ARTICLE

doi:10.1038/nature10532

# The evolution of gene expression levels in mammalian organs

David Brawand[1,2]*, Magali Soumillon[1,2]*, Anamaria Necsulea[1,2]*, Philippe Julien[1,2], Gábor Csárdi[2,3], Patrick Harrigan[4], Manuela Weier[1], Angélica Liechti[1], Ayinuer Aximu-Petri[5], Martin Kircher[5], Frank W. Albert[5]†, Ulrich Zeller[6], Philipp Khaitovich[7], Frank Grützner[8], Sven Bergmann[2,3], Rasmus Nielsen[4,9], Svante Pääbo[5] & Henrik Kaessmann[1,2]

**Changes in gene expression are thought to underlie many of the phenotypic differences between species. However, large-scale analyses of gene expression evolution were until recently prevented by technological limitations. Here we report the sequencing of polyadenylated RNA from six organs across ten species that represent all major mammalian lineages (placentals, marsupials and monotremes) and birds (the evolutionary outgroup), with the goal of understanding the dynamics of mammalian transcriptome evolution. We show that the rate of gene expression evolution varies among organs, lineages and chromosomes, owing to differences in selective pressures: transcriptome change was slow in nervous tissues and rapid in testes, slower in rodents than in apes and monotremes, and rapid for the X chromosome right after its formation. Although gene expression evolution in mammals was strongly shaped by purifying selection, we identify numerous potentially selectively driven expression switches, which occurred at different rates across lineages and tissues and which probably contributed to the specific organ biology of various mammals.**

Shared mammalian traits include lactation, hair and relatively large brains with unique structures[1]. In addition to these traits, individual lineages have evolved distinct anatomical, physiological and behavioural characteristics relating to differences in reproduction, life span, cognitive abilities and disease susceptibility. The molecular changes underlying these phenotypic shifts and the associated selective pressures have begun to be investigated using available mammalian genomes[2], the number of which is rapidly increasing. However, although genome analyses may uncover protein-coding changes that potentially underlie phenotypic alterations, regulatory mutations affecting gene expression probably explain many or even most phenotypic differences between species[3].

Until recently, comparisons of mammalian transcriptomes were essentially restricted to closely related primates[4–8] or mice[5], although human–mouse comparisons using microarrays were also attempted[9]. Nevertheless, microarrays require hybridization to species-specific probes, making between-species comparisons of transcript abundance difficult[6]. The development of RNA sequencing (RNA-seq) protocols now allows for accurate and sensitive assessments of expression levels[10]. The power of RNA-seq for transcriptome assessment was recently demonstrated for human individuals[11,12] and closely related primates[13,14].

## RNA-seq and genome reannotation

To study mammalian transcriptome evolution at high resolution, we generated RNA-seq data (~3.2 billion Illumina Genome Analyser IIx reads of 76 base pairs) for the polyadenylated RNA fraction of brain (cerebral cortex or whole brain without cerebellum), cerebellum, heart, kidney, liver and testis (usually from one male and one female per somatic tissue, and two males for testis) from nine mammalian species (Supplementary Tables 1 and 2, Methods and Supplementary Note): placental mammals (great apes, including humans; rhesus macaque; and mouse), marsupials (grey short-tailed opossum) and monotremes (platypus). Corresponding data (~0.3 billion reads) were generated for a bird (red jungle fowl, a non-domesticated chicken) and used as an evolutionary outgroup.

We refined existing Ensembl[15] genome annotations by performing an initial read mapping to detect transcribed regions and splice junctions (Methods and Supplementary Note), which resulted in modified boundaries for ~31,000–44,500 exons and the addition of 20,000–34,500 new exons and 66,000–125,000 new splice junctions to known protein-coding genes (Supplementary Note Tables 4 and 5). We also searched *de novo* for multi-exonic transcribed loci; our results validated most Ensembl-annotated protein-coding genes, pseudogenes and long non-coding RNA genes (Supplementary Note Table 11), but we also detected thousands of multi-exonic transcribed loci (possibly representing protein-coding or non-coding RNA genes) in previously unannotated regions (Supplementary Note Table 10).

Newly detected exons are transcribed at lower levels and are significantly less conserved, at the sequence level, than Ensembl-annotated exons (two-tailed $P < 10^{-8}$, Mann–Whitney *U*-test; Supplementary Fig. 1). However, the sequence conservation level is higher for new exons than for flanking introns, with visible peaks around splice sites, indicating that many of these exon sequences are preserved by purifying selection[16].

Depending on the species, 11–30% of the total genomic length is covered by unambiguously mapped RNA-seq reads (Table 1). Much of the covered length is explained by retained introns, but substantial coverage is also found outside annotated regions (Table 1). Our data suggest that large proportions (>34–61%) of amniote (that is, mammal and bird) genomes are transcribed, consistent with previous work[17].

[1]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland. [2]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland. [3]Department of Medical Genetics, University of Lausanne, 1005 Lausanne, Switzerland. [4]Department of Integrative Biology, University of California, Berkeley, California 94720, USA. [5]Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany. [6]Chair of Systematic Zoology, Humboldt-University, 10099 Berlin, Germany. [7]CAS-MPG Partner Institute for Computational Biology, 200031 Shanghai, China. [8]The Robinson Institute, School of Molecular and Biomedical Science, University of Adelaide, Adelaide, South Australia 5005, Australia. [9]The Bioinformatics Center, University of Copenhagen, 2200 Copenhagen, Denmark. †Present address: Lewis Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544, USA.
*These authors contributed equally to this work.

**Table 1 | Assessment of the transcribed fraction of amniote genomes based on RNA-seq**

| Species | Ensembl genes (Mb) | | Intergenic multi-exonic transcribed loci (Mb) | | Other intergenic islands (Mb) | Total covered‡ (Mb) | Total transcribed§ (Mb) | No. reads‖ ($10^6$) |
|---|---|---|---|---|---|---|---|---|
| | Exons* | Introns† | Exons* | Introns† | | | | |
| Human | 80 (84%) | 410 (29%) | 3.5 (100%) | 12 (15%) | 121 | 627 (20%) | 1,685 (54%) | 255 |
| Chimp | 58 (93%) | 404 (36%) | 8.4 (100%) | 30 (20%) | 185 | 685 (19%) | 1,508 (43%) | 199 |
| Gorilla | 53 (90%) | 297 (32%) | 7.6 (100%) | 30 (19%) | 179 | 567 (19%) | 1,314 (43%) | 163 |
| Orangutan | 43 (89%) | 260 (27%) | 6.2 (100%) | 22 (20%) | 144 | 475 (14%) | 1,259 (37%) | 131 |
| Macaque | 52 (92%) | 365 (35%) | 9.0 (100%) | 35 (22%) | 202 | 663 (21%) | 1,455 (47%) | 156 |
| Mouse | 71 (88%) | 433 (42%) | 5.1 (100%) | 18 (21%) | 127 | 654 (24%) | 1,311 (48%) | 278 |
| Opossum | 50 (92%) | 263 (23%) | 6.7 (100%) | 23 (16%) | 156 | 499 (14%) | 1,499 (42%) | 158 |
| Platypus | 31 (89%) | 72 (16%) | 10.0 (100%) | 16 (13%) | 107 | 234 (11%) | 712 (34%) | 137 |
| Chicken | 40 (94%) | 177 (38%) | 6.0 (100%) | 17 (26%) | 95 | 336 (30%) | 676 (61%) | 146 |

