. Phylogenetic means of the clade were calculated by assuming Brownian motion and were used as expression level at the new tip. Although the use of PhylogeneticEM and the tree skimming reduced the number of detected regime shifts, the positions of detected shifts were largely overlapped with those inferred with l1ou without tree skimming (Fig. S17). In addition, this strategy allowed us to analyze the largest gene tree with 6,553 genes. Therefore, we used PhylogeneticEM and/or the tree skimming if the gene tree cannot be analyzed by l1ou with non-skimmed topology (Table S4; 8/15,282 trees). The upper limit of regime shifts was set as \( \max \left[ \min(N/2, 100), \sqrt[3]{N} \right] \). In most gene trees (15,278/15,282), the number of detected regime shifts was smaller than the upper limit (Fig. S18).

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

**Estimation of protein evolution rate.** Parameters for codon substitution matrix, shape parameter of discrete gamma distribution for rate heterogeneity (\( \alpha \)), equilibrium transition/transversion rate ratio (\( \kappa \)), equilibrium nonsynonymous/synonymous substitution rate ratio (\( \omega \)) were estimated using IQ-TREE 1.6.5 (Nguyen et al., 2015) by fitting GY+F3X4+G4 models to each gene tree. Equilibrium base composition (\( \theta \)) was estimated from empirical codon state frequencies which are calculated from the alignment by counting. To obtain \( \theta \) at the root node, we calculated \( \theta \) at subroot nodes by
taking advantage of IQ-TREE’s empirical Bayesian method for ancestral sequence
reconstruction. Considering subroot branch length, a weighted average of the subroot $\theta$
values were calculated as the $\theta$ value at the root node. Using all those parameters,
branch-wise nonsynonymous/synonymous substitution rate ratios were estimated by
stochastic substitution mapping ($mapdNdS$) (Guéguen and Duret, 2018) using bio++
library (Guéguen et al., 2013). To examine the robustness of the $mapdNdS$-based $\omega$
estimation, we compared the results with those obtained by maximum-likelihood $\omega$
estimation by fitting MG94W9 models in HyPhy 2.3.11 (Pond et al., 2005). The two
methods yielded a high correlation of $\omega$ values (Spearman’s $\rho = 0.71, N = 622,906$
branches) and consistent results on the effect of branching events and expression shifts
(Fig. S10), suggesting methodological robustness. We reported $mapdNdS$-based results
in the main text.

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

**Analysis of branching events.** Speciation and duplication nodes were classified by
a species-overlap method (Huerta-Cepas et al., 2007) and were mapped to the species
tree on the basis of species coverages. Duplication confidence scores (Vilella et al.,
2009) were calculated for the duplication nodes and scores greater than 0 were
considered duplication events. A transition from intron-containing to intronless states
with a posterior probability greater than 0.5 was classified as a retrotransposition event.

Although our classification cannot detect RNA-based duplication from originally intronless genes, we expect such situations would be rare because most vertebrate genes contain at least one intron (e.g., 20,160/21,242 human genes in our dataset).

Inter-chromosomal translocation was detected by considering chromosomal locations with the highest posterior probability as the ancestral states. Because of the difficulty in determining rooting positions of deep phylogenies, gene tree nodes older than the root node of the species tree were removed from the analysis.

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

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

**Acknowledgements:**

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

**Author Contributions:**

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

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

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

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

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

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

Bastide, P., Ané, C., Robin, S., Mariadassou, M., and Alfaro, M. (2018). Inference of adaptive shifts for multivariate correlated traits. Syst. Biol. 67, 662–680.

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

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

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

Borrowiec, M.L. (2016). AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ 4, e1660.

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

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

Breschi, A., Djebali, S., Gillis, J., Pervouchine, D.D., Dobin, A., Davis, C.A., Gingeras, T.R., and Guigó, R. (2016). Gene-specific patterns of expression variation across
organs and species. Genome Biol. 17, 151.
Brunner, E., and Munzel, U. (2000). The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation. Biometrical J. 42, 17–25.
Budd, G.E. (2007). On the origin and evolution of major morphological characters. Biol. Rev. 81, 609–628.
Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973.
Carelli, F.N., Hayakawa, T., Go, Y., Imai, H., Warnefors, M., and Kaessmann, H. (2016). The life history of retrocopies illuminates the evolution of new mammalian genes. Genome Res. 26, 301–314.
Castillo-Davis, C.I., Hartl, D.L., and Achaz, G. (2004). cis-Regulatory and protein evolution in orthologous and duplicate genes. Genome Res. 14, 1530–1536.
Chen, X., and Zhang, J. (2012). The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. PLoS Comput. Biol. 8, e1002784.
Chen, J., Swofford, R., Johnson, J., Cummings, B.B., Rogel, N., Lindblad-Toh, K., Haerty, W., Palma, F. di, and Regev, A. (2017). A quantitative model for characterizing the evolutionary history of mammalian gene expression. BioRxiv 229096.
Chen, K., Durand, D., and Farach-Colton, M. (2000). NOTUNG: a program for dating gene duplications and optimizing gene family trees. J. Comput. Biol. 7, 429–447.
Chen, S., Krinsky, B.H., and Long, M. (2013). New genes as drivers of phenotypic evolution. Nat. Rev. Genet. 14, 645–660.
Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. BioRxiv 274100.
Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P.D., Grützner, F., and Kaessmann, H. (2014). Origins and functional evolution of Y chromosomes across mammals. Nature 508, 488–493.
Danshina, P. V., Geyer, C.B., Dai, Q., Goulding, E.H., Willis, W.D., Kitto, G.B., McCarrey, J.R., Eddy, E.M., and O’Brien, D.A. (2010). Phosphoglycerate kinase 2 (PGK2) is essential for sperm function and male fertility in mice. Biol. Reprod. 82, 136–145.
de-Leon, S.B.-T., and Davidson, E.H. (2007). Gene regulation: gene control network in development. Annu. Rev. Biophys. Biomol. Struct. 36, 191–212.

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

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

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

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

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

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

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

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

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

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

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

Huerta-Cepas, J., Dopazo, J., Huyinen, M.A., and Gabaldón, T. (2011). Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. Brief. Bioinform. 12, 442–448.
Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. Mol. Biol. Evol. 33, 1635–1638.

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

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

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

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

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

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

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

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

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

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

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

Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 3, e161.
Leinonen, R., Sugawara, H., and Shumway, M. (2011). The Sequence Read Archive. 
Nucleic Acids Res. 39, D19–D21.

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

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

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

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

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

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

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

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

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

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

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

Popescu, A.-A., Huber, K.T., and Paradis, E. (2012). ape 3.0: New tools for 
distance-based phylogenetics and evolutionary analysis in R. Bioinformatics 28, 
1536–1537.
Potrzebowski, L., Vinckenbosch, N., Marques, A.C., Chalmel, F., Jégou, B., and Kaessmann, H. (2008). Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biol. 6, e80.

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

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

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

Rohlf, R. V., Harrigan, P., and Nielsen, R. (2014). Modeling gene expression evolution with an extended Ornstein–Uhlenbeck process accounting for within-species variation. Mol. Biol. Evol. 31, 201–211.

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

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

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

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

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

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

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

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in
vertebrates. Genome Res. 19, 327–335.

Warnefors, M., and Kaessmann, H. (2013). Evolution of the correlation between expression divergence and protein divergence in mammals. Genome Biol. Evol. 5, 1324–1335.