* Exonic length covered by unambiguous RNA-seq reads (percentage of the total exonic length of expressed genes).
† Intronic length covered by unambiguous RNA-seq reads (percentage of the total intronic length of expressed genes).
‡ Total length covered by unambiguous RNA-seq reads (percentage of the total genomic length).
§ Estimation of the total transcribed length in our data set (percentage of the total genomic length). This estimation includes all exonic and intergenic regions covered by unambiguous RNA-seq reads, as well as the intronic length of intergenic multi-exonic transcribed loci and of Ensembl-annotated genes detected as expressed in our samples.
‖ Number of unambiguously mapping reads that were used to compute the read coverage statistics.
Mb, megabase.

On the basis of the refined genome annotations, we remapped our RNA-seq reads and resolved read mapping ambiguities (Methods). In this Article, we focus on comparative analyses of expression levels of protein-coding genes. For comparisons among all ten amniote species, we used a set of 5,636 one-to-one (1:1) orthologues (Methods). A corresponding set of 13,277 1:1 orthologues was used for the six primates. Expression values were normalized to render the data comparable across species (Methods).

## Mammalian gene expression phylogenies

To obtain an initial overview of gene expression patterns, we performed a principal-component analysis, which clearly separates the data according to tissue (only the neural tissues do not perfectly separate), although a substantial part of the variance is also explained by differences among lineages (Fig. 1a).

To reconstruct global evolutionary trends in more detail, we built expression distance matrices for each tissue (Methods) and reconstructed gene expression trees (Fig. 1b and Supplementary Figs 2 and 3). The trees are highly consistent with the known mammalian phylogeny: they correctly resolve the three major mammalian lineages (placentals, or eutherians; marsupials; and monotremes), separate the two eutherian lineages (primates and rodents) and group humans and the other great apes to the exclusion of the macaque (an Old World monkey). This suggests that regulatory changes accumulate over evolutionary time, such that closely related species have more similar expression levels. Our results are remarkable given the within-species variation (including sometimes substantial sex-biased gene expression; Supplementary Note, Supplementary Table 3 and Supplementary Fig. 4) and the fact that age, feeding status and other characteristics of the individuals could not be perfectly matched between species, for practical and biological reasons. Thus, evolutionary signals inherent in our data, which may reflect changes in cellular gene expression levels or changes in the cellular composition of organs between species (Supplementary Note), outweigh the gene expression variability resulting from sampling differences.

However, branching patterns within the great ape clade do not always reflect the known phylogeny (Supplementary Fig. 2), in particular for human, chimpanzee–bonobo and gorilla, which diverged only ~5–7 million years ago[18]. Bootstrapping analyses show that the branching order of these four species cannot be robustly established for the somatic tissues on the basis of the 5,636 amniote 1:1 orthologues (bootstrap values <0.9; Supplementary Fig. 2).

To resolve great ape gene expression relationships, we built expression phylogenies based on the 13,277 primate 1:1 orthologues (Supplementary Fig. 5), which robustly resolve the great ape clade (that is, bootstrap values >0.9, except for brain) and reveal surprising patterns. In half of the expression trees (testis, heart and brain), humans and gorillas group together, whereas chimpanzees and bonobos

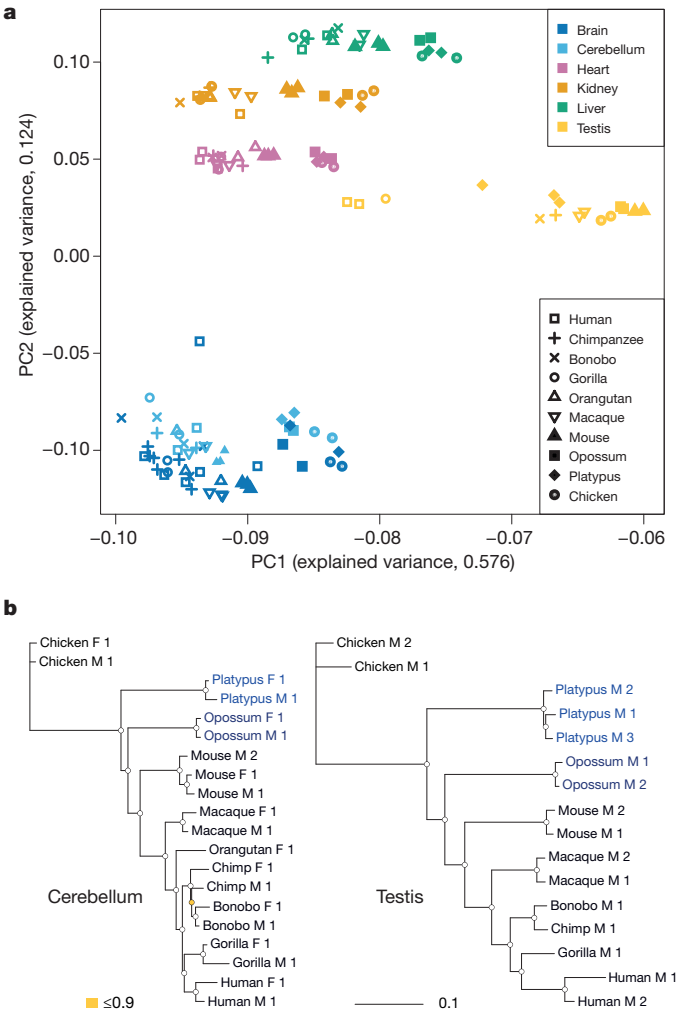


**Figure 1 | Global patterns of gene expression differences among mammals.** **a**, Factorial map of the principal-component analysis of messenger RNA expression levels. The proportion of the variance explained by the principal components is indicated in parentheses. **b**, Mammalian gene expression phylogenies. Neighbour-joining trees based on pairwise distance matrices ($1 - \rho$, Spearman's correlation coefficient) for cerebellum and testis (see Supplementary Fig. 2 for all six organs). Bootstrap values (5,636 1:1 orthologous amniote genes were randomly sampled with replacement 1,000 times) are indicated by circles: white, >0.9; yellow, ≤0.9. Species colour codes: platypus, light blue; opossum, dark blue; eutherians (mice and primates), black.

fall outside of this clade. These two species always group together, as expected given their recent divergence[19], although chimpanzees and bonobos are not always monophyletic. The testis tree groups human with gorilla (bootstrap value of 1; Supplementary Fig. 5), consistent with the evolution of male physiology and mating patterns among the African apes: the highly promiscuous chimpanzees and bonobos evolved larger testicles relative to body size and higher sperm production rates than did the less promiscuous humans and gorillas[20]. Whereas the kidney and cerebellum trees are consistent with the known species phylogeny, the liver tree has an interesting pattern: humans fall outside a clade comprising the other great apes, within which gorillas and orangutans group together. Given the role of the liver in metabolic control and detoxification, these patterns may reflect dietary variation among the great apes, although they may also reflect feeding status patterns of the individuals sampled.

## Rates of expression change in lineages and organs

The branch lengths from the common ancestor of all species to the tips of the tree are remarkably similar (Supplementary Fig. 2), suggesting that gene expression evolution has proceeded at comparable rates in different mammalian lineages. However, the branches leading to mouse are significantly shorter in several tissues, particularly in comparison with those leading to great apes and monotremes (Bonferroni-corrected two-tailed $P < 0.05$ in four or three of the six tissues, respectively; randomization test), in spite of the high rodent DNA mutation rates[21] (Supplementary Fig. 6; see Supplementary Fig. 7 for results that remove within-species variation differences). This is consistent with the strong purifying selection affecting the rodent lineage because of large long-term effective population sizes[22,23]. Our observations, confirmed by another phylogenetic method (Supplementary Note), agree with previous inferences from gene expression studies[24] and protein sequence evolution[22,23] and lend support to previous models of gene expression evolution[6], which assign a dominant role to purifying selection.

The total branch lengths of the trees vary widely among tissues (Fig. 2a). The two neural tissues evolve significantly more slowly than the other organs in both amniotes as a whole and primates (Bonferroni-corrected two-tailed $P < 0.001$; randomization test), suggesting that they may have experienced stronger purifying selection and/or less positive selection than other tissues during mammalian evolution. This observation is remarkable in view of the substantial changes in the size, structure and cellular composition of the brain that occurred during mammalian evolution[25], but is consistent with previous findings[6,26,27] which suggested that nervous tissues may have more fine-tuned expression networks than other organs.

Liver, heart and kidney show similar rates of gene expression change in amniotes (randomization test not significant, $P > 0.1$), whereas in primates kidney evolves significantly more slowly than heart and liver ($P < 0.05$; Fig. 2a). Notably, the testis, previously shown to evolve rapidly both at the phenotypic[20] and molecular levels[6,28], potentially owing to positive selection associated with sperm competition and other sex-related evolutionary forces[29], is the most rapidly evolving tissue for both data sets ($P < 0.001$).

Pairwise species comparisons confirm that gene expression divergence overall increases with evolutionary time (Fig. 2b), consistent with the expression phylogeny results (see above). However, for most tissues, expression levels are approximately as similar between human and chicken as they are between human and platypus, although the bird lineage diverged ~110 million years before the separation of monotremes and therian mammals (that is, eutherians and marsupials). This suggests that the conservation of core organ functions restricts transcriptome divergence.

## Gene expression evolution on the X chromosome

Next we investigated the rate of gene expression change on the different types of chromosome. Sex chromosomes of therians are derived

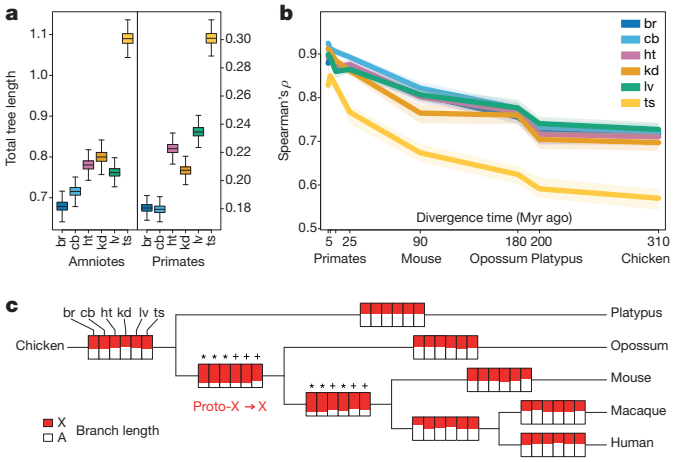


**Figure 2 | Expression divergence rates across tissues and chromosomes.** **a**, Comparisons of total branch lengths of expression trees among the six tissues (br, brain; cb, cerebellum; ht, heart; kd, kidney; lv, liver; ts, testis), for the all-amniote and primate data sets. Errors, 95% confidence intervals based on bootstrapping analysis (1,000 replicates, with one individual per species sampled in each replicate). **b**, Spearman's correlations between humans and the other species. Coloured envelopes show ranges of values obtained in 100 bootstrap replicates. **c**, Expression evolution rates on therian X chromosome versus autosomes. Rectangles reflect median branch lengths (1,000 bootstrap replicates) in X-chromosome expression trees (102 1:1 orthologues located in the X-chromosome conserved region[34]; red) relative to those in autosome trees (5,494 autosomal orthologues; white). *P* values are based on bootstrap replicates: an asterisk indicates two-tailed $P < 0.05$ (that is, branch longer in X-chromosome tree in more than 97.5% of replicates) and a plus sign indicates $P < 0.1$.

from the same ancestral autosomes[30,31], whereas the multiple X and Y chromosomes of the monotremes are distinct and partly homologous to the sex chromosomes of birds[30,32]. To test whether gene expression evolution on the therian X chromosome accelerated after sex chromosome differentiation[33], we compared rates of expression change for genes that are X-linked in both eutherians and marsupials (that is, genes in the X-chromosome conserved region[34]) and autosomal genes, on the basis of branch lengths in expression trees reconstructed for the two types of chromosome (Fig. 2c).

This analysis suggests that gene expression evolution was faster on the X chromosome than on autosomes in the common ancestor of therian mammals (two-tailed $P < 0.05$ for brain, cerebellum and heart; $P < 0.1$ for kidney, liver and testis; randomization test), which corresponds to the time when the original proto-XY chromosomes evolved into sex chromosomes, and remained accelerated in the common eutherian ancestor (two-tailed $P < 0.05$ for brain, cerebellum and kidney; $P < 0.1$ for heart, liver and testis). By contrast, the rate of X-chromosome expression evolution was similar to that of autosomes more recently, as reflected by the terminal eutherian branches ($P > 0.1$ for all tissues and branches), consistent with our hypothesis that gene expression evolution proceeded at a higher rate only on the newly formed X chromosome.