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

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

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

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

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

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

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

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

Zhang, Y.E., Landback, P., Vibranovski, M.D., Long, M., and Stark, A. (2011). Accelerated recruitment of new brain development genes into the human genome. PLoS Biol. 9, e1001179.
Figure 1. Transcriptome corrections to amalgamate heterogeneous RNA-seq experiments. (A) A simplified flow chart of the transcriptome curation. The full chart is available in Fig. S1. (B–D) Transcriptome curation within species. Data from Monodelphis domestica are shown as an example. The heatmaps show Pearson’s correlation coefficients among RNA-seq samples (B). Pearson’s correlation coefficients were calculated using expression levels of all genes. Note that project-specific correlations visible in the uncorrected data are absent in the corrected data. The box plots show distinct distributions of Pearson’s correlation coefficients depending on whether a pair of samples are the same organ or whether they are from the same research project (C, bw, between categories; wi, within category). The correlation coefficients are largely improved in within-organ comparisons when surrogate variables are removed, while within-project biases are attenuated. In this species, 7 surrogate variables were removed.
variables were detected against 52 RNA-seq data from 8 projects (D). Analysis of those variables by linear regression highlights the BioProject feature as the strongest source of removed biases. For full description of predictors, see Fig. S7. (E) Nonlinear dimensionality reduction using expression levels of 1,377 conserved orthologs from 21 species. Each point corresponds to the averaged transcriptome of one organ in one species. The six organs are colored in the same way as in B. Species are identified by border colors (see Fig. S8).
Figure 2. Expression evolution in a complex history of gene family evolution. (A) A phylogenetic simulation of univariate Ornstein–Uhlenbeck process. Regime shifts (indicated by colors) appear as a substantial change in optimal trait values. (B) Branching events on gene trees were categorized into speciation, DNA duplication and RNA duplication. (C) The gene tree of phosphoglycerol kinases (orthogroup ID: OG0002332) is shown as an example. This orthogroup shows ortholog-specific expression patterns as well as regime shifts after speciation or lineage-specific gene duplication. Branch colors indicate expression regimes. Node colors mean diversification events: blue, speciation; red, DNA duplication; orange, RNA duplication. Bootstrap supports are given above branches. The heatmap shows expression levels in, from the left, brain, heart, kidney, liver, ovary, and testis. To the right, the number of introns and located chromosomes (A, autosome; X, X chromosome; Y, Y chromosome) are indicated.
Figure 3. Characteristics of expression shifts in 15,282 gene trees. (A) The species tree showing analyzed genomes. (B) Mapping of 34,513 expression shifts in the species tree. The number and proportion of expression shifts in S, D, and R branches (see Fig. 2B) are shown. Corresponding branches in the species tree are indicated in A. Note that speciation only occurs at the left node of a branch in the species tree, whereas duplication can occur anywhere in the branch. (C) Organ-specific stationary variances of expression level evolution in vertebrates. (D–F) Cumulative frequency of change in expression specificity (D), change in maximum expression level (E), and expression complementarity between sister lineages (F) among detected expression shifts.
Fig. 4. **Evolutionary dynamics of gene expression.** (A–B) Transitions of primary-expressed organs (PEOs). Arrows represent transitions from ancestral PEOs to derived PEOs. Networks show the rank with the number of transitions (A) and its statistical significance based on 10,000 permutations (B). In A, the patterns of transitions among the three branching categories were statistically compared by a $\chi^2$ test.
Supplementary Materials for

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

Kenji Fukushima and David D. Pollock

List of Supplementary Materials:

- Tables S1–S5 (separate file)
- Figs. S1–S18
- Supplementary Data (separate file)
Fig. S1. A flow-chart of transcriptome amalgamation, gene tree inference, and evolutionary characterization in this study.
Fig. S2. Changes in number of reads, total nucleotide length, mean read length, and mapping rate by RNA-seq read filterings. Mapping rates were, in almost every case, improved by both read quality filtering and read filtering by mapping to miscellaneous genomic features.
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. The number of RNA-seq experiments, BioProjects, and detected surrogate variables in each species. The counts were derived from the final dataset. The numbers of surrogate variables are correlated with the numbers of RNA-seq experiments and BioProjects.
Fig. S5. Correlation heatmaps of corrected transcriptomes. See Supplementary Data for full descriptions including RNA-seq sample IDs and BioProject IDs.
Fig. S6. Distinct distributions of Pearson’s correlation coefficients depending on whether a pair of RNA-seq samples are the same organ or whether they are from the same research project. bw, between categories; wi, within category.
Fig. S7. Predictor analysis of detected surrogate variables. The predictive power was analyzed by linear regression using different properties of RNA-seq experiments: organ (brain, heart, kidney, liver, ovary, and testis), BioProject (e.g., PRJNA176589), library selection (e.g., cDNA and polyA), library layout (single and paired), instrument (e.g., Illumina HiSeq 2500 and NextSeq 550), number of read (e.g., 91,641,467 reads), % lost, fastp (percentage of reads that are removed by fastp; e.g., 5%), % lost, misc feature (percentage of reads that are mapped to non-nuclear-mRNA features and are removed from the analysis; e.g., 5%), minimum read length (e.g., 25 nt), average read length (e.g., 70 nt), maximum read length (e.g., 75 nt), and mapping rate (e.g., 80%). The predictors are summarized in Table S1.
Fig. S8. Multispecies correlation analysis of averaged organ expression. Corrected expression levels of 1,377 conserved orthologs were used to calculate pairwise Pearson’s correlation coefficients.
Fig. S9. The relationships between expression shifts and chromosomal location.
The heatmap shows the frequency of expression shifts observed among the branches with or without a change in the chromosomal category (non-diagonal or diagonal, respectively). Chromosomal locations were categorized into autosomes (A), X chromosome (X) and Y chromosome (Y), and the ancestral locations were inferred by stochastic character mapping.
Fig. S10. Acceleration of protein evolution rate in correlation with expression regime shifts. Results from two different approaches are shown: stochastic substitution mapping (mapdNdS) and maximum-likelihood estimation (HyPhy). A plus (+) indicates branches with expression shifts, whereas minus (–) branches are sisters to the ‘plus’ branches. Statistical differences between pairs of distributions were tested using a two-sided Brunner–Munzel test (Brunner and Munzel, 2000). Non-log-transformed median values are shown above the boxplots. For visualization purpose, extreme values were clipped (minimum = –10; maximum = 10).
Fig. S11. Frequency of organs in which highest expression level is observed before and after expression shifts. The stacked bar plot was generated from regime shifts involving altered primary-expressed organs. The numbers of regime shifts are shown in Fig. 4. S, speciation; D, DNA-based duplication; R, RNA-based duplication.
Fig. S12. The global polarity of transitions of primary-expressed organs. The global polarity is defined by the sum of asymmetry between two primary-expressed organs. Boxplots show 95% confidence intervals estimated by 10,000 bootstrap resamplings. The numbers of regime shifts are shown in Fig. 4. S, speciation; D, DNA-based duplication; R, RNA-based duplication.
**Fig. S13. Evolutionary dynamics of specifically expressed genes.** Expression shifts with high ancestral and derived specificity ($\tau > 0.67$) are analyzed. The panels A, B, C, D correspond to Fig. 4A, Fig. 4B, Fig. S11, Fig. S12, respectively.
Fig. S14. Best-fit nucleotide substitution models, rooting methods, and constrained nodes for divergence time estimation in gene tree reconstructions. (A) Complexity of best-fit nucleotide substitution matrices. (B) Complexity of rate heterogeneity among nucleotide sites. Both discrete Gamma models (Yang, 1994) and FreeRate models (Soubrier et al., 2012; Yang, 1995) were included to count the number of categories for rate heterogeneity. (C) Selected rooting positions in reconciliation-assisted gene tree rooting. MAD, minimal ancestor deviation; NTG, ‘rooting mode’ of NOTUNG; MID, midpoint between the longest path. (D) Time-constrained nodes in tree dating. R, root node; S, speciation node; D, duplication node. All available constraints (RDS) are used in the first trial and then successively relaxed if estimation fails. In the category ‘allS’, all nodes are speciation nodes and therefore no divergence time estimation was performed for the trees.
Fig. S15. Use of phylogeny reconciliation for gene tree rooting and dating. (A) Reconciliation-assisted gene tree rooting. Rooting points were estimated with two different methods: the minimum ancestor deviation (MAD) method and midpoint rooting. If they were compatible with the event parsimony involving gene duplication and loss, the MAD- or midpoint-rooted tree was reported in sequence. If not, one of event parsimony trees was reported. (B) Reconciliation-assisted gene tree dating. Speciation nodes in the dated species tree were mapped onto the non-dated gene tree by phylogeny reconciliation. By using those nodes as calibration points, the other node ages were estimated using the penalized likelihood method.
Fig. S16. Gene tree skimming. Gene tree clades are collapsed if character states are highly correlated. Resultant trees contain a smaller number of leaves than the original trees while preserving drastic changes in character evolution.
**Fig. S17. The UpSet plot of four methods for expression shift detection.** The overlap of expression shift branches is shown among those detected by l1ou or PhylogeneticEM in combination with the original trees or skimmed trees (Fig. S16). Results are shown for expression shift detection that was successfully completed on 3,862 orthogroups.
**Fig. S18.** The number of detected regime shifts in gene expression evolution. Each point corresponds to one gene tree. The red line shows the upper limits in the regime shift search.