The observed pattern is unlikely to reflect new general properties of the X chromosome as a sex chromosome (for example its reduced effective population size or reduced recombination rates), as such a property would lead to accelerated evolution on all branches following X-chromosome origination. It may instead reflect an increased rate of functional adaptation on the newly formed X chromosome, potentially driven by sex-related selective forces that started to shape this chromosome after sex chromosome differentiation[35], and/or selective pressures associated with the X-chromosome dosage reduction in males resulting from Y-chromosome degradation[36] (see also below). In this context, it is noteworthy that the rate of protein sequence change[37] (except for X-linked genes with Y-chromosome counterparts[38]) and

the rate of fixation of new genes on the X chromosome[33] seem to have increased after differentiation of the sex chromosome. Thus, similarly to *Drosophila*[39], early X-chromosome evolution in mammals seems to be characterized by increased rates of functional adaptation of genes.

## Modular gene expression change

Given that genes commonly function together, concerted expression changes of distinct sets of genes may often be phenotypically relevant. To identify such expression shifts, we identified groups of genes that have coherent expression patterns over subsets of samples[40] (Supplementary Note). These 'modules' were screened for statistically significant enrichments of functional categories.

Among a total of 639 modules in the all-amniote data and 197 modules in the primate-specific data set (Supplementary Tables 4–7; see also the searchable database with comprehensive module details at http://www.unil.ch/cbg/ISA/species), there are 33 organ-specific modules with conserved expression levels among species (17 for amniotes and 16 for primates), 145 modules specific to an organ (or organ pair; see below) with distinct lineage-specific expression patterns (124 for amniotes and 21 for primates), and 658 modules that show no clear relation to specific phylogenetic groups and/or affect multiple organs (498 for amniotes and 160 for primates) (Supplementary Tables 8 and 9).

The 33 organ-specific conserved modules are enriched with genes involved in typical processes (for example synaptic transmission for brain; Benjamini–Hochberg corrected $P < 0.05$), and thus define common primate/mammalian organ functions (Supplementary Tables 8–10).

The 145 organ-specific modules with lineage-specific expression patterns provide clues to the organ biology of different mammals. For example, the all-amniote data reveal 25 nervous tissue modules that evolved distinct expression levels along the major terminal branches of the mammalian phylogeny (Fig. 3a and Supplementary Tables 8 and 10). Notably, modules specific to the central nervous system in the non-primate mammals often (14 of 16 cases) show altered expression in both brain (or cerebral cortex) and cerebellum (Supplementary Table 8), suggesting a tight functional and evolutionary link between them in mammals. Similarly, modular expression changes in kidney and liver often (14 of 28 cases) affected both of these organs, which may reflect their close functional interactions regarding detoxification and waste excretion. The only terminal lineages with distinct testis modules are primates and monotremes (Supplementary Tables 8 and 10).

Among the 32 modular gene expression changes that occurred along the internal branches of the mammalian phylogeny, eight modules in brain, cerebellum and/or testis are highly enriched with X-linked genes (Benjamini–Hochberg corrected $P < 0.05$) and became strongly down-regulated along the common therian or eutherian branch during sex chromosome differentiation (Supplementary Tables 8 and 10), consistent with observations that reduced gene dosage on the newly evolved therian X chromosome in males was not compensated for by global transcriptional upregulation of X-linked genes (P. Julien *et al.*, submitted manuscript, and ref. 41).

Modular expression changes between mammals and chicken occurred only in the neural tissues and in kidney and liver (Supplementary Tables 8 and 10). Four of these modules are significantly enriched with X-linked genes (Benjamini–Hochberg corrected $P < 0.01$). Our results suggest that the early evolution of the mammalian brain was strongly associated with X-chromosome expression changes, perhaps because of an overrepresentation of proto-X-linked brain genes[42].

The only lineage with brain-specific (that is, prefrontal cortex; Supplementary Note) expression modules in the primate data set is that of humans (Supplementary Tables 9 and 10). The 259 genes in the largest of the four human-specific brain modules (ID #p173 in Supplementary Table 10) are involved in various neurological processes, several of which (for example cell adhesion molecules; Benjamini–Hochberg corrected $P < 0.05$) were previously found to be enriched in analyses of regulatory sequence differences between humans and chimpanzees[43]. Notably, the large number of gene ontology categories (12 of 39) related to neuron insulation probably reflects the larger proportion of myelinated axons (white matter) in the human prefrontal cortex than in that of other primates, implying an increased connectivity of this region with other cortical areas[44].

## Expression shifts of individual genes

To detect biologically relevant expression changes of individual genes, we developed a maximum-likelihood framework for modelling gene expression evolution along a phylogeny. We compared several models that incorporate selection and genetic drift and take into account within-species variations and measurement errors (Supplementary Note). To detect relevant lineage-specific changes, we compared a model that assumes a single optimum expression level for a given gene in all branches of the phylogeny with a model in which this gene evolved a different expression optimum in a specified lineage.

Using this approach, we identified 9,255 significant expression changes (Benjamini–Hochberg corrected two-tailed $P < 0.05$; log-likelihood ratio tests; total number of tests, 577,105; Table 2 and Supplementary Tables 11–26). Notably, 2,452 (~63%) of 3,909 tested

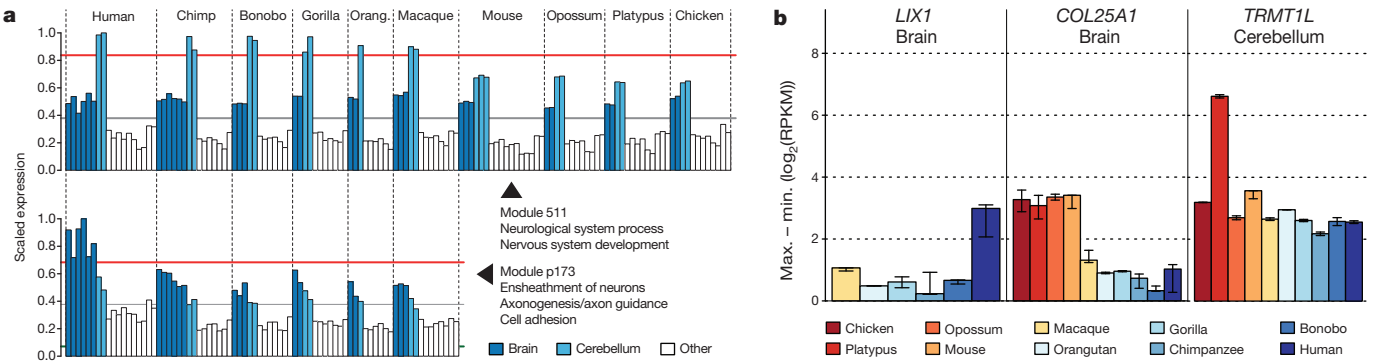


**Figure 3 | Lineage-specific expression shifts of transcription modules and individual genes.** **a**, Modules with specific expression states in human brain (prefrontal cortex; 259 genes) and primate cerebellum (189 genes) are shown. Bars represent the weighted average expression of all genes in a module, for each sample (horizontal grey line indicates average bar height). The horizontal red line represents the cut-off of the biclustering algorithm; samples above the red line are considered to have a distinct expression state. See Supplementary Note and our searchable database (http://www.unil.ch/cbg/ISA/species) for details. **b**, Examples of genes that evolved new optimal expression levels in human prefrontal cortex (*LIX1*; ENSG00000145721), primate cortex (*COL25A1*; ENSG00000188517) and platypus cerebellum (*TRMT1L*; ENSG00000121486). Expression levels are indicated as $\log_2$-transformed RPKM (reads per kilobase of exon model per million mapped reads) (see Supplementary Tables 11–26 for details). Errors, range of expression values for the different individuals for a given species or tissue.

**Table 2 | Numbers of genes with significant lineage-specific expression switches***

| Gene set | Lineage | Brain | Cerebellum | Heart | Kidney | Liver | Testis |
|---|---|---|---|---|---|---|---|
| Primate orthologues | Human | 4 | 303 | 13 | 253 | 93 | 628 |
| | Chimp–bonobo | 11 | 16 | 4 | 7 | 1 | 202 |
| | Chimp | 0 | 2 | 0 | 4 | 0 | 25 |
| | Bonobo | 0 | 2 | 0 | 8 | 11 | 29 |
| | Gorilla | 18 | 73 | 140 | 62 | 37 | 227 |
| | Orangutan | 91 | 159 | 73 | 95 | 52 | — |
| | Human–chimp–bonobo | 0 | 1 | 0 | 0 | 0 | 4 |
| | African apes | 0 | 46 | 0 | 0 | 0 | 554 |
| | Great apes† (macaque) | 30 | 339 | 18 | 215 | 232 | 556 |
| All-amniote orthologues | Primates | 141 | 5 | 2 | 3 | 3 | 3 |
| | Mouse | 33 | 81 | 96 | 154 | 90 | 205 |
| | Opossum | 90 | 195 | 100 | 154 | 113 | 222 |
| | Platypus | 323 | 358 | 264 | 280 | 207 | 326 |
| | Eutherians | 0 | 5 | 11 | 5 | 2 | 0 |
| | Therians | 54 | 50 | 31 | 36 | 19 | 59 |
| | Mammals‡ (chicken) | 135 | 153 | 106 | 141 | 145 | 217 |

* Changes detected using our phylogenetic maximum-likelihood procedure at a false-discovery rate of $<0.05$.
† Expression switch on the branch connecting great apes and macaque. Changes cannot be polarized owing to a lack of outgroup species.
‡ Expression switch on the branch connecting mammals and chicken. Changes cannot be polarized owing to a lack of outgroup species.

amniote orthologues and 3,314 (~33%) of 9,969 tested primate orthologues experienced at least one significant expression shift in one of the six organs during amniote or primate evolution, respectively. Our method is designed to detect selectively driven expression shifts, but might also detect shifts due to genetic drift or other non-adaptive forces (for example, biased gene conversion[45]), gene dosage alterations (for example, during sex chromosome differentiation; see below and P. Julien *et al.*, submitted manuscript) or cellular composition changes. In any case, our results provide an extensive list of candidates for potentially adaptive expression changes.

Although the data are not always directly comparable between tissues and lineages, owing to differences in statistical power associated with sampling and/or branch length differences, some global patterns stand out. In 11 of 15 mammalian lineages, the testis features the largest number of significant changes (Table 2), indicating that the rapid divergence of the testis (Fig. 2a, b) is at least partly explained by specific selective regimes acting on this organ, consistent with previous human–chimpanzee expression comparisons[26] and protein-coding sequence analyses[46]. By contrast, the brain shows few expression shifts in mammals, in agreement with the low rate of gene expression change in this tissue (Fig. 2a, b), with one notable exception: along the primate ancestral branch, by far the largest number of significant shifts (141 of 157) is found for the brain, which may be explained by the evolution of more complex cognitive functions, alterations of cellular composition and/or sampling differences (prefrontal cortex was sampled for primates, and whole brain, except cerebellum, was sampled for non-primates; Supplementary Note). The evolution of the cerebellum involved larger numbers of significant expression shifts than did that of the rest of the brain in 14 of 16 lineages (Table 2); this might underlie alterations in motor control functions among mammals.

Literature searches provide insights into potential functional implications of expression shifts for some of the most significant changes (see Supplementary Tables 27–42 for gene ontology analysis results). For example, the top candidate for adaptive expression change in the cortex on the terminal human branch, *LIX1* (Fig. 3b), which is strongly upregulated in humans (Benjamini–Hochberg corrected $P = 0.0183$), has a crucial role in motor neuron development and maintenance[47]. This gene is also upregulated in the human cerebellum ($P = 0.0209$) and is a component of the human-specific brain expression module (ID #p173 in Supplementary Table 10 and database at http://www.unil.ch/cbg/ISA/species). Other examples (Supplementary Figs 8 and 9) include *COL25A1*, which has the most significant expression shift in the brain of the common primate ancestor (reduced expression, corrected $P = 0.0032$; Fig. 3b) and leads to behavioural abnormalities when overexpressed in animal models[48]; and *TRMT1L*, the *TRM1*-like gene that affects motor coordination and exploratory behaviour[49] and shows higher expression levels in platypus brain (corrected $P < 10^{-8}$; Fig. 3b).

Our analysis of tissue transcriptomes from all major mammalian lineages refines previous hypotheses and provides many new clues to the function and evolution of mammalian genomes. This work marks the beginning of the exploitation of the reported transcriptome data, which will facilitate future investigations of mammalian genome biology.

## METHODS SUMMARY

We extracted high-quality RNA and prepared 131 polyadenylated RNA-seq libraries using standard protocols. Libraries were sequenced on Illumina Genome Analyser IIx platforms. We refined existing genome annotations to resolve potentially confounding factors and establish constitutive and alternative exons using a segmentation–clustering approach. We constructed orthologous gene sets based on retrieved lists of 1:1 orthologous genes for each species pair. To assess the influence of annotation heterogeneities on between-species variation, we determined sets of constitutive exon sequences that perfectly align across all species. On the basis of our refined annotations, final read mapping positions were established using a procedure that resolves read mapping ambiguities. We calculated standard expression values (RPKM) that were normalized across species and tissues on the basis of rank-conserved genes and a median-scaling procedure. Various biological analyses were performed using these data, including the development of a phylogenetic maximum-likelihood approach to detect significant expression shifts of individual genes.

Sequencing data have been deposited in the Gene Expression Omnibus and have been made directly available to Ensembl for annotation purposes. The different expression level datasets are provided as Supplementary Data Sets 1 and 2. All intermediate and final results and data are available from the authors on request.

**Full Methods** and any associated references are available in the online version of the paper at www.nature.com/nature.

**Received 11 April; accepted 5 September 2011.**

1. Kemp, T. S. *The Origin and Evolution of Mammals* (Oxford Univ. Press, Oxford, 2005).
2. Ponting, C. P. The functional repertoires of metazoan genomes. *Nature Rev. Genet.* **9,** 689–698 (2008).
3. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188,** 107–116 (1975).
4. Caceres, M. *et al.* Elevated gene expression levels distinguish human from non-human primate brains. *Proc. Natl Acad. Sci. USA* **100,** 13030–13035 (2003).
5. Enard, W. *et al.* Intra- and interspecific variation in primate gene expression patterns. *Science* **296,** 340–343 (2002).
6. Khaitovich, P., Enard, W., Lachmann, M. & Paabo, S. Evolution of primate gene expression. *Nature Rev. Genet.* **7,** 693–702 (2006).
7. Gilad, Y., Oshlack, A., Smyth, G. K., Speed, T. P. & White, K. P. Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* **440,** 242–245 (2006).
8. Uddin, M. *et al.* Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles. *Proc. Natl Acad. Sci. USA* **101,** 2957–2962 (2004).
9. Liao, B. Y. & Zhang, J. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol. Biol. Evol.* **23,** 530–540 (2006).

10. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10,** 57–63 (2009).
11. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464,** 773–777 (2010).
12. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).
13. Blekhman, R., Marioni, J. C., Zumbo, P., Stephens, M. & Gilad, Y. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.* **20,** 180–189 (2010).
14. Babbitt, C. C. *et al.* Both noncoding and protein-coding RNAs contribute to gene expression evolution in the primate brain. *Genome Biol. Evol.* **2,** 67–79 (2010).
15. Hubbard, T. J. *et al.* Ensembl 2009. *Nucleic Acids Res.* **37,** D690–D697 (2009).
16. Chodroff, R. A. *et al.* Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biol.* **11,** R72 (2010).
17. Clark, M. B. *et al.* The reality of pervasive transcription. *PLoS Biol.* **9,** e1000625 (2011).
18. Goodman, M. The genomic record of humankind's evolutionary roots. *Am. J. Hum. Genet.* **64,** 31–39 (1999).
19. Caswell, J. L. *et al.* Analysis of chimpanzee history based on genome sequence alignments. *PLoS Genet.* **4,** e1000057 (2008).
20. Harcourt, A. H., Harvey, P. H., Larson, S. G. & Short, R. V. Testis weight, body weight and breeding system in primates. *Nature* **293,** 55–57 (1981).
21. Li, W. H., Ellsworth, D. L., Krushkal, J., Chang, B. H. & Hewett-Emmett, D. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol. Phylogenet. Evol.* **5,** 182–187 (1996).
22. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
23. Warren, W. C. *et al.* Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453,** 175–183 (2008).
24. Keightley, P. D., Lercher, M. J. & Eyre-Walker, A. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* **3,** e42 (2005).
25. Marcus, G. *The Birth of the Mind* (Basic Books, 2004).
26. Khaitovich, P. *et al.* Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309,** 1850–1854 (2005).
27. Chan, E. T. *et al.* Conservation of core gene expression in vertebrate tissues. *J. Biol.* **8,** 33 (2009).
28. Kaessmann, H. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20,** 1313–1326 (2010).
29. Birkhead, T. R. & Pizzari, T. Postcopulatory sexual selection. *Nature Rev. Genet.* **3,** 262–273 (2002).
30. Veyrunes, F. *et al.* Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome Res.* **18,** 965–973 (2008).
31. Potrzebowski, L. *et al.* Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol.* **6,** e80 (2008).
32. Grützner, F. *et al.* In the platypus a meiotic chain of ten sex chromosomes shares genes with the bird Z and mammal X chromosomes. *Nature* **432,** 913–917 (2004).
33. Potrzebowski, L., Vinckenbosch, N. & Kaessmann, H. The emergence of new genes on the young therian X. *Trends Genet.* **26,** 1–4 (2010).
34. Ross, M. T. *et al.* The DNA sequence of the human X chromosome. *Nature* **434,** 325–337 (2005).
35. Rice, W. R. Sex chromosomes and the evolution of sexual dimorphism. *Evolution* **38,** 735–742 (1984).
36. Charlesworth, B. Model for evolution of Y chromosomes and dosage compensation. *Proc. Natl Acad. Sci. USA* **75,** 5618–5622 (1978).
37. Zhang, Y. E., Vibranovski, M. D., Landback, P., Marais, G. A. & Long, M. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol.* **8,** e1000494 (2010).
38. Wilson, M. A. & Makova, K. D. Evolution and survival on eutherian sex chromosomes. *PLoS Genet.* **5,** e1000568 (2009).
39. Bachtrog, D., Jensen, J. D. & Zhang, Z. Accelerated adaptive evolution on a newly formed X chromosome. *PLoS Biol.* **7,** e82 (2009).
40. Ihmels, J., Bergmann, S. & Barkai, N. Defining transcription modules using large-scale gene expression data. *Bioinformatics* **20,** 1993–2003 (2004).
41. Xiong, Y. *et al.* RNA sequencing shows no dosage compensation of the active X-chromosome. *Nature Genet.* **42,** 1043–1047 (2010).
42. Kemkemer, C., Kohn, M., Kehrer-Sawatzki, H., Fundele, R. H. & Hameister, H. Enrichment of brain-related genes on the mammalian X chromosome is ancient and predates the divergence of synapsid and sauropsid lineages. *Chromosome Res.* **17,** 811–820 (2009).
43. Haygood, R., Babbitt, C. C., Fedrigo, O. & Wray, G. A. Contrasts between adaptive coding and noncoding changes during human evolution. *Proc. Natl Acad. Sci. USA* **107,** 7853–7857 (2010).
44. Schoenemann, P. T., Sheehan, M. J. & Glotzer, L. D. Prefrontal white matter volume is disproportionately larger in humans than in other primates. *Nature Neurosci.* **8,** 242–252 (2005).
45. Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.* **10,** 285–311 (2009).
46. Nielsen, R. *et al.* A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3,** e170 (2005).
47. Fyfe, J. C. *et al.* An approximately 140-kb deletion associated with feline spinal muscular atrophy implies an essential LIX1 function for motor neuron survival. *Genome Res.* **16,** 1084–1090 (2006).
48. Tong, Y., Xu, Y., Scearce-Levie, K., Ptacek, L. J. & Fu, Y. H. *COL25A1* triggers and promotes Alzheimer's disease-like pathology *in vivo*. *Neurogenetics* **11,** 41–52 (2010).
49. Vauti, F. *et al.* The mouse *Trm1-like* gene is expressed in neural tissues and plays a role in motor coordination and exploratory behaviour. *Gene* **389,** 174–185 (2007).


**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** We thank K. Harshman and the LGTF for high-throughput sequencing support; I. Xenarios and the Vital-IT computational facility (Swiss Institute of Bioinformatics) for computational support; P. Jensen and L. Andersson for the red jungle fowl samples; E. Ait Yahya Graison and A. Reymond for C57BL/6J mouse RNA-seq data from male brain; C. Henrichsen and A. Reymond for wild-mouse samples; T. Daish, A. Casey, S. Lim, R. Jones and Glenrock station for platypus tissue collection and sample preparation; all other people and institutions that provided samples (Supplementary Table 1); W. Enard for ape sample organization; the members of the Kaessmann group for discussions; J. Meunier for statistical support; D. Cortez and M. Warnefors for comments on the manuscript; and R. Durbin and the Gorilla Genome Analysis Consortium for making the gorilla genome data available and for granting permission to use them for RNA-seq read mapping before publication. This research was supported by grants from the European Research Council (Starting Independent Researcher Grant: 242597, SexGenTransEvolution) and the Swiss National Science Foundation (grant 31003A_130287), to H.K. S.B. was supported by the Swiss National Science Foundation (grant 31003A_130691/1), the Swiss Institute of Bioinformatics and the European Framework Project 6 (AnEuploidy and EuroDia projects). S.P. was supported by the European Research Council (ERC-2008-AdG, TWOPAN) and by the Max Planck Society. A.N. was supported by a long-term FEBS postdoctoral fellowship. F.G. is an ARC Australian Research Fellow.

**Author Contributions** D.B., G.C., H.K., A.N. and P.H. performed biological data analyses. M.S. organized the RNA-seq data production. D.B. and A.N. processed and mapped the reads. A.N. refined genome annotations and established definitions and alignments of constitutive exons. M.S., A.L., F.W.A. and A.A.-P. prepared samples and generated RNA-seq libraries. M.W. prepared samples. P.J. contributed ideas regarding data analyses. F.W.A. coordinated ape RNA-seq data production. M.K. processed ape RNA-seq data. U.Z. extracted and organized *Monodelphis domestica* samples and advised on this species' biology. P.K. organized *Macaca mulatta* samples and provided general advice on gene expression evolution. F.G. organized and extracted platypus RNA samples and advised on this species' biology. P.H. developed the gene expression selection method and performed all corresponding analyses under the guidance of R.N. G.C. performed analyses using the iterative signature algorithm under the guidance of S.B. S.P. guided ape RNA-seq data production and processing. The project was supervised and originally designed by H.K. The paper was written by H.K. with input from all authors.

**Author Information** Sequencing data have been deposited in the Gene Expression Omnibus under accession code GSE30352. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to H.K. (henrik.kaessmann@unil.ch).

## METHODS

**Samples and RNA sequencing.** The 131 organ samples that provided the foundation for this study were obtained from various sources (Supplementary Note, section 1.1, and Supplementary Tables 1 and 2). To ensure comparability of data derived from homologous organs between species, several measures were taken. Most of the organs studied represent heterogeneous tissues whose structural and cellular composition may vary between species. To account for this issue, major parts of each tissue (covering the different structures/cells) were sampled and homogenized before RNA extraction, where possible. Given that the brain is a particularly heterogeneous tissue, we sampled two parts of the brain (prefrontal cortex and brain without cerebellum, depending on the species; cerebellum) for each species, one of which (cerebellum) is well defined, structurally similar and easily dissectible in all species in spite of the major differences in brain size among the amniote species studied (see Supplementary Note, section 1.1, for details).

Total RNA was extracted using the Trizol (Invitrogen) procedure or RNAeasy/RNAeasy Lipid/miRNeasy (Qiagen) column purification kits as indicated in Supplementary Table 1. RNA quality was assessed using an Agilent 2100 Bioanalyser. Only samples with high RNA integrity values (Supplementary Table 1) were used in this study.

Sequencing libraries were prepared using the mRNA-Seq Sample Prep Kit (Illumina) according to the manufacturer's instructions. Briefly, polyadenylated RNA was isolated using a poly-dT bead procedure and then chemically fragmented and randomly primed for reverse transcription. After second-strand synthesis, the ends of the double-stranded complementary DNA were repaired. After 3′-end adenylation of these products, Illumina Paired-End Sequencing adapters were ligated to the blunt ends of the cDNA fragments. Ligated products were run on gels; 250–300-bp fragments were excised and then PCR-amplified (15 cycles). After column purification, qualities of the resulting libraries were assessed using Agilent 2100 Bioanalysers. Potential influences on RNA sequencing results due to different experimenters preparing the libraries were ruled out on the basis of RNA-seq data analysis of replicate libraries prepared by the different experimenters (Supplementary Note, section 1.2). The RNA-seq libraries were each sequenced (76 cycles) in at least one lane of the Illumina Genome Analyser IIx platform according to the manufacturer's specifications. Technical replicates (that is, running the same library on different machines) were performed to rule out potential biases during the sequencing step (Supplementary Note, section 1.2).

After sequencing, we processed the fluorophore intensity files with the IBIS base caller[50], in addition to applying the standard Illumina base-calling algorithms. All subsequent analyses were performed on the IBIS-called reads, as the number of correctly mapped reads was significantly increased with this base-calling approach (Supplementary Note Fig. 1 and Supplementary Note Table 1).

**Initial read mapping and refinement of genomic annotations.** We used TOPHAT[51] to align the reads to the reference genome sequences and to extract splice junctions, without relying on the genome annotations. We developed a specific procedure to improve the splice junction detection for genes that have recent retrocopies (Supplementary Note, section 1.3). We filtered the alignments provided by TOPHAT to extract unambiguously mapping reads, from which we built a set of transcribed islands and splice junctions for each RNA-seq sample. Only junctions with GT–AG and GC–AG splice sites, for which the sense strand can be reliably inferred, were included in this analysis.

We extracted genomic annotations from Ensembl[52], release 57. These annotations were further extended by recursively adding the coordinates of the transcribed islands to the gene models, for those islands that were connected by splice junctions to previously known exons of the same gene. For this procedure, we only considered transcribed islands that were supported by at least two unambiguously mapping reads, that were in proximity to genes (distance to previously known gene boundaries, $<100$ kb) and that were not connected to multiple genes.

As the annotation extension procedure can result in the inclusion of (possibly non-functional) retained introns in the gene models, we further refined the annotations by using the splice junction coordinates to define exon boundaries precisely (Supplementary Note, section 1.4).

**Constitutive exon definition.** Before evaluating gene expression levels, we sought to eliminate minor splice isoforms from the gene models, to reduce the level of splicing-related noise in our data. As proposed previously[53], we used information on splice junctions to detect four types of alternative transcription event: skipped (or cassette) exons, retained introns and alternative 5′ and 3′ splice sites. In addition, we analysed the read coverage variation within genes with a segmentation–clustering approach[54] to define and remove regions with unusually low coverage (Supplementary Note, section 1.5). For each alternative transcription event, we quantified the two possible isoforms on the basis of the splice junction frequencies and on the read coverage (Supplementary Note, section 1.5). We then defined 'constitutive' exon segments as those segments that belong to isoforms with frequencies of more than 15%, in all samples. For genes with low expression (that is, with an average per-base read coverage of less than three), all exons were considered to be constitutive, because the minor isoform identification is not feasible in these cases.

**Orthologous gene sets and exon sequence alignment.** We retrieved the list of 1:1 orthologous genes for each pair of species in our set from the Ensembl database[55], release 57. From these pairwise orthology relationships, we extracted 5,636 gene families that have 1:1 orthology relationships among all the species in our set, as well as 13,277 gene families for primates.

Given the heterogeneity of genomic annotations, we wanted to exclude the possibility that gene expression variation between species might be due to the fact that gene expression levels are computed on heterogeneous sequences. Thus, we built a set of constitutive aligned exons for the 1:1 orthologous families. To do so, we aligned the cDNA sequences of the orthologous gene families using TBA[56]. We filtered these alignments to extract perfectly aligned blocks of sequence (no gaps were permitted) that corresponded to exon parts considered to be 'constitutive' in all species.

**Final read mapping.** To ensure unambiguous read mapping and optimal subsequent calculations of expression levels, the final read mapping procedure was based on our refined genome annotations (see above) and involved several steps (Supplementary Note Fig. 14). To prepare for the mapping of reads, we first built a library of splice junction sequences on the basis of the refined exon annotations. As a further preparation step, we then sought to assess the number of theoretically possible unique reads per given annotation element (exon, exon part and so on). Specifically, we derived all possible read sequences for each annotation (~200 million reads, depending on the genome) and mapped each of these artificial reads onto the respective genome sequence as well as the sequences from the splice junction library, using BOWTIE[57]. We then calculated the unique read coverage per genomic element and stored this information for the mapping procedure.

The final mapping positions of RNA-seq reads for a given genome were established as follows. We first mapped each read onto the genome sequence and (in parallel) the sequences from the splice junction library, using BOWTIE[57]. This mapping information served as input for an algorithm that was designed to resolve ambiguities of reads with multiple mapping positions in the genome and calculate basic expression level values for each gene. Specifically, in the case of overlapping mappings, the mappings with the lowest numbers of mismatches were chosen (in the case of identical numbers of mismatches, spliced reads were favoured). Reads that mapped equally well to different genomic loci (for example to different duplicate gene copies) were resolved in the following way. We first calculated preliminary transcription levels by dividing the number of reads that map uniquely to each locus by its unique read coverage (see above). Non-unique reads were then distributed among annotated genomic elements according to these ratios (that is, loci received reads in proportion to their unique read mapping ratios). If two or several loci had identical sequences (that is, they had no uniquely mapping reads), reads were distributed evenly among them—if they were all multi-exonic. In the case of multi-exonic 'parental' genes and their identical retroposed gene copies, reads were assigned exclusively to the parental genes, given that the majority of retrocopies (in particular recent ones) were likely to be non-functional or at least expressed at very low levels[58].

**Expression levels and normalization.** On the basis of final read assignments described in the previous section, we calculated standard RPKM expression values[59] (that were then $\log_2$-transformed) for the orthologous genes as defined above. To render the data comparable across species and tissues, we then normalized these expression values by a scaling procedure. Specifically, among the genes with expression values in the interquartile range, we identified the 1,000 genes that have the most-conserved ranks among samples and assessed their median expression levels in each sample. We then derived scaling factors that adjust these medians to a common value. Finally, these factors were used to scale expression values of all genes in the samples (Supplementary Note Fig. 16). We note that other normalization procedures led to similar results.

**Biological analyses.** Various biological analyses were performed using these data, which also involved the development of a new method (a phylogenetic maximum-likelihood approach to detect selectively driven expression change). For a full description of these analyses and newly developed approaches, see Supplementary Note.

The principal-component analysis on gene expression levels was performed using the 'prcomp' function in the 'stats' package in R (http://www.R-project.org/).

We constructed expression trees using the neighbour-joining approach, based on pairwise distance matrices between samples. The distance between samples was computed as $1 - \rho$, where $\rho$ is Spearman's correlation coefficient, because it is insensitive to outliers and any potential inaccuracies in the normalization

procedure. Euclidean distances were used as a control (Supplementary Fig. 3). The neighbour-joining trees were constructed using functions in the 'ape' package[60] in R. The reliability of branching patterns was assessed with bootstrap analyses (the 5,636 amniote 1:1 orthologous genes and the 13,277 primate 1:1 orthologous genes were randomly sampled with replacement 1,000 times). The bootstrap values are the proportions of replicate trees that share the branching pattern of the majority-rule consensus tree shown in the figures.


50. Kircher, M., Stenzel, U. & Kelso, J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol.* **10,** R83 (2009).
51. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
52. Hubbard, T. J. *et al.* Ensembl 2009. *Nucleic Acids Res.* **37,** D690–D697 (2009).
53. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476 (2008).
54. Picard, F., Robin, S., Lebarbier, E. & Daudin, J. J. A segmentation/clustering model for the analysis of array CGH data. *Biometrics* **63,** 758–766 (2007).
55. Vilella, A. J. *et al.* EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19,** 327–335 (2009).
56. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14,** 708–715 (2004).
57. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
58. Kaessmann, H., Vinckenbosch, N. & Long, M. RNA-based gene duplication: mechanistic and evolutionary insights. *Nature Rev. Genet.* **10,** 19–31 (2009).
59. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (2008).
60. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20,** 289–290 (2004).